frontiers

Real-Time Tracking of Selective Auditory Attention from M/EEG: A Bayesian Filtering Approach

Sina Miran [1], **Sahar Akram** [2], **Alireza Sheikhattar** [1], **Jonathan Z. Simon** [1,3,4], **Tao Zhang** [5], **and Behtash Babadi** [1,3,*]

[1] *Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA*
[2] *Facebook, Menlo Park, CA 94025, USA*
[3] *Institute for Systems Research, University of Maryland, College Park, MD 20742, USA*
[4] *Department of Biology, University of Maryland, College Park, MD 20742, USA*
[5] *Starkey Hearing Technologies, Eden Prairie, MN 55344, USA*

Correspondence*:
Behtash Babadi
behtash@umd.edu

ABSTRACT

Humans are able to identify and track a target speaker amid a cacophony of acoustic interference, an ability which is often referred to as the cocktail party phenomenon. Results from several decades of studying this phenomenon have culminated in recent years in various promising attempts to decode the attentional state of a listener in a competing-speaker environment from non-invasive neuroimaging recordings such as magnetoencephalography (MEG) and electroencephalography (EEG). To this end, most existing approaches compute correlation-based measures by either regressing the features of each speech stream to the M/EEG channels (the decoding approach) or vice versa (the encoding approach). To produce robust results, these procedures require multiple trials for training purposes. Also, their decoding accuracy drops significantly when operating at high temporal resolutions. Thus, they are not well-suited for emerging real-time applications such as smart hearing aid devices or brain-computer interface systems, where training data might be limited and high temporal resolutions are desired. In this paper, we close this gap by developing an algorithmic pipeline for real-time decoding of the attentional state. Our proposed framework consists of three main modules: 1) Real-time and robust estimation of encoding or decoding coefficients, achieved by sparse adaptive filtering, 2) Extracting reliable markers of the attentional state, and thereby generalizing the widely-used correlation-based measures thereof, and 3) Devising a near real-time state-space estimator that translates the noisy and variable attention markers to robust and statistically interpretable estimates of the attentional state with minimal delay. Our proposed algorithms integrate various techniques including forgetting factor-based adaptive filtering, ℓ_1-regularization, forward-backward splitting algorithms, fixed-lag smoothing, and Expectation Maximization. We validate the performance of our proposed framework using comprehensive simulations as well as application to experimentally acquired M/EEG data. Our results reveal that the proposed real-time

algorithms perform nearly as accurately as the existing state-of-the-art offline techniques, while providing a significant degree of adaptivity, statistical robustness, and computational savings.

Keywords: attention, auditory, real-time, dynamic estimation, EEG, MEG, state-space models, Bayesian filtering

1 INTRODUCTION

The ability to select a single speaker in an auditory scene, consisting of multiple competing speakers, and maintain attention to that speaker is one of the hallmarks of human brain function. This phenomenon has been referred to as the cocktail party effect (Brungart, 2001; McDermott, 2009; Haykin and Chen, 2005). The mechanisms underlying the real-time process by which the brain segregates multiple sources in a cocktail party setting, have been the topic of active research for decades (Cherry, 1953; Middlebrooks et al., 2017). Although the details of these mechanisms are for the most part unknown, various studies have pointed to the role of specific neural processes involved in this function. As the acoustic signals propagate through the auditory pathway, they are decomposed into spectrotemporal features at different stages, and a rich representation of the complex auditory environment reaches the auditory cortex. It has been hypothesized that the perception of an auditory object is the result of adaptive binding as well as discounting of these features (Bregman, 1994; Griffiths and Warren, 2004; Fishman and Steinschneider, 2010; Shamma et al., 2011).

From a computational modeling perspective, there have been several attempts at designing so-called "attention decoders", where the goal is to reliably decode the attentional focus of a listener in a multi-speaker environment using non-invasive neuroimaging techniques like electroencephalography (EEG) (O'Sullivan et al., 2015; Power et al., 2012; Mirkovic et al., 2015; Zink et al., 2017) and magnetoencephalography (MEG) (Ding and Simon, 2012a,b; Akram et al., 2014, 2016, 2017). These methods are typically based on reverse correlation or estimating linear encoding/decoding models using off-line regression techniques, and thereby detecting specific lags in the model coefficients that are modulated by the attentional state (Kaya and Elhilali, 2017). For instance, encoding coefficients comprise salient peaks at a typical lag of ~ 100 ms for MEG (Ding and Simon, 2012a), and envelope reconstruction performance is optimal at a lag of ~ 200 ms for EEG (O'Sullivan et al., 2015).

Although the foregoing approaches have proven successful in reliable attention decoding, they have two major limitations that make them less appealing for emerging real-time applications such as Brain-Computer Interface (BCI) systems and smart hearing aids. First, the temporal resolution of existing approaches for reliable attention decoding is on the order of ~ 10 s, and their decoding accuracy drops significantly when operating at temporal resolutions of ~ 1 s, i.e., the time scale at which humans are able to switch attention from one speaker to another (Zink et al., 2016, 2017). Second, approaches based on linear regression (e.g., reverse correlation) need large training datasets, often from multiple subjects and trials, to estimate the decoder/encoder reliably. Access to such training data is only possible through repeated calibration stages, which may not always be possible in real-time applications with potential variations in recording settings. While recent results (Akram et al., 2014, 2016) address the first shortcoming by employing state-space models and thereby producing robust estimates of the attentional state from limited data at high temporal resolutions, they are not yet suitable for real-time applications as they operate in the so-called "batch-mode" regime, i.e., they require the entire data from a trial at once in order to estimate the attentional state.

In this paper, we close this gap by designing a modular framework for real-time attention decoding from non-invasive M/EEG recordings that overcomes the aforementioned limitations using techniques from Bayesian filtering. Our proposed framework includes three main modules. The first module pertains to

estimating *dynamic* models of decoding/encoding in *real-time*. To this end, we use the forgetting factor mechanism of the Recursive Least Squares (RLS) algorithm together with the ℓ_1 regularization penalty from Lasso to capture the dynamics in the data while preventing overfitting (Akram et al., 2017; Sheikhattar et al., 2015a). The real-time inference is then efficiently carried out using a Forward-Backward Splitting (FBS) procedure (Combettes and Pesquet, 2011). In the second module, we extract an attention-modulated feature, which we refer to as "attention marker", as a function of the M/EEG recordings, the estimated encoding/decoding coefficients, and the auditory stimuli. For instance, the attention marker can be a correlation-based measure or the magnitude of certain peaks in the model coefficients. We carefully design the attention marker features to capture the attention modulation and thereby maximally separate the contributions of the attended and unattended speakers in the neural response in both MEG and EEG applications.

The extracted features are then passed to a novel state-space estimator in the third module, and thereby are translated into probabilistic, robust, and dynamic measures of the attentional state, which can be used for soft-decision making in real-time applications. The state-space estimator is based on Bayesian fixed-lag smoothing, and operates in *near real-time* with controllable delay. The fixed-lag design creates a trade-off between real-time operation and robustness to stochastic fluctuations. In addition, we modify the Expectation-Maximization algorithm and the nonlinear filtering and smoothing techniques of (Akram et al., 2016) for real-time implementation. Compared to existing techniques, our algorithms require minimal supervised data for initialization and tuning, which makes them more suitable for the applications of real-time attention decoding with limited training data. In order to validate our real-time attention decoding algorithms, we apply them to both simulated and experimentally recorded EEG and MEG data in dual-speaker environments. Our results suggest that the performance of our proposed framework is comparable to the state-of-the-art results of (O'Sullivan et al., 2015; Mirkovic et al., 2015; Akram et al., 2016), while operating in near real-time with $\sim 2\,\mathrm{s}$ delay.

The rest of the paper is organized as follows: In Section 2, we develop the three main modules in our proposed framework as well as the corresponding estimation algorithms. We present the application of our framework to both synthetic and experimentally recorded M/EEG data in Section 3, followed by discussion and concluding remarks in Section 4.

2 MATERIAL AND METHODS

Figure 1 summarizes our proposed framework for real-time tracking of selective auditory attention from M/EEG. In the *Dynamic Encoder/Decoder Estimation* module, the encoding/decoding models are fit to neural data in real-time. The *Attention Marker* module uses the estimated model coefficients as well as the recorded data to compute a feature that is modulated by the instantaneous attentional state. Finally, in the *State-Space Model* module, the foregoing features are refined through a linear state-space model with nonlinear observations, resulting in robust and dynamic estimates of the attentional state.

In Section 2.1, we formally define the dynamic encoding and decoding models, and develop low-complexity and real-time techniques for their estimation. This is followed by Section 2.2, in which we define suitable attention markers for M/EEG inspired by existing literature. In Section 2.3, we propose a state-space model that processes the extracted attention markers in order to produce near real-time estimates of the attentional state with minimal delay.

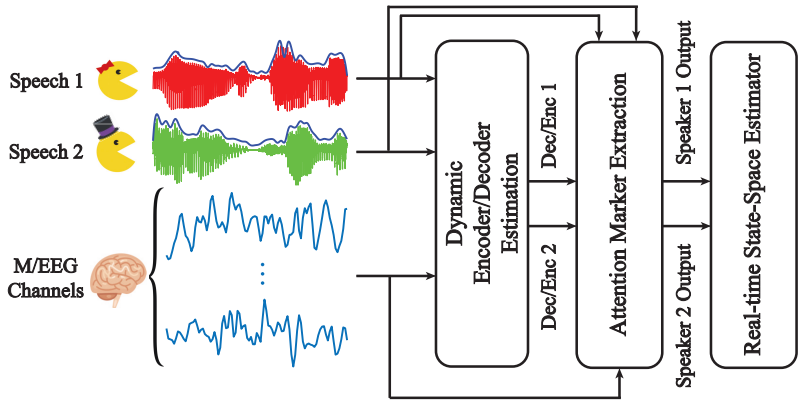

Figure 1. A schematic depiction of our proposed framework for real-time tracking of selective auditory attention from M/EEG.

2.1 Dynamic Encoding and Decoding Models

The role of a neural encoding model is to map the stimulus to the neural response. Inspired by existing literature on attention decoding (Ding and Simon, 2012a; O'Sullivan et al., 2015; Akram et al., 2016), we take the speech envelopes as covariates representing the stimuli. The neural response is manifested in the M/EEG recordings. Encoding models can be used to predict the neural response from the stimulus. In contrast, in a neural decoding model, the goal is to express the stimulus as a function of the neural response. Inspired by previous studies, we consider linear encoding and decoding models in this work.

The encoding and decoding models can be cast as mathematically dual formulations. In a dual-speaker environment, let $s_t^{(1)}$ and $s_t^{(2)}$ denote the speech envelopes (in logarithmic scale), corresponding to speakers 1 and 2, respectively, for $t = 1, 2, \ldots, T$. Also, let e_t^c denote the neural response recorded at time t and channel c, for $c = 1, 2, \ldots, C$. Throughout the paper, we assume the same sampling frequency f_s for both the M/EEG channels and the envelopes. Consider consecutive and non-overlapping windows of length W, and define $K := \lfloor \frac{T}{W} \rfloor$. We consider piece-wise constant dynamics for the encoding and decoding coefficients, in which the coefficients assume to be constant over each window. Note that we define the *temporal resolution* in an attention decoding procedure as the duration of a data segment to which a measure of the attentional state is attributed. Therefore, $\frac{W}{f_s}$ determines the temporal resolution in our attention decoding framework.

In the encoding setting, we define the vector $\mathbf{s}_t^{(i)} := [s_t^{(i)}, s_{t-1}^{(i)}, \ldots, s_{t-L_e}^{(i)}]^\top$ for $i = 1, 2$, where L_e is the total lag considered in the model. Also, let E_t denote a generic linear combination of $e_t^1, e_t^2, \ldots, e_t^C$ with some fixed set of weights. These weights can be set to select a single channel, i.e., $E_t = e_t^c$ for some c, or they can be pre-estimated from training data so that E_t represents the dominant auditory component of the neural response (de Cheveigne and Simon, 2008). The encoding coefficients then relate $\mathbf{s}_t^{(i)}$ to E_t. In the decoding setting, we define the vector $\mathbf{e}_t := [e_t^1, e_t^2, \ldots, e_t^C]^\top$ and $\boldsymbol{\mathcal{E}}_t := \left[1, \mathbf{e}_t^\top, \mathbf{e}_{t+1}^\top, \ldots, \mathbf{e}_{t+L_d}^\top\right]^\top$, where L_d is the total lag in the decoding model and determines the extent of future neural responses affected by the current stimuli. The decoding coefficients then relate $\boldsymbol{\mathcal{E}}_t$ to $s_t^{(i)}$.

Our goal is to recursively estimate the encoding/decoding coefficients in a real-time fashion as the new data samples become available. In addtion, we aim to simultaneously induce adaptivity of the parameter

estimates and capture their sparsity. To this end, we employ the following generic optimization problem:

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{k} \lambda^{k-j} \parallel \mathbf{y}_j - \mathbf{X}_j \boldsymbol{\theta} \parallel_2^2 + \gamma \parallel \boldsymbol{\theta} \parallel_1, \quad k = 1, 2, \ldots, K \tag{1}$$

where $\mathbf{y}_j$ and $\mathbf{X}_j$ are respectively the vector of response variables and the matrix of covariates pertinent to window j, $\boldsymbol{\theta}$ is the parameter vector, $\lambda \in (0, 1]$ is the forgetting factor, and γ is a regularization parameter. The optimization problem of Eq. 1 is a modified version of the LASSO problem (Tibshirani, 1996).

For the encoding problem, we define $\mathbf{y}_k \coloneqq \left[E_{(k-1)W+1}, E_{(k-1)W+2}, \ldots, E_{kW}\right]^\top$ and $\mathbf{X}_k^{(i)} \coloneqq \left[\mathbf{s}_{(k-1)W+1}^{(i)}, \mathbf{s}_{(k-1)W+2}^{(i)}, \ldots, \mathbf{s}_{kW}^{(i)}\right]^\top$, for $k = 1, 2, \ldots, K$ and $i = 1, 2$. Therefore, the full encoding covariate matrix at the k^{th} window is defined as $\mathbf{X}_k \coloneqq \left[\mathbb{1}_{W\times 1}, \mathbf{X}_k^{(1)}, \mathbf{X}_k^{(2)}\right]$, where the all-ones vector $\mathbb{1}_{W\times 1}$ corresponds to the regression intercept. In the decoding problem, we define $\mathbf{y}_k = \mathbf{s}_k^{(i)} \coloneqq \left[s_{(k-1)W+1}^{(i)}, s_{(k-1)W+2}^{(i)}, \ldots, s_{kW}^{(i)}\right]^\top$, where $i \in \{1, 2\}$. Also, the full decoding covariate matrix at the k^{th} window is $\mathbf{X}_k \coloneqq \left[\boldsymbol{\mathcal{E}}_{(k-1)W+1}, \boldsymbol{\mathcal{E}}_{(k-1)W+1}, \ldots, \boldsymbol{\mathcal{E}}_{kW}\right]^\top$, for $k = 1, 2, \ldots, K$.

The optimization problem of Eq. (1) has a useful Bayesian interpretation: if the observation noise were i.i.d. Gaussian, and the parameters were exponentially distributed, it is akin to the maximum *a posteriori* (MAP) estimate of the parameters. The quadratic terms correspond to the exponentially-weighted log-likelihood of the observations up to window k, and the ℓ_1-norm corresponds to the log-density of an independent exponential prior on the elements of $\boldsymbol{\theta}$. The exponential prior serves as an effective regularization to promote sparsity of the estimate $\hat{\boldsymbol{\theta}}_k$. Note that we have $\boldsymbol{\theta} \in \mathbb{R}^{1+2(L_e+1)}$ for the encoding model and $\boldsymbol{\theta} \in \mathbb{R}^{1+C(L_d+1)}$ for the decoding model in (1).

Remark 1. The hyperparameter λ provides a tradeoff between the adaptivity and the robustness of estimated coefficients, and it can be determined based on the inherent dynamics in the data. The case of $\lambda = 1$ corresponds to the natural data log-likelihood, i.e., the batch-mode parameter estimates. It has been shown that $\frac{W}{1-\lambda}$ can serve as the *effective* number of recent samples used to calculate $\hat{\boldsymbol{\theta}}_k$ in (1) (Sheikhattar et al., 2015b). The parameter $\frac{W}{1-\lambda}$ can also be viewed as the dynamic integration time: it needs to be chosen long enough so that the estimation is stable, but also short enough to be able to capture the dynamics of neural process involved in switching attention. The hyperparameter γ controls the tradeoff between the Maximum Likelihood (ML) fit and the sparsity of estimated coefficients, and it is usually determined through cross-validation.

Remark 2. In the decoding problem, Eq. (1) is solved separately at each window for each speech envelope, resulting in a set of decoding coefficients per speaker. In the encoding setting, we combine the stimuli as explained and solve Eq. (1) once at each window to obtain both of the encoder estimates. If the encoding/decoding coefficients are expected to be sparse in a basis represented by the columns of a matrix $\mathbf{G}$, such as the Haar or Gabor bases, we can replace $\mathbf{X}_j$ in (1) by $\mathbf{X}_j\mathbf{G}$, for $j = 1, 2, \ldots, k$, and solve for $\hat{\boldsymbol{\theta}}_k$ as before. Then, the final encoding/decoding coefficients are given by $\mathbf{G}\hat{\boldsymbol{\theta}}_k$. In the context of encoding models, the coefficients are referred to as the Temporal Response Function (TRF) (Ding and Simon, 2012a; Akram et al., 2017). The TRFs are known to exhibit some degree of sparsity on a basis consisting of shifted Gaussian kernels (see Akram et al. (2017) for details).

Remark 3. It is worth discussing the rationale behind the dynamic updating of the encoding/decoding models, as opposed to considering fixed *canonical* encoding/decoding models common in existing work.

First, estimation of the canonical encoding/decoding models in existing literature requires large training datasets. In emerging real-time applications of attention decoding, access to such large supervised training datasets may not be feasible. In addition, slight changes to the electrode placement may require recalibration of the canonical encoders/decoders. Thus, by dynamic updating of the encoding/decoding models we aim at minimizing the amount of supervised training data, which can be a bottleneck in emerging real-time applications.

Second, recent results have shown that dynamics of the encoding/decoding models indeed carry important information regarding the underlying attention process (Ding and Simon, 2012a,b; Power et al., 2012; Golumbic et al., 2013; Akram et al., 2017). Therefore, dynamic estimates of these models can be beneficial in attention decoding. In order to mitigate the variability of our dynamic estimates of the encoding/decoding models, we have employed the ℓ_1-regularized least squares estimation framework with a forgetting factor.

In summary, we argue that the dynamic framework used here is more preferable for real-time applications with limited training data and in the presence of attention dynamics. It is worth noting that our modular framework can still be used if the encoder/decoder models are pre-estimated and fixed. We refer the reader to Section 2.3 and Remark 6 for more details.

Remark 4. Throughout the paper, we assume that the envelopes of the clean speech are available. Given that this assumption does not hold in practical scenarios, recent algorithms on the extraction of speech envelopes from acoustic mixtures (Biesmans et al., 2015; Aroudi et al., 2016; Biesmans et al., 2017; O'Sullivan et al., 2017; Van Eyndhoven et al., 2017) can be added as a pre-processing module to our framework.

Among the many existing algorithms for solving the modified LASSO problem of Eq. (1), we choose the Forward-Backward Splitting (FBS) algorithm (Combettes and Pesquet, 2011), also known as the proximal gradient method. When coupled with proper step-size adjustment methods, FBS is well-suited for real-time and low-complexity updates of $\hat{\boldsymbol{\theta}}_k$ at each window. In this work, we have used the FASTA software package (Goldstein et al., 2014) available online (Goldstein et al., 2015), which has built-in features for all the FBS stepsize adjustment methods. A detailed overview of the FBS algorithm and its properties is given in Section 1 of the Supplementary Material.

2.2 Attention Markers

We define the *attention marker* as a mapping function from the estimated encoding/decoding coefficients for each speaker as well as the data in each window to positive real numbers. To be more precise, at window k and for speaker i, in the context of encoding models, the attention marker takes the speaker's estimated encoding coefficients $\hat{\boldsymbol{\theta}}_k^{(i)}$, the speaker's covariate matrix $\mathbf{X}_k^{(i)}$, and the M/EEG responses $\mathbf{y}_k$ as inputs; similarly, in the context of decoding models, the attention marker takes the speaker's estimated decoding coefficients $\hat{\boldsymbol{\theta}}_k^{(i)}$, the M/EEG covariate matrix $\mathbf{X}_k$, and the speaker's speech envelope vector $\mathbf{y}_k^{(i)}$ as inputs. In both cases, the attention marker outputs a positive real number, which we denote by $m_k^{(i)}$ henceforth, for $i = 1, 2$ and $k = 1, 2, \ldots, K$. Thus, in the modular design of Fig. 1, at each window k, the two outputs $m_k^{(1)}$ and $m_k^{(2)}$ are passed from the Attention Marker module to the State-Space Model module as measures of the attentional state at window k.

In (O'Sullivan et al., 2015), a correlation-based measure has been adopted in the decoding model to classify the attended and the unattended speeches in a dual-speaker environment. The approach in (O'Sullivan et al., 2015) is based on estimating an *attended* (resp. *unattended*) decoder from the training data to reconstruct the attended (resp. unattended) speech envelope from EEG for each trial. Then, the correlation of this reconstructed envelope with each of the two speech envelopes is computed, and the

speaker with the larger correlation coefficient is deemed as the attended (resp. unattended) speaker. This method cannot be directly applied to the real-time setting, since the lack of abundant training data hinders reliable estimation of these decoders. However, assuming that the auditory M/EEG response is more influenced by the attended speaker than the unattended one, we can expect that the decoder corresponding to the *attended* speaker exhibits a higher performance in reconstructing the speech envelope it has been trained on. This can be inferred from the findings in (O'Sullivan et al., 2015), where a trained *attended* decoder results in 10% more attention decoding accuracy than a trained *unattended* decoder, as well as the findings in (Ding and Simon, 2012a). Inspired by these results, we can define the attention marker in the decoding scenario as the correlation magnitude between the speech envelope and its reconstruction by the corresponding decoder, i.e., $m_k^{(i)} = f\left(\hat{\boldsymbol{\theta}}_k^{(i)}, \mathbf{X}_k, \mathbf{y}_k^{(i)}\right) := \left|\mathrm{corr}\left(\mathbf{y}_k^{(i)}, \mathbf{X}_k\hat{\boldsymbol{\theta}}_k^{(i)}\right)\right|$ for $i = 1, 2$ and $k = 1, 2, \ldots, K$. As we will demonstrate later in Section 3, this attention marker is suitable for the analysis of EEG recordings.

In the context of cocktail party studies using MEG, it has been shown that the magnitude of the negative peak in the TRF of the attended speaker around a lag of 100 ms, referred to as the M100 component, is larger than that of the unattended speaker (Ding and Simon, 2012a; Akram et al., 2017, 2016). Inspired by these findings, in the encoding scenario applied to MEG data, we can define the attention marker $m_k^{(i)}$ to be the magnitude of the $\hat{\boldsymbol{\theta}}_k^{(i)}$ coefficients corresponding to the M100 component, for $i = 1, 2$ and $k = 1, 2, \ldots, K$.

Due to the inherent uncertainties in the M/EEG recordings, the limitations of non-invasive neuroimaging in isolating the relevant neural processes, and the unknown and likely nonlinear processes involved in auditory attention, the foregoing attention markers derived from linear models are not readily reliable indicators of the attentional state. Given ample training data, nevertheless, these attention markers have been validated using batch-mode analysis. However, their usage in a real-time setting at high temporal resolution requires more care, as the limited data in real-time applications and computation over small windows add more sources of uncertainty to the foregoing list. To address this issue, a state-space model is required in the real-time setting to correct for the uncertainties and stochastic fluctuations of the attention markers caused by the limited integration time in real-time application. We will discuss in detail the formulation and advantages of such a state-space model in the following subsection.

2.3 State-Space Model

In order to translate the attention markers $m_k^{(1)}$ and $m_k^{(2)}$, for $k = 1, 2, \ldots, K$, into a robust and statistically interpretable measure of the attentional state, we employ state-space models. Inspired by the models used in (Akram et al., 2016), we design a new state-space model and a corresponding estimator that operates in a fixed-lag smoothing fashion, and thereby admits real-time processing while maintaining the benefits of batch-mode state-space models. Recall that the index k corresponds to a window in time ranging from $t = (k-1)W + 1$ to $t = kW$; however, we refer to each index k as an *instance* when talking about the state-space model, so as not to conflate it with the sliding window in the forthcoming treatment.

Figure 2 displays the fixed-lag smoothing design of the state-space estimator. Suppose that we are at the instance $k = k_0$. We consider an *active* sliding window of length $K_A := K_B + K_F + 1$ as shown in Fig. 2, where K_F and K_B are respectively called the forward-lag and the backward-lag. In order to carry out the computations in real-time, we assume all of the attentional state estimates to be fixed prior to this window and only update our estimates for the instances within, based on $m_k^{(1)}$'s and $m_k^{(2)}$'s inside the window. In a fixed-lag framework, at $k = k_0$, the goal is to provide an estimate of the attentional state at instance

$k = k^*$, where $k^* = k_0 - K_F$. Thus, when using a decoding (resp. encoding) model, the *built-in* attention decoding delay of our framework is $(L_d + K_F W)/f_s$ (resp. $K_F W/f_s$) seconds. It is worth noting that in addition to the built-in delay, our attention decoding results are affected by another source of delay, which we refer to as the *transition* delay. The transition delay is due to the forgetting factor mechanism as well as the smoothing effect in the state-space estimation, which we will discuss further in Section 3.1. The parameter K_F creates a tradeoff between real-time and robust estimation of the attentional state. For $K_F = 0$, the estimation is carried out fully in real-time; however, the estimates lack robustness to the fluctuations of the outputs of the attention marker block. The backward-lag K_B determines the attention marker samples prior to k^* that are used in the inference procedure, and it controls the computational cost of the state-space model for fixed values of K_F. Throughout the rest of the paper, we use the expression *real-time* for referring to algorithms that operate with a fixed forward-lag of K_F. We will discuss specific choices of K_F and K_B and their implications in Section 3.

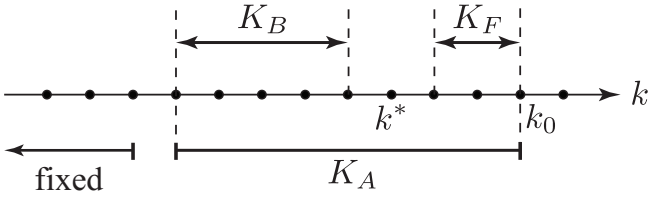

Figure 2. The parameters involved in state-space fixed-lag smoothing.

Suppose we have a window of length K_A where the instances are indexed by $k = 1, 2, \ldots, K_A$. Inspired by (Akram et al., 2016), we assume a linear state-space model on the logit-probability of attending to speaker 1. We define the binary random variable $n_k = 1$ when speaker 1 is attended and $n_k = 2$ when speaker 2 is attended, at instance k. The goal is to obtain estimates of $p_k := \mathrm{P}\,(n_k = 1)$ together with its confidence intervals for $1 \leq k \leq K_A$. The state dynamics are given by:

$$\begin{cases} p_k = \mathrm{P}\,(n_k = 1) = 1 - \mathrm{P}\,(n_k = 2) = \frac{1}{1+\exp(-z_k)} \\ z_k = c_0 z_{k-1} + w_k \\ w_k \sim \mathcal{N}(0, \eta_k) \\ \eta_k \sim \text{Inverse-Gamma}\,(a_0, b_0) \end{cases} \tag{2}$$

The dynamics of the main latent variable z_k are controlled by its transition scale c_0 and state variance η_k. The hyperparameter $0 \leq c_0 \leq 1$ ensures the stability of the updates for z_k. The state variance η_k is modeled using an Inverse-Gamma conjugate prior with hyper-parameters a_0 and b_0. The log-prior of the Inverse-Gamma density takes the form $\ln \mathrm{P}\,(\eta_k) = -(a_0 + 1) \ln \eta_k - \frac{b_0}{\eta_k} + C$ for $\eta_k > 0$, where C is a normalization constant. By choosing a_0 greater than and sufficiently close to 2, the variance of the Inverse-Gamma distribution takes large values and therefore can serve as a non-informative conjugate prior. Considering the fact that we do not expect the attentional state to have high fluctuations within a small window of time, we can further tune the hyperparameters a_0 and b_0 for the prior to promote smaller values of η_k's. This way, we can avoid large consecutive fluctuations of the z_k's, and consequently the p_k's.

Next, we develop an observation model relating the state dynamics of Eq. (2) to the observations $m_k^{(1)}$ and $m_k^{(2)}$ for $k = 1, 2, \ldots, K_A$. To this end, we use the latent variable n_k as the link between the states and observations:

$$\begin{cases} \begin{cases} m_k^{(i)} \;\middle|\; n_k = i \sim \text{Log-Normal}\left(\rho^{(a)}, \mu^{(a)}\right) \\ m_k^{(i)} \;\middle|\; n_k \neq i \sim \text{Log-Normal}\left(\rho^{(u)}, \mu^{(u)}\right) \end{cases}, \quad i = 1, 2 \\ \rho^{(a)} \sim \text{Gamma}\left(\alpha_0^{(a)}, \beta_0^{(a)}\right), \quad \mu^{(a)} \;\middle|\; \rho^{(a)} \sim \mathcal{N}\left(\mu_0^{(a)}, \rho^{(a)}\right) \\ \rho^{(u)} \sim \text{Gamma}\left(\alpha_0^{(u)}, \beta_0^{(u)}\right), \quad \mu^{(u)} \;\middle|\; \rho^{(u)} \sim \mathcal{N}\left(\mu_0^{(u)}, \rho^{(u)}\right) \end{cases} \tag{3}$$

When speaker $i = 1, 2$ is attended to, we use a Log-Normal distribution on $m_k^{(i)}$'s, with log-density given by $\ln \mathrm{P}\left(m_k^{(i)} \mid n_k = i\right) = -\ln m_k^{(i)} + \frac{1}{2}\ln \rho^{(a)} - \frac{\rho^{(a)}}{2}\left(\ln m_k^{(i)} - \mu^{(a)}\right)^2 + C^{(i)}$, where $\mu^{(a)} \in \mathbb{R}$, $\rho^{(a)} \in \mathbb{R}_{>0}$, and $C^{(i)}$ is a normalization constant, for $i = 1, 2$, and $k = 1, 2, \ldots, K_A$. Similarly, when speaker $i = 1, 2$ is *not* attended to, we use a Log-Normal distribution on $m_k^{(i)}$ with parameters $\rho^{(u)}$ and $\mu^{(u)}$. As mentioned before, choosing an appropriate attention marker results in a statistical separation between $m_k^{(1)}$ and $m_k^{(2)}$, if only one speaker is attended. The Log-Normal distribution is a unimodal distribution on $\mathbb{R}_{>0}$ which lets us capture this concentration in the values of $m_k^{(i)}$'s. In contrast to (Akram et al., 2016), this distribution also leads to closed form update rules, which significantly reduces computational costs. We have also imposed conjugate priors on the joint distribution of (ρ, μ)'s, which factorizes as $\ln \mathrm{P}(\rho, \mu) = \ln \mathrm{P}(\rho) + \ln \mathrm{P}(\mu \mid \rho)$. The hyperparameters α_0, β_0, and μ_0 serve to tune the attended and the unattended Log-Normal distributions to create separation between the attended and unattended cases. These hyperparameters can be determined based on the mean and variance information of $m_k^{(i)}$'s in a supervised manner, in which the attended speaker labels are known, while enforcing large enough variances for the priors not to be too restrictive in estimating the Log-Normal distribution parameters. As will be discussed in our simulation and real-data analysis, this tuning step can be performed using a minimal amount of labeled data, which is significantly less than those required for reliable pre-estimation of encoder/decoder coefficients in existing approaches.

The parameters of the state-space model are therefore $\mathbf{\Omega} = \left\{z_{1:K_A}, \eta_{1:K_A}, \rho^{(a)}, \mu^{(a)}, \rho^{(u)}, \mu^{(u)}\right\}$, which have to be inferred from $m_{1:K_A}^{(1)}$ and $m_{1:K_A}^{(2)}$. As mentioned before, our goal in the fixed-lag smoothing approach is to estimate $p_{k^*} = 1/\left(1 + \exp\left(-z_{k^*}\right)\right)$ as well as its confidence intervals in each window, where $k^* = K_A - K_F$. However, in order to do so in our model, we perform the inference step over all the parameters in $\mathbf{\Omega}$ and output the estimate of $z_{k^*} \in \mathbf{\Omega}$ and its confidence intervals. The calculation of confidence intervals is discussed in detail at the end of Section 2 of the Supplementary Material. In short, the density of each z_k given the set of observed attention markers, estimated variances, and estimated Log-Normal distribution parameters is recursively approximated by a Gaussian density. Then, the mean of this Gaussian approximation is reported as the estimated z_k and its confidence intervals are determined based on the corresponding variance. The estimated $\mathbf{\Omega}$ would then serve as the initialization for parameter estimation in the next window. The parameters in $\mathbf{\Omega}$ can be inferred through two nested EM algorithms as in (Akram et al., 2016). In Section 2 of the Supplementary Material, we have given a detailed derivation of the EM framework and update rules in the real-time setting, as well as solutions to further reduce the computational costs thereof. From here on, we refer to the output of the introduced framework, which

operates with the discussed built-in delay, as the *real-time (state-space) estimator*. In Section 3.1, we compare the performance of the real-time estimator against that of the *batch-mode (state-space) estimator*. We define the batch-mode estimator as applying the state-space model in Eq. (2) and (3) on all the computed attention markers in a trial *at once*, i.e., $K_A = K$, rather than in a fixed-lag sliding window fashion. In other words, the batch-mode estimator observes all the attention marker samples in a trial, i.e., $m_k^{(i)}$ for $i = 1, 2$ and $k = 1, \ldots, K$, and then infers the attention probabilities. In this sense, it is similar to the state-space estimator used in (Akram et al., 2016). The batch-mode estimator provides a robust estimate of the attentional state at any instance by having access to all the future and past attention markers. Thus, it can serve as a performance benchmark for tuning the fixed-lag sliding window hyperparameters in the real-time estimator. We will further discuss this point in Section 3.1.4.

Remark 5. The state-space models given in Eqs. 2 and 3 have two major differences with the one used in (Akram et al., 2016). First, in (Akram et al., 2016), the distribution over the correlative measure for the *unattended* speaker is assumed to be uniform. However, this assumption may not hold for other attention markers in general. For instance, the M100 magnitude of the TRF estimated from MEG data is a positive random variable, which is concentrated on higher values for the attended speaker compared to the unattended speaker. In order to address this issue, we consider a parametric distribution in Eq. (3) over the attention marker corresponding to the unattended speaker and infer its parameters from the data. If this distribution is indeed uniform and non-informative, the variance of the unattended distribution, which is estimated from the data, would be large enough to capture the flatness of the distribution. Second, the parametrization of the observations using Log-Normal densities and their corresponding priors factorized using Gamma and Gaussian priors, admits fast and closed-form update equations in the real-time setting. As we have shown in Section 2 of the Supplementary Material, these models also have the advantage of incorporating low-complexity updates by simplifying the EM procedure. In addition, the Log-Normal distribution as a generic unimodal distribution allows us to model a larger class of attention markers.

Remark 6. As mentioned in Section 1, one limitation of existing approaches based on reverse-correlation is that their decoding accuracy drops significantly when operating at high temporal resolutions. The major source for this performance deterioration is the stochastic fluctuations and uncertainties in correlation values when computed over small windows of length $\sim 1\,\text{s}$. Therefore, when enough training data is available for reliable pre-estimation of decoders/encoders, our real-time state-space module can be added as a complementary final step to the foregoing approaches in order to correct for the stochastic fluctuations in the calculated correlation values.

2.4 EEG Recording and Experiment Specifications

64-channel EEG was recorded using the actiCHamp system (Brain Vision LLC, Morrisville, NC, US) and active EEG electrodes with Cz channel being the reference. The data was digitized at a 10 kHz sampling frequency. Insert earphones ER-2 (Etymotic Research Inc., Elk Grove Village, IL, US) were used to deliver sound to the subjects while sitting in a sound-attenuated booth. The earphones were driven by the clinical audiometer Piano (Inventis SRL, Padova, Italy), and the volume was adjusted for every subject's right and left ears separately until the loudness in both ears was matched at a comfortably loud listening level. Three normal-hearing adults participated in the study. The mean age of subjects was 49.5 years with the standard deviation of 7.18 years. The study included a constant-attention experiment, where the subjects were asked to sit in front of a computer screen and restrict motion while any audio was playing. The data used in this paper corresponds to 3 subjects, 24 trials each.

The stimulus set contained eight story segments, each approximately ten minutes long. Four segments were narrated by male speaker 1 (M1) and the other four by male speaker 2 (M2). The stimuli were presented to the subjects in a dichotic fashion, where the stories read by M1 were played in the left ear, and stories read by M2 were played in the right ear for all the subjects. Each subject listened to twenty four trials of the dichotic stimulus. Each trial had a duration of approximately one minute, and for each subject, no storyline was repeated in more than one trial. During each trial, the participants were instructed to look at an arrow at the center of the screen, which determined whether to attend to the right-ear story or to the left one. The arrow remained fixed for the duration of each trial, making it a constant-attention experiment. At the end of each trial, two multiple choice semantic questions about the attended story were displayed on the screen to keep the subjects alert. The responses of the subjects as well as their reaction time were recorded as a behavioral measure of the subjects' level of attention, and above eighty percent of the questions were answered correctly by each subject. Breaks and snacks were given between stories if requested. All the audio recordings, corresponding questions, and transcripts were obtained from a collection of stories recorded at Hafter Auditory Perception Lab at UC Berkeley.

2.5 MEG Recording and Experiment Specifications

MEG signals were recorded with a sampling rate of $1\,\text{kHz}$ using a 160-channel whole-head system (Kanazawa Institute of Technology, Kanazawa, Japan) in a dimly lit magnetically shielded room (Vacuumschmelze GmbH & Co. KG, Hanau, Germany). Detection coils were arranged in a uniform array on a helmet-shaped surface on the bottom of the dewar with $25\,\text{mm}$ between the centers of two adjacent $15.5\,\text{mm}$ diameter coils. The sensors are first-order axial gradiometers with a baseline of $50\,\text{mm}$, resulting in field sensitivities of $5\,\frac{\text{fT}}{\sqrt{\text{Hz}}}$ or better in the white noise region.

The two speech signals were presented at $65\,\text{dB}$ SPL using the software package Presentation (Neurobehavioral Systems Inc., Berkeley, CA, US). The stimuli were delivered to the subjects' ears with $50\,\Omega$ sound tubing (E-A-RTONE 3A; Etymotic Research), attached to E-A-RLINK foam plugs inserted into the ear canal. Also, the whole acoustic delivery system was equalized to give an approximately flat transfer function from $40\,\text{Hz}$ to $3000\,\text{Hz}$. A $200\,\text{Hz}$ low-pass filter and a notch filter at $60\,\text{Hz}$ were applied to the magnetic signal in an online fashion for noise removal. Three of the 160 channels are magnetometers separated from the others and used as reference channels. Finally, to quantify the head movement, five electromagnetic coils were used to measure each subject's head position inside the MEG machine once before and once after the experiment.

Nine normal-hearing, right-handed young adults (ages between 20 and 31) participated in this study. The study includes two sets of experiments: the constant-attention experiment and the attention-switch experiment, in each of which six subjects participated. Three subjects took part in both of the experiments. The experimental procedure were approved by the University of Maryland Institutional Review Board (IRB), and written informed consent was obtained from each subject before the experiment.

The stimuli included four non-overlapping segments from the book *A Child's History of England* by Charles Dickens. Two of the segments were narrated by a man and the other two by a woman. Three different mixtures, each $60\,\text{s}$ long, were generated and used in the experiments to prevent reduction in the attentional focus of the subjects. Each mixture included a segment narrated by the male speaker and one narrated the the female speaker. In all trials, the stimuli were delivered diotically to both ears using tube phones inserted into the ear canals at a level of approximately $65\,\text{dB}$ SPL. The constant-attention experiment consisted of two conditions: 1) attending to the male speaker in the first mixture, 2) attending to the female speaker in the second mixture. In the attention-switch experiment, subjects were instructed

to focus on the female speaker in the first 28 s of the trial, switch their attention to the male speaker after hearing a 2 s pause (28th to 30th seconds), and maintain their focus on the latter speaker through the end of the trial. Each mixture was repeated three times in the experiments, resulting in six trials per speaker for the constant-attention experiment and three trials per speaker for the attention-switch experiment. After the presentation of each mixture, subjects answered comprehensive questions related to the segment they were instructed to focused on, as a way to keep them motivated to attend to the target speaker. Eighty percent of the questions were answered correctly on average. Furthermore, a preliminary experiment for each of the nine participating subjects was performed prior to the main experiments. In this study, the subjects listened to a single speech stream, first segment in the stimuli set narrated by the male speaker, for three trials each 60 s long. The MEG recordings in the pilot study were used to calculate the subject-specific linear combination of MEG channels which forms the auditory component of the response, as will be explained next. Note that for each subject, all the recordings were performed in a single session resulting in a minimal change of the subject's head position with respect to the MEG sensors.

3 RESULTS

In this section, we apply our real-time attention decoding framework to synthetic data as well as M/EEG recordings. Subsection 3.1 includes the simulation results, and subsections 3.2 and 3.3 demonstrate the results for the analysis of EEG and MEG recordings, respectively.

3.1 Simulations

In order to validate our proposed framework, we perform two sets of simulations. The first simulation pertains to our EEG analysis and employs a decoding model, which we describe below in full detail. The second simulation, for our MEG analysis using an encoding model, is deferred to the Supplementary Material Section 4, in the interest of space.

3.1.1 Simulation Settings

In order to simulate EEG data under a dual-speaker condition, we use the following generative model:

$$e_t = w_t^{(1)}\left(s_t^{(1)} * h_t\right) + w_t^{(2)}\left(s_t^{(2)} * h_t\right) + \mu + u_t \tag{4}$$

where $s_t^{(1)}$ and $s_t^{(2)}$ are respectively the speech envelopes of speakers 1 and 2 at time t; the output e_t is the simulated neural response, which denotes an auditory component of the EEG or the EEG response at a given channel at time t for $t = 1, 2, \ldots, T$. Motivated by the analysis of LTI systems, h_t can be considered as the impulse response of the neural process resulting in e_t, and $*$ represents the convolution operator; the scalar μ is an unknown constant mean, and u_t denotes a zero-mean i.i.d Gaussian noise. The weight functions $w_t^{(1)}$ and $w_t^{(2)}$ are signals modulated by the attentional state which determine the contributions of speakers 1 and 2 to e_t, respectively. In order to simulate the attention modulation effect, we assume that when speaker 1 (resp. 2) is attended to at time t, we have $w_t^{(1)} > w_t^{(2)}$ (resp. $w_t^{(1)} < w_t^{(2)}$).

We have chosen two 60 s-long speech segments from those used in the MEG experiment (See section 2.5) and calculated $s_t^{(1)}$ and $s_t^{(2)}$ as their envelopes for a sampling rate of $f_s = 200$ Hz. Also, we have set $\mu = 0.02$ and $u_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 2.5 \times 10^{-5})$ in Eq. (4). Fig. 3-A shows the location and amplitude of the lag components in the impulse response, which is then smoothed using a Gaussian kernel with standard deviation of 10 ms to result in the final impulse response h_t, shown in Fig. 3–B. The significant components

of h_t are chosen at 50 ms and 100 ms lags, with a few smaller components at higher latencies (Akram et al., 2016). It is noteworthy that existing results (Ding and Simon, 2012a; Power et al., 2012; Akram et al., 2017) suggest that this impulse response (i.e., the TRF) is not the same for the attended and unattended speakers, as discussed in Section 2.2. However, we have considered the same h_t for both speakers in this simulation for simplicity, given that our focus here is to model the stronger presence of the attended speaker in the neural response in terms of the extracted attention markers. In Section 4 of the Supplementary Material, we indeed use an encoding model consisting of different and attention-modulated TRFs for the two speakers. The weight signals $w_t^{(1)}$ and $w_t^{(2)}$ in Eq. (4) are chosen to favor speaker 1 in the $[0\,\text{s}, 30\,\text{s})$ interval and speaker 2 in the $(30\,\text{s}, 60\,\text{s}]$ interval.

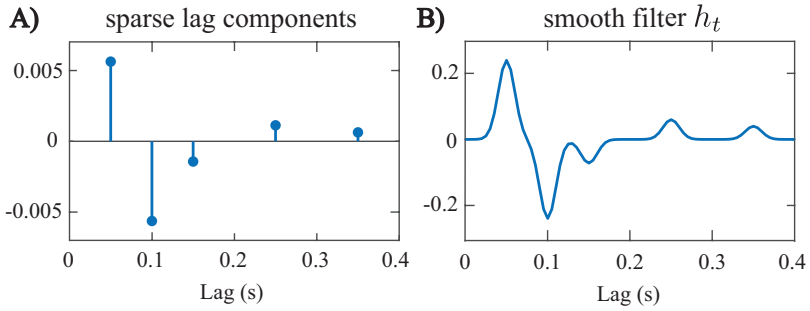

Figure 3. Impulse response h_t used in Eq. (4). A) sparse lag components, B) the smooth impulse response.

3.1.2 Parameter Selection

We aim at estimating decoders in this simulation, which linearly map e_t and its lags to $s_t^{(1)}$ and $s_t^{(2)}$. To estimate the decoders, we have considered consecutive non-overlapping windows of length 0.25 s resulting in $K = 240$ windows of length $W = 50$ samples. Also, we have chosen $\gamma = 0.001$, through cross-validation, and $\lambda = 0.95$ in estimating the decoding coefficients, which results in an *effective* data length of 5 s for decoder estimation. The forward lags of the neural response have been limited to a 0.4 s window, i.e., $L_d = 80$ samples. Given that the decoder corresponds to the inverse of a smooth kernel h_t, it may not have the same smoothness properties of h_t. Hence, we do not employ a smooth basis for decoder estimation. We have used the FASTA package (Goldstein et al., 2014) with Nesterov's acceleration method to implement the forward-backward splitting algorithm for encoder/decoder estimation. As for the state-space model estimators, we have considered 20 (inner and outer) EM iterations for the batch-mode estimators, while for the real-time estimators, we use 1 inner EM iteration and 20 outer EM iterations (See Section 2 of the Supplementary Material for more details).

There are three criteria for choosing the fixed-lag smoothing parameters: First, how close to the true real-time analysis the system operates is determined by K_F. Second, the computational cost of the system is determined by K_A. Third, how close the output of the system is to that of the batch-mode estimator is determined by both K_F and K_A. These three criteria form a tradeoff in tuning the parameters K_A and K_F. Specific choices of these parameters are given in the next subsection.

For tuning the hyperparameters of the priors on the attended and unattended distributions, we have used a separate 15 s sample trial generated from the same simulation model in Eq. (4) for each of the three cases. The parameters $\left(\alpha_0^{(a)}, \alpha_0^{(u)}, \beta_0^{(a)}, \beta_0^{(u)}, \mu_0^{(a)}, \mu_0^{(u)}\right)$ have been chosen by fitting the Log-Normal distributions to the attention marker outputs from the sample trials in a supervised manner (with known attentional state). The variance of the Gamma priors $\frac{\alpha_0^{(a)}}{{\beta_0^{(a)}}^2}$ and $\frac{\alpha_0^{(u)}}{{\beta_0^{(u)}}^2}$ have been chosen large enough such

that the priors are non-informative. This step can be thought of as the initialization of the algorithms prior to data analysis. For the Inverse-Gamma prior on the state-space variances, we have chosen $a_0 = 2.008$ and $b_0 = 0.2016$, resulting in a mean of 0.2 and a variance of 5. This prior favors small values of η_k's to ensure that the state estimates are immune to large fluctuations of the attention markers, while the large variance (compared to the mean) results in a non-informative prior for smaller values of η_k's.

3.1.3 Estimation Results

Fig. 4 shows the results of our estimation framework for a correlation-based attention marker. Row A in Fig. 4 shows three cases considered for modulating the weights $w_t^{(1)}$ and $w_t^{(2)}$, where the weights are contaminated with Gaussian noise $\mathcal{N}(0, 4\times 10^{-4})$ to model extra uncertainties in determining the contribution of each speech to the neural response, arising from irrelevant or background neural processes. In order to probe the transition delay of the state-space estimates due to abrupt changes in the attentional state, the two weight vectors undergo step-like transition at 30 s. Cases 1, 2, and 3 exhibit increasing levels of difficulty in discriminating the contributions of the two speakers to the neural response. Rows B and C in Fig. 4 respectively show the decoder estimates for speakers 1 and 2. As expected, the significant components of the decoders around 50 ms, 100 ms, and 150 ms lags, are modulated by the attentional state, and the modulation effect weakens as we move from Case 1 to 3. In Case 1, these components are less significant overall for the decoder estimates of speaker 2 in the $[0\,\text{s}, 30\,\text{s}]$ time interval and become larger as the attention switches to speaker 2 during the rest of the trial (red boxes in row C of Case 1). On the other hand, in Case 3, the magnitude of said components do not change notably across the 30 s mark. The TRF h_t in the forward generative model of Eq. (4) is an FIR filter with significant components at lags which are multiples of 0.05 s (See Fig. 3-B). Therefore, the decoder estimates in Fig. 4 correspond to truncated IIR filters, which form approximate inverse filters of the TRF. Therefore, it is expected that they comprise significant components at lags which are multiples of 0.05 s as well, but decay exponentially fast.

We have considered two different attention markers for this simulation. Row D in Fig. 4 displays the output of a correlation-based attention marker for speakers 1 and 2, which is calculated as $m_k^{(i)} = \left|\text{corr}\left(\mathbf{y}_k^{(i)}, \mathbf{X}_k\hat{\boldsymbol{\theta}}_k^{(i)}\right)\right|$ for $i = 1, 2$ and $k = 1, 2, \ldots, K$. As discussed in subsection 2.2, this attention marker is a measure of how well a decoder can reconstruct its target envelope. As observed in row D of Fig. 4, the attention marker is a highly variable surrogate of the attentional state at each instance, i.e., *on average* the attention marker output for speaker 1 is higher then that of speaker 2 in the $[0\,\text{s}, 30\,\text{s})$ interval and vice versa in the $(30\,\text{s}, 60\,\text{s}]$ interval. The reliability of the attention marker significantly degrades going from Case 1 to 3. This highlights the need for state-space modeling and estimation in order to optimally exploit the attention marker.

Rows E and F in Fig. 4 respectively show the batch-mode and real-time estimator outputs as the inferred attentional state probabilities $p_k = \text{P}\left(n_k = 1\right)$ for $k = 1, \ldots, K$, for the correlation-based attention marker, where colored hulls indicate 90% confidence intervals. Row F in Fig. 4 corresponds to the fixed-lag smoother, using a window of length $15\,\text{s}$ ($K_A = \lfloor 15 f_s/W \rfloor$), and a forward-lag of $1.5\,\text{s}$ ($K_F = \lfloor 1.5 f_s/W \rfloor$). By accounting for the lag in the decoder (L_d), the built-in delay in estimating the attentional state is $1.9\,\text{s}$. Note that all the relevant figures showing the outputs of the *real-time* estimator are calibrated with respect to the built-in delay for the sake of illustration. Thus, these figures must be interpreted as non-causal when $K_F > 0$, since the estimated attentional state at each time depends on the future K_F samples of the attention marker. Recall that in the batch-mode estimator, all of the attention marker outputs across the trial are available to the state-space estimator, as opposed to the fixed-lag real-time estimator which has access to a limited number of the attention markers. Therefore, the output

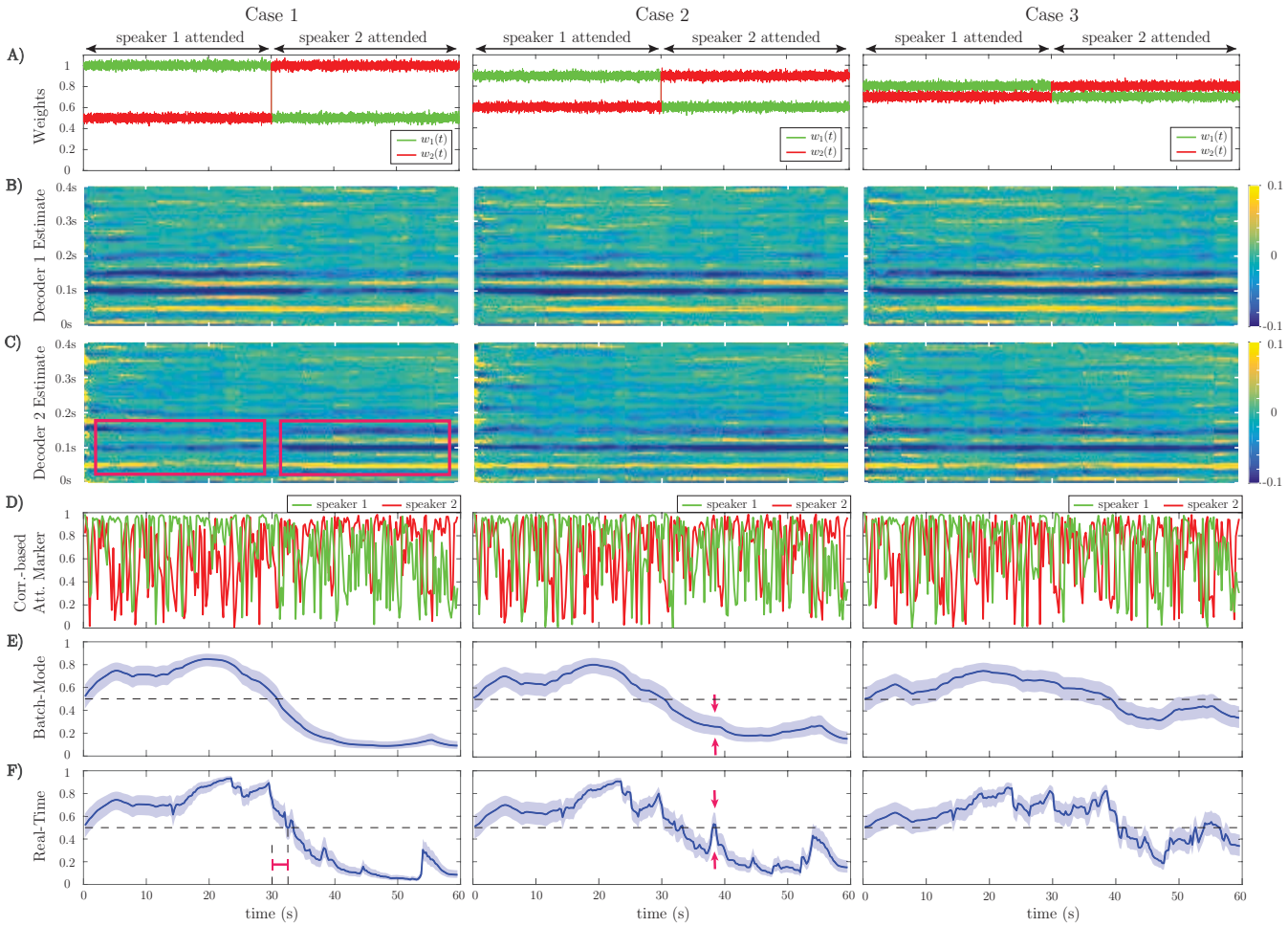

Figure 4. Estimation results of application to simulated EEG data for the correlation-based attention marker: A) Input weights $w_t^{(1)}$ and $w_t^{(2)}$ in Eq. (4), which determine the relative effect of the two speeches on the neural response. Based on our generative model, the attention is on speaker 1 for the first half of each trial and on speaker 2 for the second half. Case 1 corresponds to a scenario where the effects of the attended and unattended speeches in the neural response are well-separated. This separation decreases as we move from Case 1 to Case 3. B) Estimated decoder for speaker 1. C) Estimated decoder for speaker 2. In Case 1, the significant components of the estimated decoders near the 50 ms, 100 ms, and 150 ms lags are notably modulated by the attentional state as highlighted by the red boxes. This effect weakens in Case 2 and visually disappears in Case 3. D) Output of the correlation-based attention marker for each speaker. E) Output of the batch-mode state-space estimator for the correlation-based attention marker as the estimated probability of attending to speaker 1. F) Output of the real-time state-space estimator, i.e., fixed-lag smoother, for the correlation-based attention marker as the estimated probability of attending to speaker 1. The real-time estimator is not as robust as the batch-mode estimator to the stochastic fluctuations of the attention marker in row D and is more prone to misclassifications. The red arrows in rows E and F of Case 2 show that the batch-mode estimator correctly classifies the instance as attending to speaker 2, while the real-time estimator is unable to determine the attentional state.

of the batch-mode estimator (Row E) is a more robust measure of the instantaneous attentional state as compared to the real-time estimator (Row F), since it is less sensitive to the stochastic fluctuations of the attention markers in row D. For example, in the instance marked by the red arrows in rows E and F of Case 2 in Fig. 4, the batch-mode estimator classifies the instance correctly as attending to speaker 2, while the real-time estimator cannot make an informed decision since $p_k = 0.5$ falls within the 90% confidence interval of the estimate at this instance. However, the real-time estimator exhibits performance closely

matching that of the batch-mode estimator for most instances, while operating in real-time with limited data access and significantly lower computational complexity. Comparing the state-space estimators with the raw attention markers in Fig. 4-D, we observe the smoothing effect of the state-space model which makes its output robust to the stochastic fluctuations in the attention marker at high temporal resolution. Section 3 of the Supplementary Material includes a comparison of this smoothing effect with that of a typical Gaussian smoothing kernel applied directly to the attention markers.

Row A in Fig. 5 exhibits the output of another attention marker computed as the ℓ_1-norm of the decoder given by $m_k^{(i)} := \left\| \hat{\boldsymbol{\theta}}_k^{(i)} \right\|_1$ for $i = 1, 2$ and $k = 1, 2, \ldots, K$, where the first element of $\hat{\boldsymbol{\theta}}_k^{(i)} \in \mathbb{R}^{L_d+2}$ (the intercept parameter) is discarded in computing the ℓ_1-norm. This attention marker captures the effect of the significant peaks in the decoder. The rationale behind using the ℓ_1-norm based attention marker is the following: in the extreme case that the neural response is solely driven by the attended speech, we expect the unattended decoder coefficients to be small in magnitude and randomly distributed across the time lags. The attended decoder, however, is expected to have a sparse set of informative and significant components corresponding to the specific latencies involved in auditory processing. Thus, the ℓ_1-norm serves to distinguish between these two cases by capturing such significant components. Rows B and C in Fig. 5 show the batch-mode and real-time estimates of the attentional state probabilities for the ℓ_1-based attention marker, respectively, where colored hulls indicate 90% confidence intervals. Consistent with the results of the correlation-based attention marker (Rows E and F in Fig. 4), the real-time estimator exhibits performance close to that of the batch-mode estimator. Comparing Figs. 4 and 5 reveals the dependence of the attentional state estimation performance on the choice of the attention marker: while the correlation-based attention marker is more widely used, the ℓ_1-based attention marker provides smoother estimates of the attention probabilities, and can be used as an alternative to the correlation-based attention marker. Overall, this simulation illustrates that if the attended stimulus has a stronger presence in the

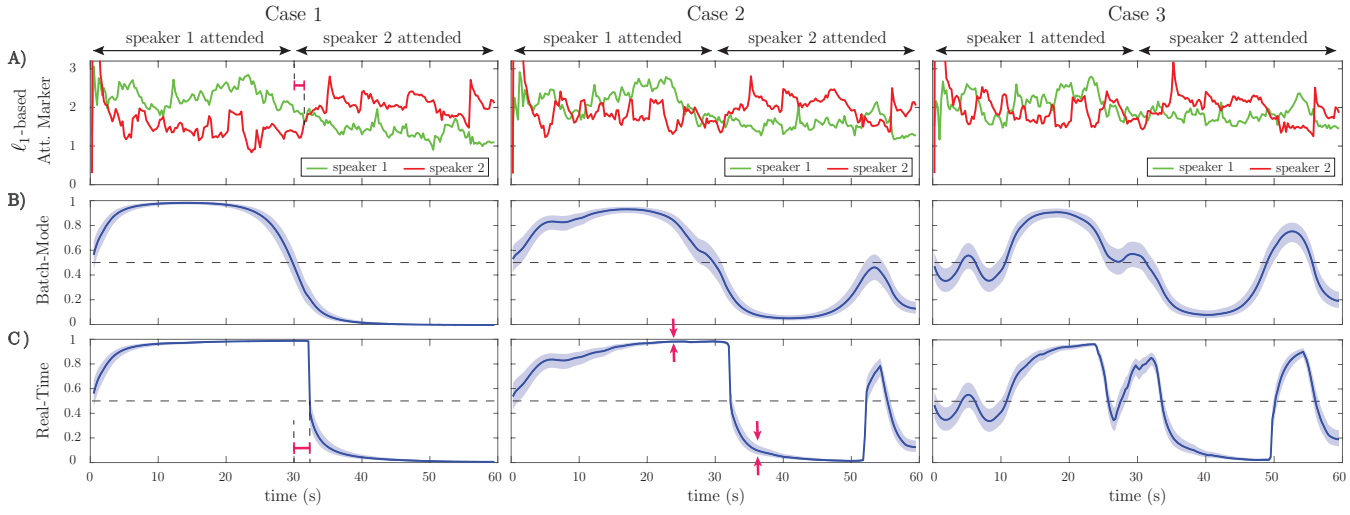

Figure 5. Estimation results of application to simulated EEG data for the ℓ_1-based attention marker: A) Output of the ℓ_1-based attention marker for each speaker, corresponding to the three cases in Figure 4. B) Output of the batch-mode state-space estimator for the ℓ_1-based attention marker as the estimated probability of attending to speaker 1. C) Output of the real-time state-space estimator for the ℓ_1-based attention marker as the estimated probability of attending to speaker 1. Similar to the preceding correlation-based attention marker, the classification performance degrades when moving from Case 1 (strong attention modulation) to Case 3 (weak attention modulation).

neural response than the unattended one, both the correlation-based and ℓ_1-based attention markers can be attention modulated and can therefore potentially be used in real M/EEG analysis.

3.1.4 Discussion and Further Analysis

Going from Case 1 to Case 3 in Fig. 4 and Fig. 5, we observe that the performance of all estimators degrades, causing a drop in the classification accuracy and confidence. This performance degradation is due to the declining power of the attention markers in separating the contributions of the attended and unattended speakers. However, comparing the outputs of the real-time and batch-mode estimators with their corresponding attention marker outputs in row D of Fig. 4 and row A of Fig. 5, highlights the role of the state-space model in suppressing the stochastic fluctuations of the attention markers and thereby providing a robust and smooth measure of the attentional state.

In response to abrupt step-like changes in the attentional state, we define the *transition* delay as the time it takes for the output of the real-time estimator to reach the $p_k = 0.5$ level, which marks the point at which the classification label of the attended speaker changes. We calculate the transition delay after calibrating for the built-in delay, for all the real-time estimator outputs. Thus, the overall delay of the system in detecting abrupt attentional state changes is equal to the sum of the built-in and transition delays. The red intervals in Case 1 of row F in Fig. 4 and row C of Fig. 5 mark the transition delay of the real-time estimator corresponding to the correlation-based and ℓ_1-based attention markers, respectively. From the deflection point at $30\,\mathrm{s}$, this delay is given by $\sim 2.3\,\mathrm{s}$. The transition delay is due to the forgetting factor mechanism and the smoothing effect of the state-space estimation given the backward- and forward-lags, which have been set in place to increase the robustness of the decoding framework to stochastic fluctuations of the extracted attention markers. As a result, such classification delays in response to a sudden attention switches are expected by design. Specifically, the sole contribution of the forgetting factor mechanism to this delay can be observed as the red interval in Case 1 of row A in Fig. 5, which precedes the application of the state-space estimation.

Comparing the batch-mode and the real-time estimators in Fig. 4 and Fig. 5, we observe that the real-time estimators closely follow the output of the batch-mode estimators, while having access to data in an online fashion. A significant deviation between the batch-mode and real-time performance is observed in rows B and C (Cases 1 and 2) of Fig. 5 in the form of sharp drops in the real-time estimates of the attentional state probability. Given that the real-time estimator has only access to the attention marker within K_F samples in the future, the confidence intervals significantly narrow down within the first half of the trial, as all the past and near-future observations are consistent with attention to speaker 1. However, shortly after the $30\,\mathrm{s}$ mark the estimator detects the change and the confidence bounds widen accordingly (see red arrows in row C of Case 2 in Fig. 5).

In order to further quantify the performance gap between the batch-mode and real-time estimators, we define their relative Mean Squared Error (MSE) as:

$$\mathrm{MSE} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{1+\exp\left(-\hat{z}_k^{(B)}\right)} - \frac{1}{1+\exp\left(-\hat{z}_k^{(R)}\right)}\right)^2 \tag{5}$$

where $\hat{z}_{1:K}^{(R)}$ and $\hat{z}_{1:K}^{(B)}$ denote the real-time and batch-mode state estimates over a given trial, respectively. We have considered the logistic transformation of $\hat{z}_{1:K}^{(B)}$ and $\hat{z}_{1:K}^{(R)}$, which gives the probability of attending to speaker 1. The rationale behind this MSE metric is to measure the performance and robustness of the

real-time estimator with respect to the batch-mode estimator, since they both operate on the same computed attention markers, but in different algorithmic fashions.

Figure 6 shows the effect of varying the forward-lag K_F from $0\,\text{s}$ (i.e., fully real-time) to $5\,\text{s}$ with $0.5\,\text{s}$ increments for the two attention markers in Case 2 of Fig. 4 and Fig. 5, as an example. All of the other parameters in the simulation have been fixed as before. The left panels in Fig. 6 show the MSE for different values of K_F in the real-time setting. As expected, for both attention markers, the MSE decreases as the forward-lag increases. The right panels in Fig. 6 display the incremental MSE defined as the change in MSE when K_F is increased by $0.5\,\text{s}$ at each value, starting from $K_F = 0$. The incremental MSE is basically the discrete derivative of the displayed MSE plots and shows the amount of relative performance boost between two consecutive values of K_F, if we allow for a larger built-in delay. Notice that even a $0.5\,\text{s}$ forward-lag significantly decreases the MSE from $K_F = 0$. The subsequent improvements of the MSE diminish as K_F is increased further. Our choice of K_F corresponding to $1.5\,\text{s}$ in the foregoing analysis was made to maintain a reasonable tradeoff between the MSE improvement and the built-in delay in real-time operation. In summary, Fig. 6 shows that having larger forward-lags can make our estimates more robust but it creates a larger built-in delay. Whether higher levels of delay are tolerable or not depends on the particular attention decoding application.

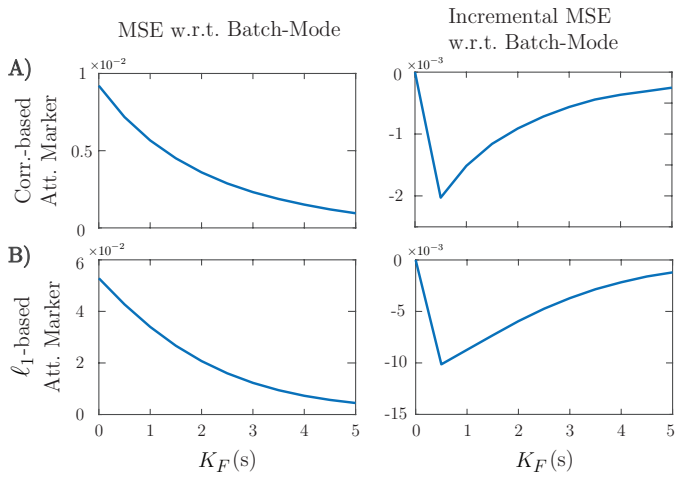

Figure 6. Effect of the forward-lag K_F on the MSE for the two attention markers in case 2 of Fig. 4 and Fig. 5. A) Correlation-based attention marker, B) ℓ_1-based attention marker. As the forward-lag increases, the MSE decreases, and the output of the real-time estimator becomes more similar to that of the batch-mode. This results in more robustness for the real-time estimator at the expense of more built-in delay in decoding the attentional state. The right panels show that the incremental improvement to the MSE decreases as K_F increases.

Finally, Fig. 7 shows the estimated attention probabilities and their 90% confidence intervals for the correlation-based attention marker in Case 2 of Fig. 4, as an example of the output of the state-space estimator. The three curves correspond to the extreme values of K_F in Fig. 6 corresponding to $0\,\text{s}$ (blue) and $5\,\text{s}$ (red) forward-lags, and the batch-mode estimate (green). All the other parameters have been fixed as described above. The fixed-lag smoothing approach with K_F of $5\,\text{s}$ is as robust as the batch-mode estimate. The fully real-time estimate with K_F of $0\,\text{s}$ follows the same trend as the other two. However, it is susceptible to the stochastic fluctuations of the attention marker, which may lead to misclassifications (see the red arrows in Fig. 7). The red interval in Fig. 7 displays the difference between the transition delays corresponding to the forward-lag of $0\,\text{s}$ and $5\,\text{s}$. Although the built-in attention decoding delay of a

5 s forward-lag is more than that of 0 s by 5 s, the transition delay corresponding to the former is smaller due to observing the future attention marker samples up to 5 s. Therefore, the parameter K_F also provides a tradeoff in the overall delay of the framework in detecting abrupt attention switches, which equals the transition delay plus the built-in delay. The choice of 1.5 s for the forward-lag in our analysis was also aimed to minimize this overall delay.

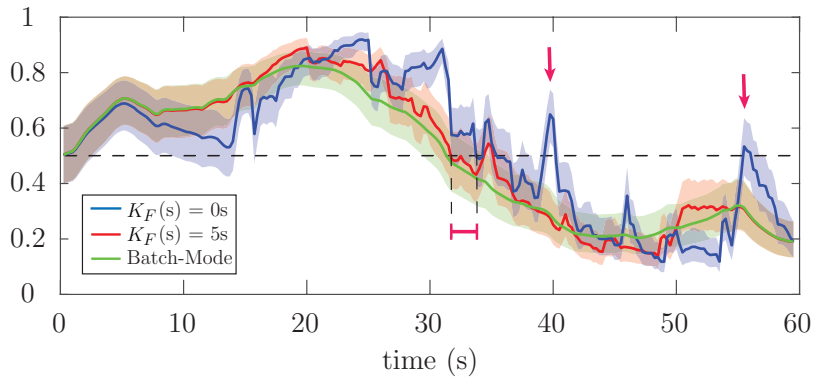

Figure 7. Estimated attention probabilities together with their 90% confidence intervals for the correlation-based attention marker in Case 2 of Fig. 4. The blue, red and green curves correspond to K_F of 0 s, K_F of 5 s, and batch-mode estimation, respectively. The estimator for K_F of 5 s is nearly as robust as the batch-mode. However, the fully real-time estimator with K_F of 0 s is sensitive to the stochastic fluctuations of the attention markers, which results in the misclassification of the attentional state at the instances marked by red arrows.

3.2 Application to EEG

In this subsection, we apply our real-time attention decoding framework to EEG recordings in a dual-speaker environment. Details of the experimental procedures are given in Section 2.4.

3.2.1 Preprocessing and Parameter Selection

Both the EEG data and the speech envelopes were downsampled to $f_s = 64$ Hz using an anti-aliasing filter. As the trials had variable lengths, we have considered the first 53 s of each trial for analysis. We have considered consecutive windows of length 0.25 s for decoder estimation, resulting in $W = 16$ samples per window and $K = 212$ instances for each trial. Also, we have considered lags up to 0.25 s for decoder estimation, i.e., $L_d = 16$. The latter is motivated by the results of (O'Sullivan et al., 2015) suggesting that the most relevant decoder components are within the first 0.25 s lags. Prior studies have argued that the effects of auditory attention and speech perception are strongest in the frontal and close-to-ear EEG electrodes (Power et al., 2012; Khalighinejad et al., 2017; Kähkönen et al., 2001; Bleichner et al., 2016). We have only considered 28 EEG channels in the decoder estimation problem, i.e., $C = 28$, including the frontal channels Fz, F1-F8, FCz, FC1-FC6, FT7-FT10, C1-C6, and the T complex channels T7 and T8. This subsampling of the electrodes is inspired by the results in Mirkovic et al. (2015), which show that using an electrode subset of the same size for decoding results in nearly the same classification performance as in the case of using all the electrodes. Note that for our real-time setting, a channel selection step can considerably decrease the computational cost and the dimensionality of the decoder estimation step, given that a vector of size $1+C(L_d+1)$ needs to be updated within each 0.25 s window.

We have determined the regularization coefficient $\gamma = 0.4$ via cross-validation and the forgetting factor $\lambda = 0.975$, which results in an *effective* data length of 10 s in the estimation of the decoder and is long

enough for stable estimation of the decoding coefficients. It is worth noting that small values of λ, and hence small effective data lengths, may result in an under-determined inverse problem, since the dimension of the decoder is given by $1+C(L_d+1)$. Finally, in the FASTA package, we have used a tolerance of 0.01 together with Nesterov's accelerated gradient descent method to ensure that the processing can be done in an online fashion.

In studies involving correlation-based measures, such as (O'Sullivan et al., 2015; Akram et al., 2016), the convention is to train attended and unattended decoders/encoders using multiple trials and then use them to calculate the correlation measures over the test trials. The correlation-based attention marker, however, did not produce a statistically significant segregation of the attended and the unattended speakers in our analysis. This discrepancy seems to stem from the fact that the estimated encoders/decoders and the resulting correlations in the aforementioned studies are more informative and robust due to the use of batch-mode analysis with multiple trials for decoder estimation, as compared to our real-time framework. The ℓ_1-based attention marker, however, resulted in a meaningful statistical separation between the attended and the unattended speakers. Therefore, in what follows, we present our EEG analysis results using the ℓ_1-based attention marker.

The parameters of the state-space models have been set similar to those used in simulations, i.e., $K_A=\lfloor 15 f_s/W \rfloor$, $K_F=\lfloor 1.5 f_s/W \rfloor$, $a_0=2.008$, $b_0=0.2016$. Considering the 0.25 s lag in the decoder model, the built-in delay in estimating the attentional state for the real-time system is 1.75 s. For estimating the prior distribution parameters for each subject, we use the first 15s of each trial. As mentioned before, considering the 15 s-long sliding window, we can treat the first 15 s of each trial as a tuning step in which the prior parameters are estimated in a supervised manner and the state-space model parameters are initialized with the values estimated using these initial windows. Thus, similar to the simulations, $\left(\alpha_0^{(a)}, \alpha_0^{(u)}, \beta_0^{(a)}, \beta_0^{(u)}, \mu_0^{(a)}, \mu_0^{(u)}\right)$ for each subject have been set according to the parameters of the two fitted Log-Normal distributions on the ℓ_1-norm of the decoders in the first 15 s of the trials, while choosing large variances for the priors to be non-informative.

3.2.2 Estimation Results

Fig. 8 shows the results of applying our proposed framework to EEG data. For graphical convenience, the data have been rearranged so that speaker 1 is always attended. The left, middle and right panels correspond to subjects 1, 2, and 3, respectively. For each subject, three example trials have been displayed in rows A, B, and C. Row A includes trials in which the attention marker clearly separates the attended and unattended speakers, while Row C contains trials in which the attention marker fails to do so. Row B displays trials in which on average the ℓ_1-norm of the estimated decoder is larger for the attended speaker; however, occasionally, the attention marker fails to capture the attended speaker.

Consistent with our simulations, the real-time estimates (third graphs in rows A, B and C) generally follow the output of the batch-mode estimates (second graphs in rows A, B and C). However, the batch-mode estimates yield smoother transitions and larger confidence intervals in general, both of which are due to having access to future observations.

Figure 9 shows the effect of forward-lag K_F on the performance of real-time estimates, similar to that shown in Fig. 6 for the simulations. The forward-lag K_F is increased from 0 s to 5 s with 0.5 s increments while all the other parameters of the EEG analysis remain the same. The MSE in Fig. 9 has been averaged over all trials for each subject. As we observe in the incremental MSE plot, even a 0.5 s lag can significantly decrease the MSE from the case of 0 s forward-lag (corresponding to the fully real-time setting). Similar

to the simulations, we have chosen a K_F of $1.5\,\mathrm{s}$ for the EEG analysis, since the incremental MSE improvements are significant at this lag, and this choice results in a tolerable built-in delay for real-time applications.

Finally, Fig. 10 summarizes the *real-time* classification results of our EEG analysis at the group level, in order to present subject-specific and individual trial performances. Fig. 10-A shows a cartoon of the estimated attention probabilities for a generic trial in order to illustrate the classification conventions. We define an instance (i.e., K consecutive windows of length W) to be correctly (incorrectly) classified if the estimated attentional state probability together with its 90% confidence intervals lie above (below) 0.5. If the 90% confidence interval at an instance includes the 0.5 attention probability line, we do not classify it

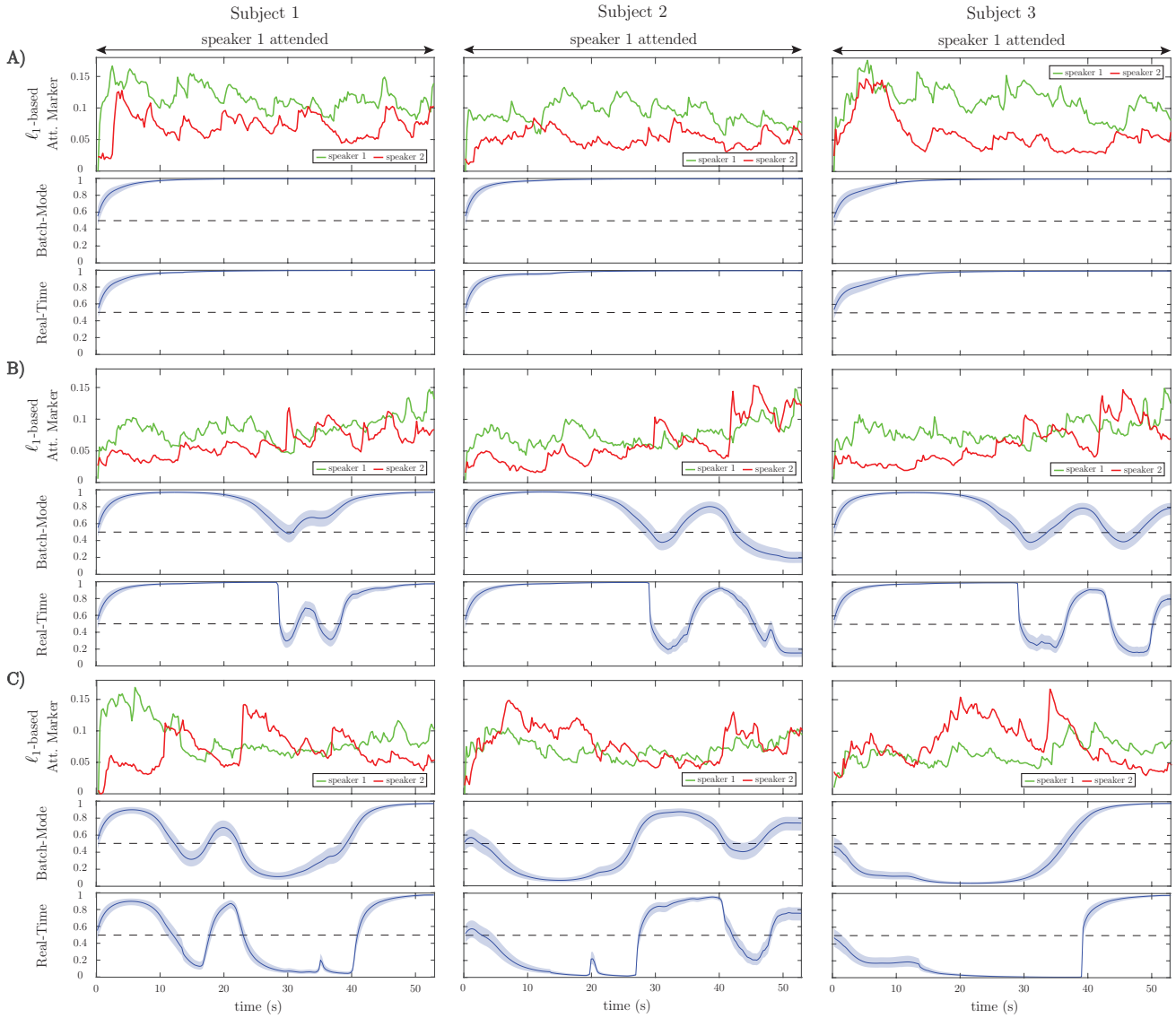

Figure 8. Examples of the ℓ_1-based attention markers (first panels), batch-mode (second panels), and real-time (third panels) state-space estimation results for nine selected EEG trials. A) Representative trials in which the attention marker reliably separates the attended and unattended speakers. B) Representative trials in which the attention marker separates the attended and unattended speakers on average over the trial. C) Representative trials in which the attention marker either does not separate the two speakers or results in a larger output for the unattended speaker.

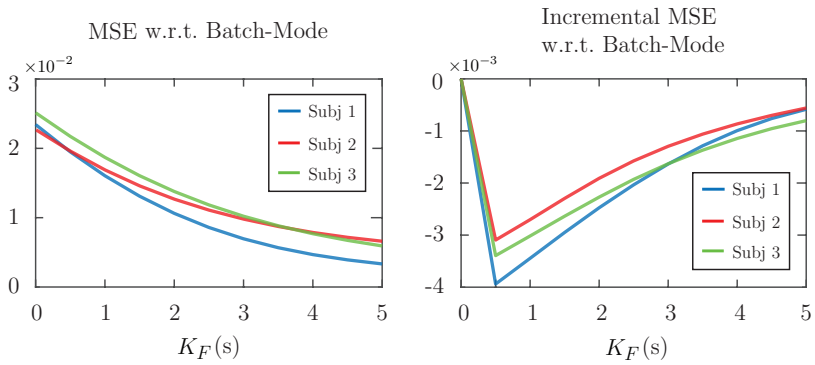

Figure 9. Effect of the forward-lag K_F on MSE in application to real EEG data. The left panel shows the MSE with respect to the batch-mode output averaged over all the trials for each subject. The right panel displays the incremental MSE at each lag, from K_F of 0 s to K_F of 5 s with 0.5 s increments.

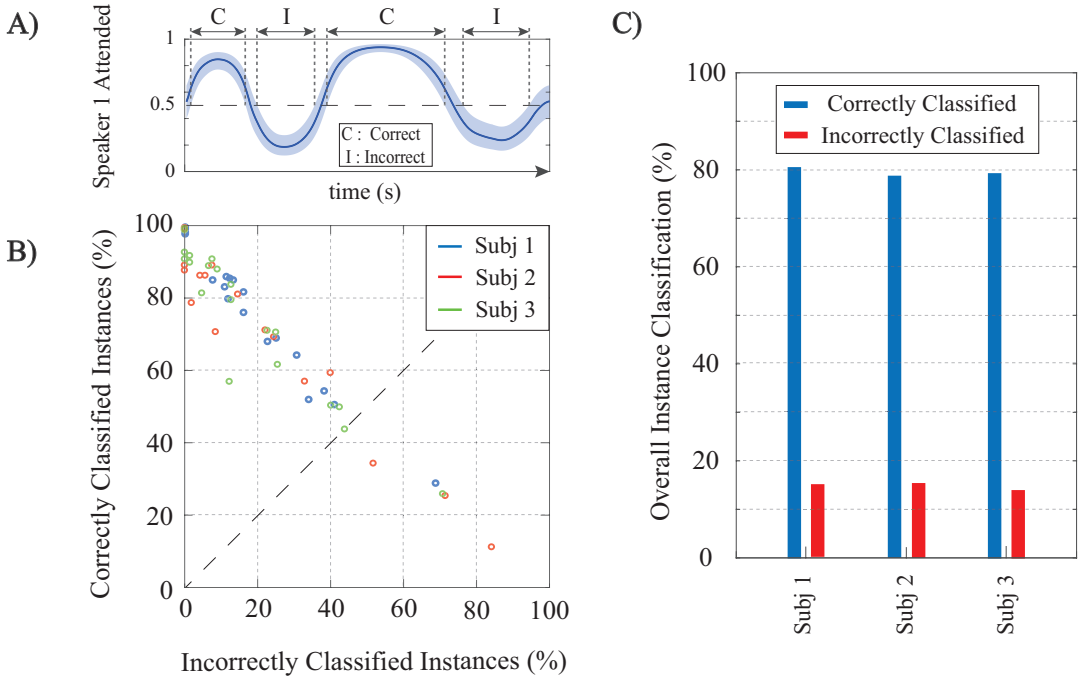

Figure 10. Summary of the real-time classification results in application to real EEG data. A) a generic example of the state-space output for a trial illustrating the classification conventions. B) Classification results per trial for all subjects; each circle corresponds to a trial and the subjects are color-coded. The trials falling below the dashed line have more incorrectly classified instances than correctly classified ones. C) Average classification performance over all trials for the three subjects.

as either correct or incorrect. Figure 10-B displays the correctly classified instances (y-axis) versus those incorrectly classified (x-axis) for each trial. The subjects are color-coded and each circle corresponds to one trial. The average classification results over all trials for each subject are shown in Figure 10-C. In summary, our framework provides $\sim 80\%$ average hit rate and $\sim 15\%$ average false-alarm per trial per subject. The group-level hit rate and false alarm rate are respectively given by 79.63% and 14.84%.

3.3 Application to MEG

In this subsection, we apply our real-time attention decoding framework to MEG recordings of multiple subjects in a dual-speaker environment. The MEG experimental procedures are discussed in Section 2.5.

3.3.1 Preprocessing and Parameter Selection

The recorded MEG responses were band-pass filtered between 1 Hz-8 Hz (delta and theta bands), corresponding to the slow temporal modulations in speech (Ding and Simon, 2012b,a), and downsampled to 200 Hz. MEG recordings, like EEG, include both the stimulus-driven response as well as the background neural activity, which is irrelevant to the stimulus. For the encoding model used in our analysis, we need to extract the stimulus-driven portion of the response, namely the auditory component. In (Särelä and Valpola, 2005; de Cheveigne and Simon, 2008), a blind source separation algorithm called the Denoising Source Separation (DSS) is described which decomposes the data into temporally uncorrelated components ordered according to their trial-to-trial phase-locking reliability. In doing so, DSS only requires the responses in different trials and not the stimuli. Similar to (Akram et al., 2017, 2016), we only use the first DSS component as the auditory component, since it tends to capture a significant amount of stimulus information and to produce a bilateral stereotypical auditory field pattern.

Since DSS is an *offline* algorithm operating on all the data at once, we cannot readily use it for real-time attention decoding. Instead, we apply DSS to the data from preliminary trials from each subject in order to calculate the *subject-specific* linear combination of the MEG channels that compose the first DSS component. We then use these channel weights to extract the MEG auditory responses during the constant-attention and attention-switch experiments in a real-time fashion. Note that the MEG sensors are not fixed with respect to the head position across subjects and are densely distributed in space. Therefore, it is not reasonable to use the same MEG channel weights for all subjects. The preliminary trials for each subject can thus serve as a training and tuning step prior to the application of our proposed attention decoding framework.

The MEG auditory component extracted using DSS is used as E_t in our encoding model. Similar to our foregoing EEG analysis, we have considered consecutive windows of length $0.25\,$s resulting in $W = 50$ samples per window and a total number of $K = 240$ instances, at a sampling frequency of $200\,$Hz. The TRF length, or the total encoder lag, has been set to $0.4\,$s resulting in $L_e = 80$ in order to include the most significant TRF components (Ding and Simon, 2012a). The ℓ_1-regularization parameter γ in Eq. (1) has been adjusted to 1 through two-fold cross-validation, and we have chosen a forgetting factor of $\lambda = 0.975$, resulting in an *effective* data length of $10\,$s, long enough to ensure estimation stability.

As for the encoder model, we have used a Gaussian dictionary $\mathbf{G_0}$ to enforce smoothness in the TRF estimates. The columns of $\mathbf{G}_0$ consist of overlapping Gaussian kernels with the standard deviation of 20 ms whose means cover the $0\,$s to $0.4\,$s lag range with $T_s = 5$ ms increments. The $20\,$ms standard deviation is consistent with the average full width at half maximum (FWHM) of an auditory MEG evoked response (M50 or M100), empirically obtained from MEG studies (Akram et al., 2017). Thus, the overall dictionary discussed in Remark 2 takes the form $\mathbf{G} = \mathrm{diag}\,(1, \mathbf{G_0}, \mathbf{G_0})$. Also, similar to (Akram et al., 2017), we have used the logarithm of the speech envelopes as the regression covariates. Finally, the parameters of the FASTA package in encoder estimation have been chosen similar to those in the foregoing EEG analysis.

The M100 component of the TRF has shown to be larger for the attended speaker than the unattended speaker (Ding and Simon, 2012a; Akram et al., 2017). Thus, at each instance k, we extract the magnitude of the negative peak close to the $0.1\,$s delay in the real-time TRF estimate of each speaker as the attention markers $m_k^{(1)}$ and $m_k^{(2)}$. For the state-space model and the fixed-lag window, we have used the same configuration as in our foregoing EEG analysis, i.e. $K_A = \lfloor 15 f_s/W \rfloor$, $K_F = \lfloor 1.5 f_s/W \rfloor$, $a_0 = 2.008$, and $b_0 = 0.2016$. Note that the built-in delay in estimating the attentional state is now only $1.5\,$s, given that we use an encoding model for our MEG analysis. Furthermore, the prior distribution parameters for each

subject were chosen according to the two fitted Log-Normal distributions on the extracted M100 values in the first 15 s of the trials, while choosing large variances for the Gamma priors to be non-informative. Similar to the preceding cases, the first 15 s of each trial can be thought of as an initialization stage.

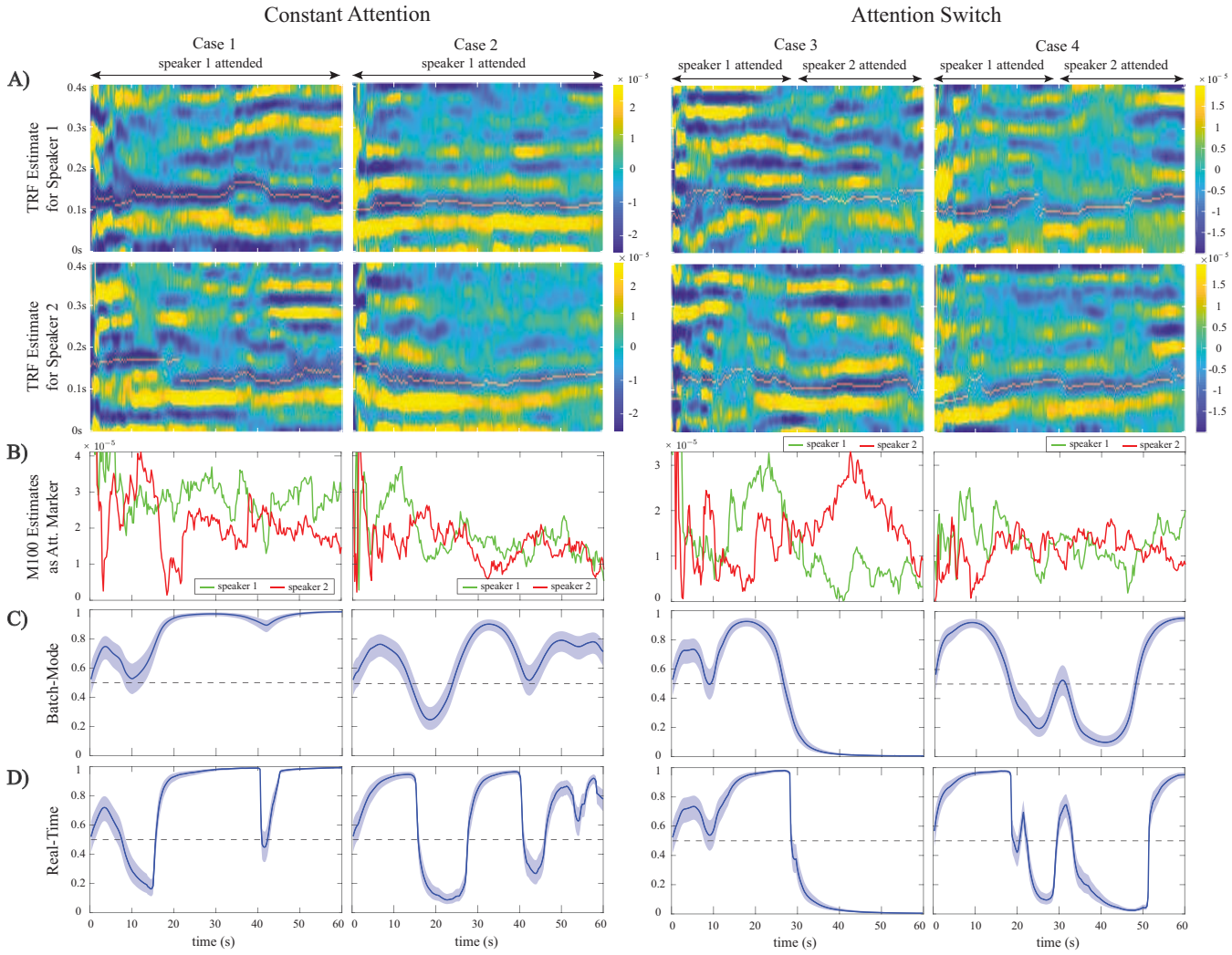

Figure 11. Examples from the constant-attention and attention-switch MEG experiments, using the M100 attention marker, for trials with reliable (cases 1 and 3) and unreliable (cases 2 and 4) separation of the attended and unattended speakers. A) TRF estimates for speakers 1 and 2 over time with the extracted M100 peak positions tracked by a narrow yellow line. B) Extracted M100 peak magnitudes over time for speakers 1 and 2 as the attention marker. In cases 1 and 3, the M100 components exhibit a strong modulation effect of the attentional state, i.e., the attended speaker has a larger M100 peak, in contrast to cases 2 and 4, where there is a weak modulation. C) Batch-mode state-space estimates of the attentional state. D) Real-time state-space estimates of the attentional state. The strong or weak modulation effects of attentional state in the extracted M100 components directly affects the classification accuracy and the width of the confidence intervals for both the batch-mode and real-time estimators.

3.3.2 Estimation Results

Figure 11 shows our estimation results for four sample trials from the constant-attention (cases 1 and 2) and attention-switch (cases 3 and 4) experiments. For graphical convenience, we have rearranged the MEG data such that in the constant-attention experiment, the attention is always on speaker 1, and in the attention-switch experiment, speaker 1 is attended from 0 s to 28 s. Cases 1 and 3 corresponds to trials in

which the extracted M100 values for the attended speaker are more significant than those of the unattended speaker during most of the trial duration. Cases 2 and 4, on the other hand, correspond to trials in which the extracted M100 values are not reliable representatives of the attentional state. Row A in Fig. 11 shows the estimated TRFs for speakers 1 and 2 in time for each of the four cases. The location of the M100 peaks is shown and tracked with a narrow line (yellow) on the extracted M100 components (blue). The M50 components are also evident as positive peaks occurring around the 50 ms lag. The M50 components do not strongly depend on the attentional state of the listener (Akram et al., 2017; Ding and Simon, 2012a; Chait et al., 2004, 2010), which is consistent with those shown in Fig. 11-A. It is worth noting that real-time estimation of the TRFs makes the estimates heavily affected by the dynamics of neural response and the background neural activity. Therefore, the estimates contain longer latency components which are typically suppressed in the offline estimates of TRFs common in the literature, which use multiple trial averaging to extract the stimulus-driven response (Ding and Simon, 2012a; Power et al., 2012). The width of the extracted components in Fig. 11 is due to the usage of a Gaussian dictionary matrix to represent the TRFs.

Row B in Fig. 11 displays the extracted M100 peak magnitudes over time for speakers 1 and 2. The attention modulation effect is more significant in cases 1 and 3. Rows C and D respectively show the batch-mode and real-time estimates of the attentional state based on the extracted M100 values. As expected, the batch-mode output is more robust to the fluctuations in the extracted M100 peak values, with smoother transitions and larger confidence intervals. Despite the poor attention modulation effect in cases 2 and 4, we observe that both the real-time and the batch-mode state-space models show reasonable performance in translating the extracted M100 peak values to a robust measure of the attentional state. This effect is notable in Rows C and D of Case 4. We performed the same analysis as in Fig. 9 to assess the effect of the forward-lag parameter K_F. Since the results were quite similar to those in Figures 6 and 9, we have omitted them for brevity and chose the same forward-lag of 1.5 s.

Finally, Fig. 12 summarizes the *real-time* classification results for the constant-attention (left panels) and attention-switch (right panels) MEG experiments. The classification convention is similar to that used in our EEG analysis, and is illustrated in Fig. 12-A for the completeness. For the attention-switch experiment, the 28 s-30 s interval is removed from the classification analysis, as it pertains to a silence period during which the subject is instructed to switch attention. Fig. 12-B shows the corresponding classification results, consisting of 36 trials for the constant-attention and 18 trials for the attention-switch experiments. Each circle corresponds to a single trial and the subjects in each experiment are color-coded. The average classification results per trial are shown in Fig. 12-C for each subject. The average hit rate and false alarm rates in the constant-attention experiments are respectively given by 71.67% and 20.81%. These quantities for the attention-switch experiment are respectively given by 64.12% and 26.16%, showing a reduction in hit rate and increase in false alarm.

4 DISCUSSION

In this work, we have proposed a framework for real-time decoding of the attentional state of a listener in a dual-speaker environment from M/EEG. This framework consists of three modules. In the first module, the encoding/decoding coefficients, relating the neural response to the envelopes of the two speech streams, are estimated in a low-complexity and real-time fashion. Existing approaches for encoder/decoder estimation operate in an offline fashion using multiple experiment trials or large training datasets (O'Sullivan et al., 2015; Akram et al., 2016; Van Eyndhoven et al., 2017; Aroudi et al., 2016), and hence are not suitable for real-time applications with limited amount of training data and potential variability in the recording setup.

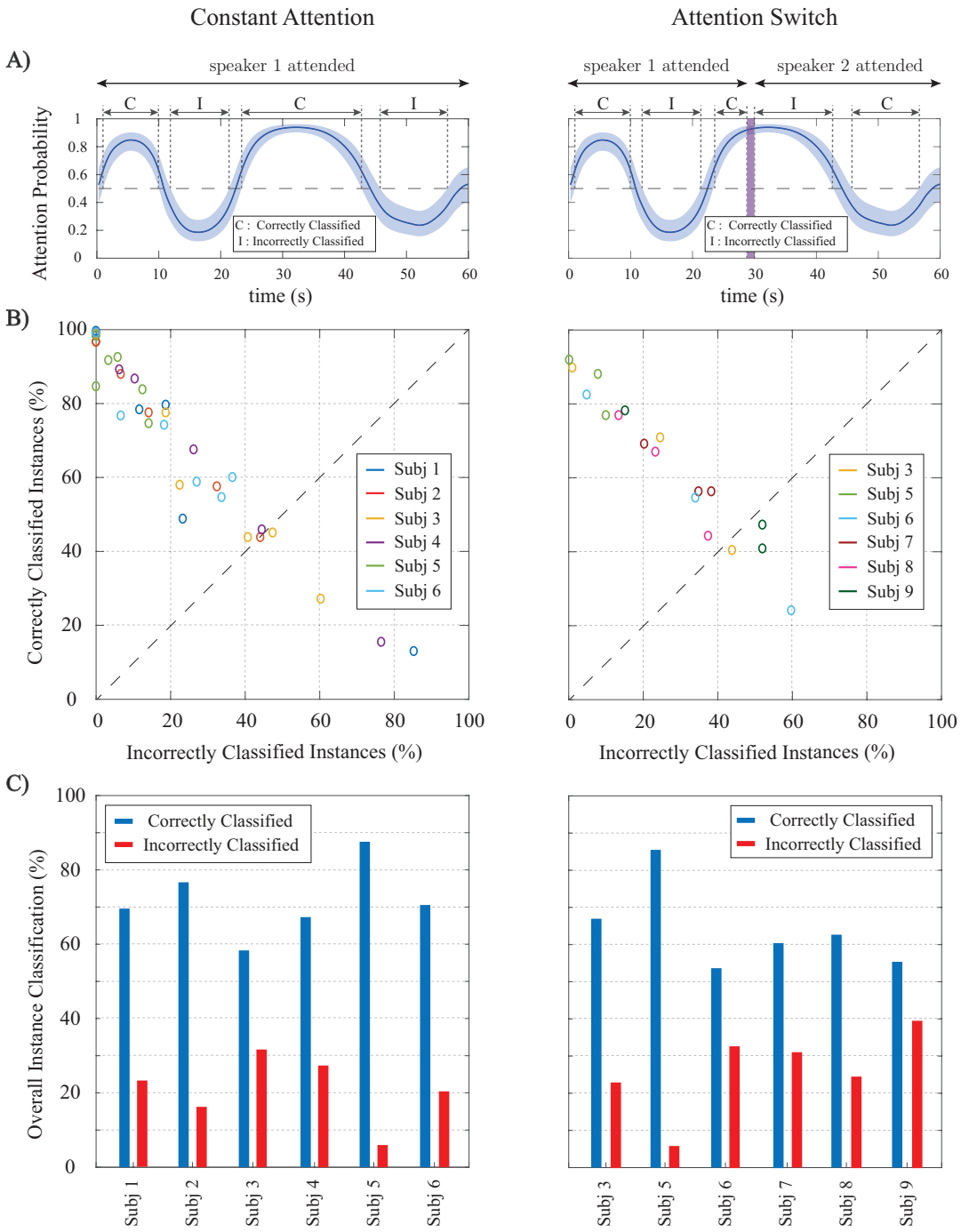

Figure 12. Summary of real-time classification results for the constant-attention (left panels) and attention-switch (right panels) MEG experiments. A) a generic instance of the state-space output for a trial illustrating the classification convention. B) Classification results per trial for all subjects; each circle corresponds to a trial and the subjects are color-coded. The trials falling below the dashed line have more incorrectly classified instances than correctly classified ones. C) Average classification performance over all trials for the six subjects.

To address this issue, we have integrated the forgetting factor mechanism used in adaptive filtering with ℓ_1-regularization, in order to capture the coefficient dynamics and mitigate overfitting.

In the second module, a function of the estimated encoding/decoding coefficients and the acoustic data, which we refer to as the *attention marker*, is calculated in real-time for each speaker. The role of the attention marker is to provide dynamic features that create statistical separation between the attended and the unattended speakers. Examples of such attention markers include correlation-based measures (e.g. correlation of the acoustic envelopes and their reconstruction from neural response), or measures solely based on the estimated decoding/encoding coefficients (e.g. the ℓ_1-norm of the decoder coefficients or the M100 peak of the encoder).

Finally, the attention marker is passed to the third module consisting of a near real-time state-space estimator. To control the delay in state estimation, we adopt a fixed-lag smoothing paradigm, in which the past and near future data are used to estimate the states. The role of the state-space model is to translate the noisy and highly variable attention markers to robust measures of the attentional state with minimal delay. We have archived a publicly available MATLAB implementation of our framework on the open source repository GitHub in order to ease reproducibility (Miran, 2017).

We validated the performance of our proposed framework using simulated EEG and MEG data, in which the ground truth attentional states are known. We also applied our proposed methods to experimentally recorded MEG and EEG data. As for a comparison benchmark to study the effect of the parameter choices in our real-time estimator, we considered the offline state-space attention decoding approach of (Akram et al., 2016). Our MEG analysis showed that although the proposed real-time estimator has access to significantly fewer data points, it closely matches the outcome of the offline state-space estimator in (Akram et al., 2016), for which the entire data from multiple trials are used for attention decoding. In particular, our analysis of the MEG data in constant-attention conditions revealed a hit rate of $\sim 70\%$ and a false alarm rate of $\sim 20\%$ at the group level. While the performance is slightly degraded compared to the offline analysis of (Akram et al., 2016), our algorithms operate in real-time with 1.5 s built-in delay, over single trials, and using minimal tuning. Similarly, our analysis of EEG data provided $\sim 80\%$ hit rate and $\sim 15\%$ false alarm rate at a single trial level. These performance measures are slightly degraded compared to the results of offline approaches such as (O'Sullivan et al., 2015).

Our proposed modular design admits the use of any attention-modulated statistic or feature as the attention marker, three of which have been considered in this work. While some attention markers perform better than the rest in certain applications, our goal in this work was to provide different examples of attention markers which can be used in the encoding/decoding models based on the literature, rather than comparing their performance against each other. The choice of the best attention marker that results in the highest classification accuracy is a problem-specific matter. Our modular design allows to evaluate the performance of a variety of attention markers for a given experimental setting, while fixing the encoding/decoding estimation and state-space modules, and to choose one that provides the desired classification performance. Our state-space module can also operate on the output of existing methods with encoder/decoder coefficients that are pre-estimated using training datasets (O'Sullivan et al., 2015; Zink et al., 2017) to provide a robust and statistically interpretable measure of the attentional state at high temporal resolutions.

A practical limitation of our proposed methodology in its current form is the need to have access to clean acoustic data in order to form regressors based on the speech envelopes. In a realistic scenario, the speaker envelopes have to be extracted from the noisy mixture of speeches recorded by microphone arrays. Thanks to a number of fairly recent results in attention decoding literature (Van Eyndhoven et al., 2017; Biesmans et al., 2015, 2017; Aroudi et al., 2016; O'Sullivan et al., 2017), it is possible to integrate our methodology with a pre-processing module that extracts the acoustic features of individual speech streams from their noisy mixtures. We view this extension as a future direction of research.

The proposed approach requires a minimal amount of *labeled* training data for tuning purposes. However, we can determine the attended speaker in an unlabeled dataset as the speaker whose speech signal best fits the EEG data or whose encoder/decoder estimates have larger peaks at certain time lags, and then train the decoders or hyperparameters with these data-driven labels. This can be done both in existing methods such as that of O'Sullivan et al. (2015) for attended decoder estimation and in our approach for capturing the statistical properties of attention markers for hyperparameter tuning. We view this extension to deal with unlabeled data as a future direction of research.

Our proposed framework has several advantages over existing methodologies. First, our algorithms require minimal amount of offline tuning or training. The subject-specific hyperparameters used by the algorithms are tuned prior to real-time application in a supervised manner. The only major offline tuning step in our framework is computing the subject-specific channel weights in the encoding model for MEG analysis in order to extract the auditory component of the neural response. This is due to the fact that the channel locations are not fixed with respect to the head position across subjects. It is worth noting that this step can be avoided if the encoding model treats the MEG channels separately in a multivariate model. Given that recent studies suggest that the M100 component of the encoder obtained from the MEG auditory response is a reliable attention marker (Ding and Simon, 2012a,b; Akram et al., 2017), we adopted the DSS algorithm for computing the channel weights that compose the auditory response in an offline fashion.

Second, our framework yields robust attention decoding performance at a temporal resolution in the order of ~ 1 second, comparable to that at which humans switch their attention from one speaker to another. The accuracy of existing methods, however, significantly degrades when they operate at these temporal resolutions (Zink et al., 2016, 2017). Our proposed framework operates in a near real-time fashion, where the attention decoding delay can be adjusted for controlling the trade-off between robustness and adaptivity of the attentional state estimates. In addition, the probabilistic output of our attentional state decoding framework can be used for further statistical analysis and soft-decision mechanisms which are desired in smart hearing aid applications. Finally, the modular design of our framework facilitates its adaptation to more complex auditory scenes (e.g. with multiple speakers and realistic noise and reverberation conditions) and integration of other covariates relevant to real-time applications (e.g. electrooculography measurements).

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. Author Tao Zhang is employed by Starkey Hearing Technologies. Author Sahar Akram is employed by Facebook. All other authors declare no competing interests.

AUTHOR CONTRIBUTIONS

TZ, JZS and BB designed the research. SM and BB performed the research, with contributions to the methods by AS, SA, and TZ and experimental data supplied by TZ and JZS. All authors participated in writing the paper.

FUNDING

This material is based on work supported in part by the National Science Foundation Awards No. 1552946 and 1734892 and a research gift from the Starkey Hearing Technologies to the University of Maryland.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Tom Goldstein for helpful remarks on adapting the FASTA package options to our decoder/encoder estimation problem.

REFERENCES

Akram, S., Presacco, A., Simon, J. Z., Shamma, S. A., and Babadi, B. (2016). Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *NeuroImage* 124, 906–917

Akram, S., Simon, J. Z., and Babadi, B. (2017). Dynamic estimation of the auditory temporal response function from MEG in competing-speaker environments. *IEEE Transactions on Biomedical Engineering* 64, 1896–1905

Akram, S., Simon, J. Z., Shamma, S. A., and Babadi, B. (2014). A state-space model for decoding auditory attentional modulation from MEG in a competing-speaker environment. In *Advances in Neural Information Processing Systems*. 460–468

Aroudi, A., Mirkovic, B., De Vos, M., and Doclo, S. (2016). Auditory attention decoding with EEG recordings using noisy acoustic reference signals. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on* (IEEE), 694–698

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2017). 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* 25, 402–412

Biesmans, W., Vanthornhout, J., Wouters, J., Moonen, M., Francart, T., and Bertrand, A. (2015). Comparison of speech envelope extraction methods for EEG-based auditory attention detection in a cocktail party scenario. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE* (IEEE), 5155–5158

Bleichner, M. G., Mirkovic, B., and Debener, S. (2016). Identifying auditory attention with ear-EEG: cEEGrid versus high-density cap-EEG comparison. *Journal of Neural Engineering* 13, 066004

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT press)

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America* 109, 1101–1109

Chait, M., de Cheveigné, A., Poeppel, D., and Simon, J. Z. (2010). Neural dynamics of attending and ignoring in human auditory cortex. *Neuropsychologia* 48, 3262–3271

Chait, M., Simon, J. Z., and Poeppel, D. (2004). Auditory m50 and m100 responses to broadband noise: functional implications. *Neuroreport* 15, 2455–2458

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America* 25, 975–979

Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering* (Springer New York). 185–212

de Cheveigne, A. and Simon, J. Z. (2008). Denoising based on spatial filtering. *Journal of Neuroscience Methods* 171, 331–339

Ding, N. and Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences* 109, 11854–11859

Ding, N. and Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology* 107, 78–89

Fishman, Y. I. and Steinschneider, M. (2010). Neural correlates of auditory scene analysis based on inharmonicity in monkey primary auditory cortex. *Journal of Neuroscience* 30, 12480–12494

Goldstein, T., Studer, C., and Baraniuk, R. (2014). A field guide to forward-backward splitting with a FASTA implementation. *arXiv eprint* abs/1411.3406

Goldstein, T., Studer, C., and Baraniuk, R. (2015). FASTA: A generalized implementation of forward-backward splitting. `http://arxiv.org/abs/1501.04979`

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* 77, 980–991

Griffiths, T. D. and Warren, J. D. (2004). What is an auditory object? *Nature reviews. Neuroscience* 5, 887

Haykin, S. and Chen, Z. (2005). The cocktail party problem. *Neural Computation* 17, 1875–1902

Kähkönen, S., Ahveninen, J., Jääskeläinen, I. P., Kaakkola, S., Näätänen, R., Huttunen, J., et al. (2001). Effects of haloperidol on selective attention: a combined whole-head MEG and high-resolution EEG study. *Neuropsychopharmacology* 25, 498–504

Kaya, E. M. and Elhilali, M. (2017). Modelling auditory attention. *Phil. Trans. R. Soc. B* 372, 20160101

Khalighinejad, B., da Silva, G. C., and Mesgarani, N. (2017). Dynamic encoding of acoustic features in neural responses to continuous speech. *Journal of Neuroscience* 37, 2176–2185

McDermott, J. H. (2009). The cocktail party problem. *Current Biology* 19, R1024–R1027

Middlebrooks, J. C., Simon, J. Z., Popper, A. N., and Fay, R. R. (2017). *The Auditory System at the Cocktail Party* (in the Springer Handbook of Auditory Research series)

Miran, S. (2017). *Real-Time Tracking of Selective Auditory Attention MATLAB Code* (Available on GitHub Repository: `https://github.com/sinamiran/Real-Time-Tracking-of-Selective-Auditory-Attention`)

Mirkovic, B., Debener, S., Jaeger, M., and De Vos, M. (2015). Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *Journal of Neural Engineering* 12, 046007

O'Sullivan, J., Chen, Z., Herrero, J., McKhann, G. M., Sheth, S. A., Mehta, A. D., et al. (2017). Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *Journal of Neural Engineering* 14

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex* 25, 1697–1706

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., and Lalor, E. C. (2012). At what time is the cocktail party? a late locus of selective attention to natural speech. *European Journal of Neuroscience* 35, 1497–1503

Särelä, J. and Valpola, H. (2005). Denoising source separation. *Journal of Machine Learning Research* 6, 233–272

Shamma, S. A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences* 34, 114–123

Sheikhattar, A., Fritz, J. B., Shamma, S. A., and Babadi, B. (2015a). Adaptive sparse logistic regression with application to neuronal plasticity analysis. In *Signals, Systems and Computers, 2015 49th Asilomar Conference on* (IEEE), 1551–1555

Sheikhattar, A., Fritz, J. B., Shamma, S. A., and Babadi, B. (2015b). Recursive sparse point process regression with application to spectrotemporal receptive field plasticity analysis. *IEEE Transactions on Signal Processing* 64, 2026–2039

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* , 267–288

Van Eyndhoven, S., Francart, T., and Bertrand, A. (2017). EEG-informed attended speaker extraction from recorded speech mixtures with application in neuro-steered hearing prostheses. *IEEE Transactions on Biomedical Engineering* 64, 1045–1056

Zink, R., Baptist, A., Bertrand, A., Van Huffel, S., and De Vos, M. (2016). Online detection of auditory attention in a neurofeedback application. In *Proc. 8th International Workshop on Biosignal Interpretation*. 1–4

Zink, R., Proesmans, S., Bertrand, A., Van Huffel, S., and De Vos, M. (2017). Online detection of auditory attention with mobile EEG: closing the loop with neurofeedback. *bioRxiv* , 218727

frontiers

Supplementary Material: Real-Time Tracking of Selective Auditory Attention from M/EEG: A Bayesian Filtering Approach

Sina Miran, Sahar Akram, Alireza Sheikhattar, Jonathan Z. Simon, Tao Zhang, and Behtash Babadi*

*Correspondence:
Author Name: Behtash Babadi
behtash@umd.edu

This supplementary document contains the derivations of our proposed estimation framework as well as additional simulation studies. In Section 1, we present the parameter estimation procedures used for the encoding and decoding models. Section 2 includes the inference algorithms for state estimation using fixed-lag smoothing, and Section 3 discusses the smoothing effect of the proposed state-space model. Finally, we apply our proposed techniques to simulated MEG data in Section 4.

1 DYNAMIC ENCODING AND DECODING MODELS: PARAMETER ESTIMATION

Recall that the encoder/decoder estimation problems can be posed as the following optimization problem:

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{k} \lambda^{k-j} \| \mathbf{y}_j - \mathbf{X}_j \boldsymbol{\theta} \|_2^2 + \gamma \| \boldsymbol{\theta} \|_1 , \quad k = 1, 2, \ldots, K. \tag{S1}$$

At each window k, for $k = 1, \ldots, K$, the encoding/decoding coefficients $\hat{\boldsymbol{\theta}}_k$ are updated based on the new measurements, i.e., $\mathbf{y}_k$ and $\mathbf{X}_k$, and previous measurements through the forgetting factor mechanism while applying sparsity-promoting priors on the coefficients.

There are several standard optimization techniques that can be used to find the minimizer in (S1). Off-line algorithms such as interior point methods do not meet the real-time requirements of our dynamic estimation. The SPARLS algorithm has been introduced in (Babadi et al., 2010) to solve the problem in (S1) through EM iterations, and it has been successfully adopted in (Akram et al., 2017) to estimate encoding coefficients in a dynamic fashion. However, the EM algorithm and the constant step-size in SPARLS may result in low convergence rates. Hence, to adapt our estimation procedure for real-time applications, we use the Forward-Backward Splitting (FBS) method (Combettes and Pesquet, 2011), also known as the proximal gradient method, to solve for $\hat{\boldsymbol{\theta}}_k$ in (S1). FBS is suited for optimization problems where the objective function can be expressed as the sum of a differentiable term, e.g., the log-likelihood term in (S1), and a simple non-differentiable term, e.g., the ℓ_1-norm in (S1). This type of problems frequently arise in signal processing and machine learning (Jenatton et al., 2010; Duchi and Singer, 2009; Figueiredo et al., 2007).

In summary, each FBS iteration for the optimization problem in (S1) includes two steps: 1) taking a descent step along the gradient of the log-likelihood term, and 2) applying a soft-thresholding shrinkage operator (Goldstein et al., 2014; Sheikhattar et al., 2015). This procedure provides an algorithm that uses recursive and low-complexity updates in an online fashion to solve Eq. (S1) upon the arrival of a new data window. The optimization problem in (S1) can be rewritten as:

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}} \ \boldsymbol{\theta}^T \mathbf{A}_k \boldsymbol{\theta} + \mathbf{b}_k^T \boldsymbol{\theta} + \gamma \parallel \boldsymbol{\theta} \parallel_1 , \quad k = 1, 2, \ldots, K, \tag{S2}$$

where $\mathbf{A}_k$ and $\mathbf{b}_k$ can be updated recursively. Algorithm 1 summarizes the steps of the FBS algorithm to solve for $\boldsymbol{\theta}_k$ in (S1), when moving from window $k-1$ to window k, as well as the required recursive update rules for $\mathbf{A}_k$ and $\mathbf{b}_k$. The parameter $\mathcal{S}_{FBS}$ in Algorithm 1 denotes the stopping condition for the FBS algorithm, which can be a maximum iteration number or a convergence criterion on the objective function.

Algorithm 1 Parameter Estimation in Dynamic Encoding and Decoding Models by Forward-Backward Splitting

Input: $\mathbf{y}_k$, $\mathbf{X}_k$, $\hat{\boldsymbol{\theta}}_{k-1}$, $\mathbf{A}_{k-1}$, $\mathbf{b}_{k-1}$, λ, γ, $\mathcal{S}_{FBS}$
Output: $\hat{\boldsymbol{\theta}}_k$, $\mathbf{A}_k$, $\mathbf{b}_k$

$\mathbf{A}_k = \lambda \mathbf{A}_{k-1} + \mathbf{X}_k^T \mathbf{X}_k$
$\mathbf{b}_k = \lambda \mathbf{b}_{k-1} - 2\mathbf{X}_k^T \mathbf{y}_k$
initialize $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}_{k-1}$
while $\neg \mathcal{S}_{FBS}$ **do**
    choose stepsize τ
    $\mathbf{u} = \boldsymbol{\theta} - \tau \left(2\mathbf{A}_k \boldsymbol{\theta} + \mathbf{b}_k\right)$
    $\boldsymbol{\theta}_i = \text{sign}(\mathbf{u}_i) \times \max\left\{|\mathbf{u}_i| - \gamma\tau, 0\right\}$, for each element of $\boldsymbol{\theta}$
end while
$\hat{\boldsymbol{\theta}}_k = \boldsymbol{\theta}$.

Remark 1. A proper step-size choice in Alg. 1 at each FBS iteration is crucial to the convergence of the algorithm. For a fixed step-size, it has been shown that $\tau < \frac{2}{L(\nabla f_k)}$ ensures the stability and convergence of the algorithm (Combettes and Pesquet, 2011), where $L(.)$ represents the Lipschitz constant, and f_k represents the log-likelihood term in (S1). Through standard Cauchy-Schwarz and triangle inequality manipulations, we can calculate the simple upper bound $L(\nabla f_k) \leq L_{\text{ub}} = 2\sum_{j=1}^{k} \lambda^{k-j} \,\text{trace}\left\{\mathbf{X}_k^T \mathbf{X}_k\right\}$, implying that $\tau < \frac{2}{L_{\text{ub}}}$ ensures stability; however, this loose upper bound may decrease the convergence rate of the algorithm. Thus, it is more beneficial to ensure stability through backtracking and employing acceleration schemes such as adaptive step-size selection or the Nesterov's method (Goldstein et al., 2014). We have used the FASTA software package (Goldstein et al., 2014) available online at (Goldstein et al., 2015) in this work, which has built-in features for all the foregoing FBS step-size adjustment methods.

2 DYNAMIC STATE-SPACE MODEL: PARAMETER ESTIMATION

Recall that p_k denotes the probability of attending to speaker 1 at instance k for $k = 1, \ldots, K_A$. Although each k corresponds to a data window in time, we refer to it as an *instance* not to conflate it with the fixed-lag

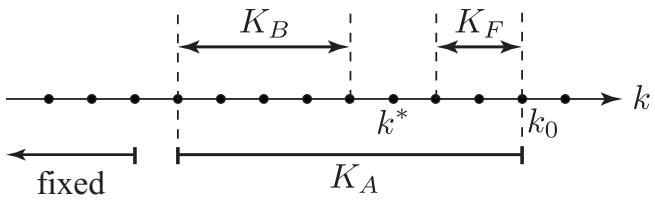

Figure S1. The parameters involved in state-space fixed-lag smoothing.

sliding window used for state estimation. The parameter K_A denotes the number of instances in fixed-lag smoothing as shown in Figure S1 (replaced from Figure 2 for completeness).

The linear state-space model which we apply on $\mathrm{logit}(p_k) = \ln\left(\frac{p_k}{1-p_k}\right)$, can be summarized as:

$$\begin{cases} p_k = \mathrm{P}\left(n_k{=}1\right) = 1 - \mathrm{P}\left(n_k{=}2\right) = \frac{1}{1+\exp(-z_k)} \\ z_k = c_0 z_{k-1} + w_k \\ w_k \sim \mathcal{N}(0, \eta_k) \\ \eta_k \sim \text{Inverse-Gamma}\left(a_0, b_0\right) \end{cases} \tag{S3}$$

Let $m_k^{(1)}$ and $m_k^{(2)}$ represent the attention markers and n_k represent a binary random variable taking values 1 or 2 depending on the attended speaker at instance k for $k = 1, \ldots, K_A$. The observation equations of the state-space model, which relate the observed $m_{1:K_A}^{(1)}$ and $m_{1:K_A}^{(2)}$ to the hidden variables of the state-space model in Eq. (S3), can be summarized as:

$$\begin{cases} m_k^{(i)} \,\Big|\, n_k{=}i \sim \text{Log-Normal}\left(\rho^{(a)}, \mu^{(a)}\right), \quad i = 1, 2 \\ m_k^{(i)} \,\Big|\, n_k{\neq}i \sim \text{Log-Normal}\left(\rho^{(u)}, \mu^{(u)}\right), \quad i = 1, 2 \\ \rho^{(a)} \sim \text{Gamma}\left(\alpha_0^{(a)}, \beta_0^{(a)}\right), \quad \mu^{(a)} \,\Big|\, \rho^{(a)} \sim \mathcal{N}\left(\mu_0^{(a)}, \rho^{(a)}\right) \\ \rho^{(u)} \sim \text{Gamma}\left(\alpha_0^{(u)}, \beta_0^{(u)}\right), \quad \mu^{(u)} \,\Big|\, \rho^{(u)} \sim \mathcal{N}\left(\mu_0^{(u)}, \rho^{(u)}\right) \end{cases} \tag{S4}$$

The parameters of the state-space model are, therefore, $\boldsymbol{\Omega} = \left\{z_{1:K_A}, \eta_{1:K_A}, \rho^{(a)}, \mu^{(a)}, \rho^{(u)}, \mu^{(u)}\right\}$, which have to be inferred from $m_{1:K_A}^{(1)}$ and $m_{1:K_A}^{(2)}$. For notational simplicity, hereafter we use the boldface version of a variable to denote a vector containing all its instances, e.g., $\boldsymbol{z} \coloneqq z_{1:K_A}$ and $\boldsymbol{m}^{(i)} \coloneqq m_{1:K_A}^{(i)}$ for $i = 1, 2$.

The inference problem for $\boldsymbol{\Omega}$ can be expressed as:

$$\widehat{\boldsymbol{\Omega}} = \arg\max_{\boldsymbol{\Omega}} \ \ln \mathrm{P}\left(\boldsymbol{\Omega} \,\Big|\, \boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}\right) = \arg\max_{\boldsymbol{\Omega}} \ \ln \mathrm{P}\left(\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)} \,\Big|\, \boldsymbol{\Omega}\right) + \ln \mathrm{P}\left(\boldsymbol{\Omega}\right), \tag{S5}$$

where the log-likelihood and the log-prior are respectively expanded as:

$$\ln \mathrm{P}\left(\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)} \mid \boldsymbol{\Omega}\right) = \ln \left(\sum_{n_{1:K_A}} \sum_{k=1}^{K_A} p_k \, \mathrm{P}\left(m_k^{(1)} \mid n_k, \boldsymbol{\Omega}\right) \mathrm{P}\left(m_k^{(2)} \mid n_k, \boldsymbol{\Omega}\right) \right), \tag{S6}$$

$$\ln \mathrm{P}(\boldsymbol{\Omega}) = \ln \mathrm{P}\left(\rho^{(a)}, \mu^{(a)}\right) + \ln \mathrm{P}\left(\rho^{(u)}, \mu^{(u)}\right) + \underbrace{\sum_{k=1}^{K_A} \left[-\frac{1}{2} \ln \eta_k - \frac{(z_k - c_0 z_{k-1})^2}{2\eta_k} + \ln \mathrm{P}(\eta_k) \right]}_{\ln \mathrm{P}(\boldsymbol{z}, \boldsymbol{\eta})} + \mathsf{cnst}. \tag{S7}$$

Similar to the treatment in (Akram et al., 2016), we use an Expectation Maximization (EM) algorithm with $\boldsymbol{n}$ as the latent variables to infer $\boldsymbol{\Omega}$. Note that the optimization problem in (S5) is non-convex in general; thus, the choice of initial conditions and hyperparameters for priors are important for reaching a desirable local maximum. Having the estimate $\widehat{\boldsymbol{\Omega}}^{(\ell)}$ for $\boldsymbol{\Omega}$ at the ℓ^{th} EM iteration, we will next derive the E-step and M-step of the $(\ell+1)^{\text{th}}$ EM iteration.

2.1 The E-step

In the E-step, the surrogate function $Q\left(\boldsymbol{\Omega} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)$ is calculated as:

$$Q\left(\boldsymbol{\Omega} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right) = \frac{1}{K_A} \underbrace{\mathbb{E}\left\{ \ln \mathrm{P}\left(\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, \boldsymbol{n} \mid \boldsymbol{\Omega}\right) \right\}}_{\mathcal{A}} + \ln \mathrm{P}(\boldsymbol{\Omega}), \tag{S8}$$

where the expectation of the *complete* log-likelihood $\ln \mathrm{P}\left(\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, \boldsymbol{n} \mid \boldsymbol{\Omega}\right)$ needs to be calculated with respect to $\boldsymbol{n}$ given $\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, \widehat{\boldsymbol{\Omega}}^{(\ell)}$. For notational simplicity, hereafter we drop the $\boldsymbol{n} \mid \boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, \widehat{\boldsymbol{\Omega}}^{(\ell)}$ subscript of the conditional expectations.

We have used a *normalized* version of the log-likelihood in Eq. (S8) for two reasons. First, the window length K_A is a hyperparameter in our framework, which we can modify to find the optimal trade-off between the dimensionality of the state-space and history-dependence of the model. Thus, to change the window length for fixed priors, it is important to normalize the contribution of the log-likelihood in (S8). Second, as noted before, we have a non-convex inference problem, which makes the resulting local maximum dependent on the conjugate priors used. We can use samples of $m_k^{(i)}$'s to estimate the attended and the unattended Log-Normal distributions and tune the hyperparameters to these distributions. By normalizing the log-likelihood term, we are enforcing informative and empirical prior distributions which would guide the inference procedure towards a plausible local maximum. For instance, for the correlation-based attention marker, we expect that a plausible solution would result in the attended Log-Normal distribution being concentrated around larger correlation values compared to the unattended distribution. Nevertheless, the forthcoming derivations can be carried out without the normalization factor $1/K_A$ in a similar fashion.

Let $\mathbb{I}_u(v)$ represent the indicator function, i.e., it is equal to one if $v = u$ and zero otherwise. Conditioning on $\boldsymbol{n}$ and using the conditional independence of $\boldsymbol{m}^{(1)}$ and $\boldsymbol{m}^{(2)}$ given $\boldsymbol{n}$ and $\boldsymbol{\Omega}$, the expected log-likelihood $\mathcal{A}$ in (S8) can be simplified as:

$$\begin{aligned}\mathcal{A} &= \sum_{i=1}^{2} \mathbb{E}\left\{\ln \mathrm{P}\left(\boldsymbol{m}^{(i)} \mid \boldsymbol{n}, \boldsymbol{\Omega}\right)\right\} + \mathbb{E}\left\{\ln \mathrm{P}\left(\boldsymbol{n} \mid \boldsymbol{\Omega}\right)\right\} \\ &= \sum_{k=1}^{K_A}\left[\sum_{i=1}^{2} \mathbb{E}\left\{\ln \mathrm{P}\left(m_k^{(i)} \mid n_k, \boldsymbol{\Omega}\right)\right\} + \mathbb{E}\left\{\ln \mathrm{P}\left(n_k \mid \boldsymbol{\Omega}\right)\right\}\right] \\ &= \sum_{k=1}^{K_A}\left[\sum_{i=1}^{2}\sum_{j=1}^{2} \mathbb{E}\left\{\mathbb{I}_j(n_k)\right\} \ln \mathrm{P}\left(m_k^{(i)} \mid n_k{=}j, \boldsymbol{\Omega}\right) + \underbrace{\mathbb{E}\left\{\mathbb{I}_1(n_k)\right\} p_k + \mathbb{E}\left\{\mathbb{I}_2(n_k)\right\}(1{-}p_k)}_{\mathbb{E}\left\{\ln \mathrm{P}\left(n_k \mid \boldsymbol{\Omega}\right)\right\}}\right].\end{aligned} \tag{S9}$$

Note that $m_k^{(i)} \mid n_k, \boldsymbol{\Omega}$ pertains to either the attended or unattended Log-Normal distributions in Eq. (S4) depending on the values of i and n_k. Considering that the n_k's are binary random variables and the expectations are with respect to $\boldsymbol{n} \mid \boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, \widehat{\boldsymbol{\Omega}}^{(\ell)}$, the term $\mathbb{E}\left\{\mathbb{I}_j(n_k)\right\}$ can be computed for $j = 1, 2$ using Bayes' rule and conditional independence as:

$$\begin{aligned}\mathbb{E}\left\{\mathbb{I}_j(n_k)\right\} &= \mathrm{P}\left(n_k{=}j \mid \boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, \widehat{\boldsymbol{\Omega}}^{(\ell)}\right) \\ &= \mathrm{P}\left(n_k{=}j \mid m_k^{(1)}, m_k^{(2)}, \widehat{\boldsymbol{\Omega}}^{(\ell)}\right) \\ &= \frac{\mathrm{P}\left(m_k^{(1)}, m_k^{(2)} \mid n_k{=}j, \widehat{\boldsymbol{\Omega}}^{(\ell)}\right) \mathrm{P}\left(n_k{=}j \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)}{\mathrm{P}\left(m_k^{(1)}, m_k^{(2)} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)} \\ &= \frac{\mathrm{P}\left(m_k^{(1)} \mid n_k{=}j, \widehat{\boldsymbol{\Omega}}^{(\ell)}\right) \mathrm{P}\left(m_k^{(2)} \mid n_k{=}j, \widehat{\boldsymbol{\Omega}}^{(\ell)}\right) \mathrm{P}\left(n_k{=}j \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)}{\sum_{n_k} \mathrm{P}\left(m_k^{(1)} \mid n_k, \widehat{\boldsymbol{\Omega}}^{(\ell)}\right) \mathrm{P}\left(m_k^{(2)} \mid n_k, \widehat{\boldsymbol{\Omega}}^{(\ell)}\right) \mathrm{P}\left(n_k \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)}.\end{aligned} \tag{S10}$$

The parameters of the Log-Normal distributions for $m_k^{(i)} \mid n_k, \widehat{\boldsymbol{\Omega}}^{(\ell)}$ are determined from the estimated $\left(\rho^{(a)}, \mu^{(a)}, \rho^{(u)}, \mu^{(u)}\right)$ in the previous EM iteration, i.e., $\widehat{\boldsymbol{\Omega}}^{(\ell)}$. Also, $\mathrm{P}\left(n_k \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right) = \frac{1}{1+\exp\left(-\hat{z}_k^{(\ell)}\right)}$ in (S10), where $\hat{z}_k^{(\ell)}$ is the estimate of z_k from the previous EM iteration. Note that $\mathbb{E}\left\{\mathbb{I}_1(n_k)\right\} = 1 - \mathbb{E}\left\{\mathbb{I}_2(n_k)\right\}$ as n_k is a binary random variable. Defining $\epsilon_k^{(\ell)} := \mathbb{E}\left\{\mathbb{I}_1(n_k)\right\}$ with the expectation over $n_k \mid m_k^{(1)}, m_k^{(2)}, \widehat{\boldsymbol{\Omega}}^{(\ell)}$, we can conclude the E-step by simplifying $Q\left(\boldsymbol{\Omega} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)$ in Eq. (S8) as:

$$
\begin{aligned}
Q\left(\boldsymbol{\Omega} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right) = \sum_{k=1}^{K_A} \frac{1}{2K_A} \Bigg\{ & -\rho^{(a)} \left[\epsilon_k^{(\ell)} \left(\ln m_k^{(1)} - \mu^{(a)} \right)^2 + \left(1 - \epsilon_k^{(\ell)} \right) \left(\ln m_k^{(2)} - \mu^{(a)} \right)^2 \right] \\
& - \rho^{(u)} \left[\left(1 - \epsilon_k^{(\ell)} \right) \left(\ln m_k^{(1)} - \mu^{(u)} \right)^2 + \epsilon_k^{(\ell)} \left(\ln m_k^{(2)} - \mu^{(u)} \right)^2 \right] \\
& + \ln \rho^{(a)} + \ln \rho^{(u)} \Bigg\} \\
& - \rho^{(a)} \left[\beta_0^{(a)} + 0.5 \left(\mu^{(a)} - \mu_0^{(a)} \right)^2 \right] + \left(\alpha_0^{(a)} - 0.5 \right) \ln \rho^{(a)} \\
& - \rho^{(u)} \left[\beta_0^{(u)} + 0.5 \left(\mu^{(u)} - \mu_0^{(u)} \right)^2 \right] + \left(\alpha_0^{(u)} - 0.5 \right) \ln \rho^{(u)} \\
& + \sum_{k=1}^{K_A} \left\{ \epsilon_k^{(\ell)} p_k + \left(1 - \epsilon_k^{(\ell)} \right) (1 - p_k) - (a_0 + 1.5) \ln \eta_k - \frac{1}{\eta_k} \left[b_0 + 0.5 (z_k - c_0 z_{k-1})^2 \right] \right\} \\
& + \mathsf{const.}
\end{aligned}
\tag{S11}
$$

where the const. term includes all the terms that are independent of $\boldsymbol{\Omega}$.

2.2 The M Step

In the M step, we maximize $Q\left(\boldsymbol{\Omega} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)$ in Eq. (S11) with respect to $\boldsymbol{\Omega}$. The maximizers form the parameter updates for the $(\ell{+}1)^{\text{th}}$ EM iteration. As we observe in Eq. (S11), having $\boldsymbol{n}$ as the latent variables separates the terms in $Q\left(\boldsymbol{\Omega} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)$ depending on the distribution parameters, i.e., $\left(\rho^{(a)}, \mu^{(a)}, \rho^{(u)}, \mu^{(u)}\right)$, and the terms depending on the state-space parameters, i.e., $\boldsymbol{z}$ and $\boldsymbol{\eta}$. The derivation of the update rules for the distribution parameters is straightforward through taking the derivatives of $Q\left(\boldsymbol{\Omega} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)$ and solving for their joint zero-crossings. Consequently, the closed-form formulas for the distribution parameters maximizing $Q\left(\boldsymbol{\Omega} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)$ can be expressed as:

$$
\mu^{(a)*} = \frac{1}{2} \left\{ \mu_0^{(a)} + \frac{1}{K_A} \sum_{k=1}^{K_A} \left[\epsilon_k^{(\ell)} \ln m_k^{(1)} + \left(1 - \epsilon_k^{(\ell)} \right) \ln m_k^{(2)} \right] \right\}, \tag{S12}
$$

$$
\mu^{(u)*} = \frac{1}{2} \left\{ \mu_0^{(u)} + \frac{1}{K_A} \sum_{k=1}^{K_A} \left[\left(1 - \epsilon_k^{(\ell)} \right) \ln m_k^{(1)} + \epsilon_k^{(\ell)} \ln m_k^{(2)} \right] \right\}, \tag{S13}
$$

$$\rho^{(a)^*} = \frac{2K_A\alpha_0^{(a)}}{\sum_{k=1}^{K_A}\left[\epsilon_k^{(\ell)}\left(\ln m_k^{(1)} - \mu^{(a)^*}\right)^2 + \left(1-\epsilon_k^{(\ell)}\right)\left(\ln m_k^{(2)} - \mu^{(a)^*}\right)^2\right] + K_A\left[2\beta_0^{(a)} + \left(\mu^{(a)^*} - \mu_0^{(a)}\right)^2\right]}, \tag{S14}$$

$$\rho^{(u)^*} = \frac{2K_A\alpha_0^{(u)}}{\sum_{k=1}^{K_A}\left[\left(1-\epsilon_k^{(\ell)}\right)\left(\ln m_k^{(1)} - \mu^{(u)^*}\right)^2 + \epsilon_k^{(\ell)}\left(\ln m_k^{(2)} - \mu^{(u)^*}\right)^2\right] + K_A\left[2\beta_0^{(u)} + \left(\mu^{(u)^*} - \mu_0^{(u)}\right)^2\right]}, \tag{S15}$$

where $\left(\rho^{(a)^*}, \mu^{(a)^*}, \rho^{(u)^*}, \mu^{(u)^*}\right)$ will be the updated distribution parameters in $\widehat{\boldsymbol{\Omega}}^{(\ell+1)}$.

The next step is to maximize $Q\left(\boldsymbol{\Omega} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)$ with respect to $\boldsymbol{z}$ and $\boldsymbol{\eta}$. Note that this joint maximization is non-convex in general. Consider the following state-space model with parameters $(\boldsymbol{z'}, \boldsymbol{\eta'})$ and binary observations $\boldsymbol{n'}$.

$$\begin{cases} n'_k \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-z'_k)}\right) \\ z'_k = c_0 z'_{k-1} + w'_k \\ w'_k \sim \mathcal{N}(0, \eta'_k) \\ \eta'_k \sim \text{Inverse-Gamma}\left(a_0, b_0\right) \end{cases} \tag{S16}$$

For the inference problem in (S16), the log-posterior can be expressed as:

$$\underset{\boldsymbol{z'},\boldsymbol{\eta'}}{\arg\max} \ln \mathrm{P}\left(\boldsymbol{z'}, \boldsymbol{\eta'} \mid \boldsymbol{n'}\right) = \underset{\boldsymbol{z'},\boldsymbol{\eta'}}{\arg\max}\left[\ln \mathrm{P}\left(\boldsymbol{\eta'} \mid \boldsymbol{n'}\right) + \mathrm{P}\left(\boldsymbol{z'} \mid \boldsymbol{\eta'}, \boldsymbol{n'}\right)\right]. \tag{S17}$$

If we replace the observations n'_k in (S17) with $\epsilon_k^{(\ell)}$, for $k = 1, 2, \ldots, K_A$, the inference problem becomes equivalent to maximizing $Q\left(\boldsymbol{\Omega} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)$ in (S11) with respect to $\boldsymbol{z}$ and $\boldsymbol{\eta}$.

In (Smith and Brown, 2003; Smith et al., 2004), the inference of the parameters in (S16) has been carried out through the EM algorithm, where in each iteration, a Kalman filtering and smoothing algorithm has been employed together with Gaussian approximations. Similar to (Akram et al., 2016), we refer to this EM algorithm as the inner EM not to confuse it with the EM algorithm we have already adopted, which we call the outer EM hereafter. The basic idea behind the inner EM is to approximate the solutions to (S17) as:

$$\begin{cases} \boldsymbol{\eta'}^* = \arg\max_{\boldsymbol{\eta'}} \mathrm{P}\left(\boldsymbol{\eta'} \mid \boldsymbol{n'}\right) \\ \boldsymbol{z'}^* = \arg\max_{\boldsymbol{z'}} \mathrm{P}\left(\boldsymbol{z'} \mid \boldsymbol{\eta'}^*, \boldsymbol{n'}\right) \end{cases}, \tag{S18}$$

where $\boldsymbol{\eta'}^*$ are estimated through the inner EM with $\boldsymbol{z'}$ as the latent variables, and $\boldsymbol{z'}^*$ are just the result of a Kalman filtering and smoothing algorithm in (S16) for $\boldsymbol{\eta'} = \boldsymbol{\eta'}^*$.

In order to make the inference procedure suitable for real-time implementation, we can avoid the inner EM and instead use crude estimates of $\boldsymbol{\eta'}^*$ in (S18). Note that $\epsilon_k^{(\ell)}$, which acts as the observation n'_k in (S16) for $k = 1, 2, \ldots, K_A$, is equal to $\mathrm{P}\left(n_k = 1 \mid m_k^{(1)}, m_k^{(2)}, \widehat{\boldsymbol{\Omega}}^{(\ell)}\right)$ calculated as in (S10). Assuming that $\epsilon_k^{(\ell)} \approx \mathrm{P}\left(n'_k = 1\right) = \frac{1}{1+\exp(-z'_k)}$, in the ℓ^{th} outer EM iteration, we can consider $\left[\mathrm{logit}\left(\epsilon_k^{(\ell)}\right) - c_0\,\mathrm{logit}\left(\epsilon_{k-1}^{(\ell)}\right)\right]$ as a sample of $\mathcal{N}\left(0, \eta'_k\right)$. Therefore, considering the Inverse-Gamma prior, a crude estimate for η'^*_k can be calculated for $k = 1, 2, \ldots, K_A$ as:

$$\eta'^*_k = \frac{2b_0 + \left[\mathrm{logit}\left(\epsilon_k^{(\ell)}\right) - c_0\,\mathrm{logit}\left(\epsilon_{k-1}^{(\ell)}\right)\right]^2}{2a_0 - 1}. \tag{S19}$$

If K_A is small enough, we can simplify the state-space model of (S16) by assuming a single variance, i.e., $\eta' = \eta'_k$ for $k = 1, 2, \ldots, K_A$, and using an estimate similar to (S19) for η'^*. However, in this model, the crude estimate would be more reliable as it is based on K_A samples rather than a single sample. Considering a normalized log-likelihood and the same Inverse-Gamma prior on η', the estimate for η'^* can be computed as:

$$\eta'^* = \frac{2b_0 + \frac{1}{K_A}\sum_{k=1}^{K_A}\left[\mathrm{logit}\left(\epsilon_k^{(\ell)}\right) - c_0\,\mathrm{logit}\left(\epsilon_{k-1}^{(\ell)}\right)\right]^2}{2a_0 - 1}. \tag{S20}$$

After estimating η'^*_k in (S19) for $k = 1, 2, \ldots, K_A$, or η'^* in (S20), we can proceed as before to estimate $\boldsymbol{z'}^*$, i.e., using a Kalman filtering and smoothing algorithm with Gaussian approximations to estimate $\boldsymbol{z'}^*$ in (S18). These estimates, namely $\boldsymbol{z}^*$ and $\boldsymbol{\eta}^*$, form approximate solutions for $\boldsymbol{z}$ and $\boldsymbol{\eta}$ in the original problem of maximizing $Q(\boldsymbol{\Omega} \mid \widehat{\boldsymbol{\Omega}}^{(\ell)})$ in (S11) with respect to the state-space parameters.

Next, we discuss the details of the inner EM algorithm, as in (Akram et al., 2016), used to solve for $\boldsymbol{z}'$ and $\boldsymbol{\eta}'$ in (S16). As mentioned before, the idea is to use an EM algorithm together with Gaussian approximations to maximize $\mathrm{P}\left(\boldsymbol{\eta}' \mid \boldsymbol{n}'\right)$, and then maximize the likelihood of $\boldsymbol{z'}$ with respect to the observations and estimated variances. Considering $\boldsymbol{z'}$ as the latent variables, the surrogate function $Q(\boldsymbol{\eta}' \mid \widehat{\boldsymbol{\eta}'}^{(\ell)})$ at ℓ^{th} EM iteration is calculated as:

$$\begin{aligned} Q\left(\boldsymbol{\eta}' \middle| \widehat{\boldsymbol{\eta}'}^{(\ell)}\right) &= \mathbb{E}\left\{\ln \mathrm{P}\left(\boldsymbol{n}', \boldsymbol{z}' \mid \boldsymbol{\eta}'\right)\right\} + \ln \mathrm{P}(\boldsymbol{\eta}') \\ &= \sum_{k=1}^{K_A}\left[\frac{\mathbb{E}\left\{(z'_k - c_0 z'_{k-1})^2\right\} + 2b_0}{2\eta'_k} + (a_0 + 1.5)\ln \eta'_k\right] + \mathsf{const.}, \end{aligned} \tag{S21}$$

where the expectations are with respect to $\boldsymbol{z}' \mid \boldsymbol{n}', \widehat{\boldsymbol{\eta}'}^{(\ell)}$, and the const. term contains all the terms that are independent of $\boldsymbol{\eta}'$.

In the M-step of the inner EM algorithm, $Q\left(\boldsymbol{\eta}' \middle| \widehat{\boldsymbol{\eta}'}^{(\ell)}\right)$ is maximized with respect to $\boldsymbol{\eta}'$ to calculate the updated variances for the next EM iteration. Taking the derivative of (S21) with respect to $\boldsymbol{\eta}'$ and equating it to zero results in the following update rule for $\widehat{\boldsymbol{\eta}'}^{(\ell+1)}$:

$$\begin{aligned}
\widehat{\eta'}_k^{(\ell+1)} &= \frac{1}{2a_0+3}\left[\mathbb{E}\left\{(z'_k - c_0 z'_{k-1})^2\right\} + 2b_0\right] \\
&= \frac{1}{2a_0+3}\left[\mathbb{E}\left\{{z'_k}^2\right\} + c_0^2\mathbb{E}\left\{{z'_{k-1}}^2\right\} - 2c_0\mathbb{E}\left\{z'_k z'_{k-1}\right\} + 2b_0\right] \\
&= \frac{1}{2a_0+3}\Big[\sigma^2_{k|K_A} + \bar{z}^2_{k|K_A} + c_0^2\sigma^2_{k-1|K_A} + c_0^2\bar{z}^2_{k-1|K_A} - 2c_0\sigma^2_{k,k-1|K_A} \\
&\qquad - 2c_0\bar{z}_{k|K_A}\bar{z}_{k-1|K_A} + 2b_0\Big],
\end{aligned} \tag{S22}$$

where the parameters $\bar{z}_{k|K_A}$ and $\sigma^2_{k|K_A}$ in Eq. (S22) are respectively the mean and the variance of $z'_k \mid \boldsymbol{n}', \widehat{\boldsymbol{\eta}'}^{(\ell)}$.

If we consider the Gaussian approximation $\mathcal{N}\left(\bar{z}_{k_1|k_2}, \sigma^2_{k_1|k_2}\right)$ to the density $z'_{k_1} \,|\, n'_{1:k_2}, \widehat{\boldsymbol{\eta}'}^{(\ell)}$ for $1 \le k_1 \le k_2 \le K_A$, these parameters can be computed in a forward and backward pass similar to the conventional Kalman filtering and smoothing algorithms. The corresponding filtering equations for $1 \le k \le K_A$ are summarized as:

$$\begin{cases}
\bar{z}_{k|k-1} = c_0\bar{z}_{k-1|k-1} \\
\sigma^2_{k|k-1} = c_0^2\sigma^2_{k-1|k-1} + {\eta'_k}^{(l)} \\
\bar{z}_{k|k} = \bar{z}_{k|k-1} + \sigma^2_{k|k-1}\left[n'_k - \frac{\exp(\bar{z}_{k|k})}{1+\exp(\bar{z}_{k|k})}\right] \\
\sigma^2_{k|k} = \left[\frac{1}{\sigma^2_{k|k-1}} + \frac{\exp(\bar{z}_{k|k})}{\left(1+\exp(\bar{z}_{k|k})\right)^2}\right]^{-1}
\end{cases} \tag{S23}$$

Note that the third equation in (S23) is a non-linear equation whose solution can be approximated through standard approaches such as the Newton's method. The last two equations in (S23) come from the Gaussian approximation: assuming that $z'_{k-1} \mid n'_{1:k-1}, \widehat{\boldsymbol{\eta}'}^{(\ell)} \backsim \mathcal{N}\left(\bar{z}_{k-1|k-1}, \sigma^2_{k-1|k-1}\right)$ we calculate the Gaussian approximation for $z'_k \mid n'_{1:k}, \widehat{\boldsymbol{\eta}'}^{(\ell)}$. The mean of the Gaussian approximation $\bar{z}_{k|k}$ is calculated as the mode of $\ln \mathrm{P}\left(z'_k \mid n'_{1:k}, \widehat{\boldsymbol{\eta}'}^{(\ell)}\right)$, and its variance $\sigma^2_{k|k}$ is computed as the negative inverse Hessian of $\ln \mathrm{P}\left(z'_k \mid n'_{1:k}, \widehat{\boldsymbol{\eta}'}^{(\ell)}\right)$ evaluated at the estimated mean $\bar{z}_{k|k}$ (Tanner, 1991). The smoothing equations are the same as those used for fixed interval smoothing. Therefore, for $1 \le k \le K_A - 1$, we have:

$$\begin{cases} s_k = \sigma^2_{k|k} \Big/ \sigma^2_{k+1|k} \\ \bar{z}_{k|K_A} = \bar{z}_{k|k} + s_k \left(\bar{z}_{k+1|K_A} - \bar{z}_{k+1|k}\right) \\ \sigma^2_{k|K_A} = \sigma^2_{k|k} + s_k^2 \left(\sigma^2_{k+1|K_A} - \sigma^2_{k+1|k}\right) \end{cases} \tag{S24}$$

The $\sigma^2_{k,k-1|K_A}$ term in (S22) is a lagged covariance term that can be computed using the covariance smoothing algorithm (De Jong and Mackinnon, 1988):

$$\sigma^2_{k,k-1|K_A} = \mathrm{Cov}\left\{z'_k, z'_{k-1} \mid \boldsymbol{n'}, \widehat{\boldsymbol{\eta'}}^{(\ell)}\right\} = \frac{\sigma^2_{k-1|k-1}\sigma^2_{k|K_A}}{\sigma^2_{k|k-1}}. \tag{S25}$$

Having calculated the variances $\boldsymbol{\eta'}^*$ from the inner EM algorithm, $\boldsymbol{z'}^*$ can be estimated using a single forward and backward pass for $\boldsymbol{\eta'} = \boldsymbol{\eta'}^*$, similar to that used in the inner EM algorithm. In summary, we have transformed the problem of maximizing (S11) with respect to $\boldsymbol{z}$ and $\boldsymbol{\eta}$ into inferring $\boldsymbol{z'}$ and $\boldsymbol{\eta'}$ in (S16) by identifying n'_k with $\epsilon_k^{(l)}$ for $k = 1, \ldots, K_A$. We have then solved the latter problem through an EM algorithm combined with Gaussian approximations and Kalman filtering and smoothing. Therefore, we have $\boldsymbol{z}^* = \boldsymbol{z'}^*$ and $\boldsymbol{\eta}^* = \boldsymbol{\eta'}^*$ in the original problem.

Algorithm 2 Parameter Estimation in Dynamic State-Space Model

Input: $m^{(1)}_{1:K_A}, m^{(2)}_{1:K_A}, \alpha_0^{(a)}, \alpha_0^{(u)}, \beta_0^{(a)}, \beta_0^{(u)}, \mu_0^{(a)}, \mu_0^{(u)}, a_0, b_0, \mathcal{S}_{EM}$

Output: $\widehat{\boldsymbol{\Omega}} = \left\{\hat{z}_{1:K_A}, \hat{\eta}_{1:K_A}, \hat{\rho}^{(a)}, \hat{\mu}^{(a)}, \hat{\rho}^{(u)}, \hat{\mu}^{(u)}\right\}$

Set $\widehat{\boldsymbol{\Omega}}^{(0)}$ as the initialization for state-space model parameter set based on estimates in the previous instance
$\ell = 0$
while $\neg\mathcal{S}_{EM}$ **do**
    calculate $\epsilon^{(\ell)}_{1:K_A}$ using (S10)
    update the parameters of the Log-Normal distributions, i.e., $\mu^{(a)}$, $\mu^{(u)}$, $\rho^{(a)}$, $\rho^{(u)}$, based on equations (S12), (S13), (S14), and (S15) respectively
    update the state-space variances, i.e., $\eta_{1:K_A}$, using the inner-EM algorithm or the crude estimates in equations (S19) and (S20)
    update the hidden states in the state-space model, i.e., $z_{1:K_A}$, using a Kalman filtering and smoothing algorithm with Gaussian approximations
    set $\widehat{\boldsymbol{\Omega}}^{(\ell+1)}$ as the updated parameter set including the updated distribution parameters, variances, and hidden states in the state-space model
    $\ell \leftarrow \ell + 1$
end while
$\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}^{(\ell)}$.

Algorithm 2 summarizes the overall inference procedure within a fixed-lag window of length K_A. Going back to Fig. S1, copied from the paper, we assume $k = k_0$ is the current instance and the

goal is to infer the attentional state at instance $k = k_0 - K_F$ based on the attention markers within the window indexed from 1 to K_A, given by $m_k^{(i)}$ for $i = 1, 2$ and $k = 1, \ldots, K_A$. We initialize the state-space model parameter set $\mathbf{\Omega}$ using the estimates at the previous instance, and the output of Algorithm 2, i.e., $\widehat{\mathbf{\Omega}}$, is used for initialization in the next instance. Defining $f(.)$ as the sigmoid function, $f\left(\hat{z}_{K_A-K_F}\right)$ determines the estimated probability of attending to speaker 1 at $k = k_0 - K_F$, and $\left[f\left(\hat{z}_{K_A-K_F} - 1.65\hat{\sigma}^2_{K_A-K_F|K_A}\right), f\left(\hat{z}_{K_A-K_F} + 1.65\hat{\sigma}^2_{K_A-K_F|K_A}\right)\right]$ represents the 90% confidence intervals of this estimate, where $\hat{\sigma}^2_{K_A-K_F|K_A}$ represents the inferred variance of $\hat{z}_{K_A-K_F}$ calculated through the discussed Gaussian approximations. The parameter $\mathcal{S}_{EM}$ in Algorithm 2 is a stopping condition for the outer EM, which can be a limit on the number of iterations.

3 SMOOTHING EFFECT OF STATE-SPACE MODELING

In this section, we discuss the smoothing effect of the proposed state-space estimation and compare it with that of sliding Gaussian kernel smoothers. Recall that the Inverse-Gamma conjugate prior on η_k's in Eq. (S3) controls the degree of smoothing in the state-space model. If this prior favors smaller values of η_k's, the consecutive changes in z_k's and thereby p_k's will be smaller, which results in a larger smoothing effect. We tune the Inverse-Gamma prior through the hyperparameters a_0 and b_0 as in Eq. (S3) to match the auditory attention dynamics. Therefore, we expect that the corresponding smoothing effect will make the state-space estimates robust to the stochastic fluctuations in the attention markers, while capturing the attention switching instances with a small transition delay.

Fig. S2-A shows the output of the correlation-based attention marker in Case 2 of the simulation study in the main manuscript (row D of Fig. 4). The output of the real-time estimator with 1.5 s forward-lag (as in row F of Fig. 4) is shown in Fig. S2-B. We also consider two *non-causal* Gaussian kernel smoothers with the same delay of 1.5 s for fairness of comparison. Fig. S2-C and S2-E show the attention markers of Fig. S2-A convolved with the two Gaussian kernels, respectively. The two kernels are shown as insets in Fig. S2-C and S2-D. Gaussian kernel 1 in Fig. S2-C favors the current values of the attention marker while Gaussian kernel 2 in Fig. S2-D gives more weight to its future values.

Both kernels provide a clearer picture of the attentional state by smoothing out the stochastic fluctuations of Fig. S2-A. However, unlike the output of the state-space estimator, they do not provide statistically interpretable results. First, based on Figs. S2-C and -D, we can only obtain a binary decision on the attended speaker at each instance. The state-space estimates, however, provide a probabilistic measure of the attentional state as shown in Fig. S2-B, together with statistical confidence intervals. The red arrows in Fig. S2-C and -D mark instances where strong fluctuations in the attention markers result in misclassification. For instance, the smoothed markers with kernel 2 imply an attention switch earlier than the 30 s mark (upward arrow, Fig. S2-D). Such abrupt classification errors could be undesirable for applications such as BCI systems or smart hearing aids, as the devices need to modify their settings back and forth in a small time period. The state-space model prevents these instances of misclassification, thanks to the confidence intervals of the estimated p_k's (the middle arrows) which help rule out such false alarm events.

4 ENCODING MODEL SIMULATION

This section provides a simulated example to motivate our MEG analysis, in which we use an encoding model and take the M100 component of the Temporal Response Function (TRF) as the attention marker.

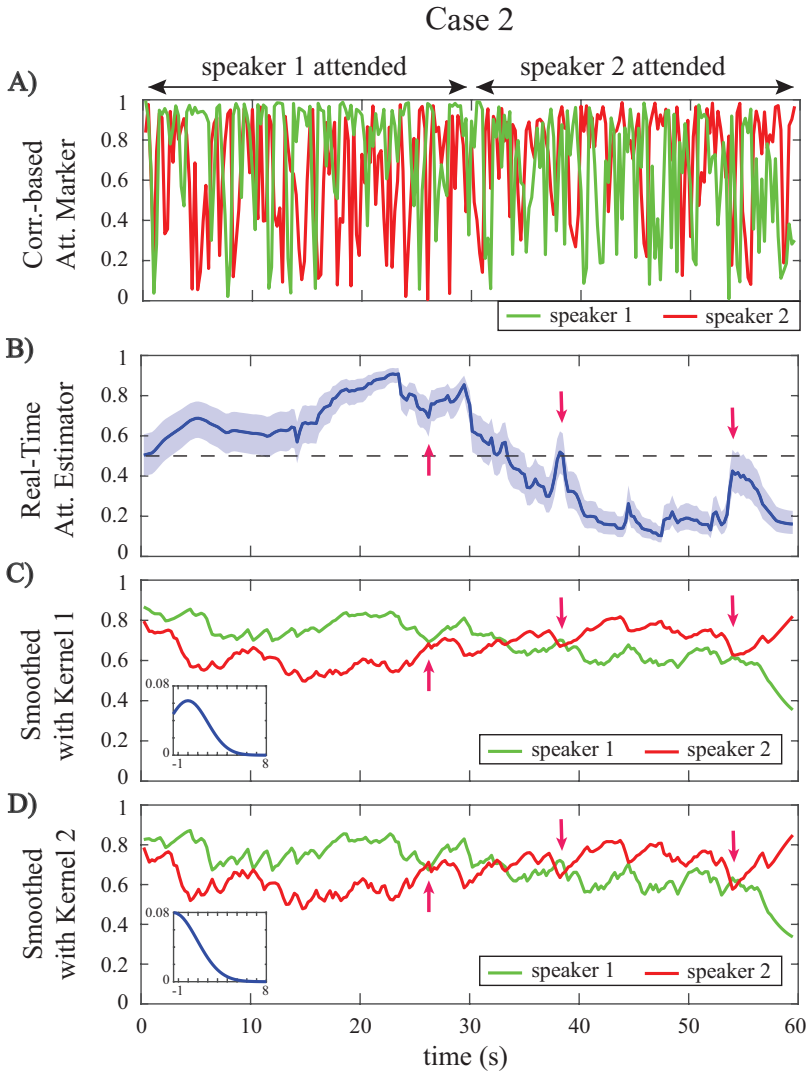

Figure S2. Smoothing effect of the state-space model in comparison to simple kernel smoothers: A) Output of the correlation-base attention marker corresponding to Case 2 of the simulation study in the main manuscript. B) Real-time estimator with 1.5 s forward-lag. C) Convolution of the correlation-based attention marker with Gaussian kernel 1 (shown as inset). D) Convolution of the correlation-based attention marker with Gaussian kernel 2 (shown as inset).

4.1 Simulation Settings

Consider the following generative model:

$$e_t = s_t^{(1)} * \tau_t^{(1)} + s_t^{(2)} * \tau_t^{(2)} + \mu + n_t, \tag{S26}$$

where e_t, $s_t^{(1)}$, and $s_t^{(2)}$ respectively denote the auditory component of the neural response, speech envelope for speaker 1, and speech envelope for speaker 2. We have used the same speech signals for $s_t^{(1)}$ and $s_t^{(2)}$ as in the EEG simulation, with the same sampling rate of $f_s = 200$ Hz. In the context of MEG processing, $\tau_t^{(1)}$ and $\tau_t^{(2)}$ are referred to as the TRF for speakers 1 and 2. We have set $\mu = 0.001$ as the unknown constant

mean and $n_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 2.5\times10^{-7})$ as the observation noise. We assume an attention modulation effect on the M100 component of the TRFs.

Figure S3 shows two cases for the TRFs $\tau_t^{(1)}$ and $\tau_t^{(2)}$: In the left panels (case 1), there is a strong attention modulation effect on the M100 components, and in the right panels (case 2), this effect is weakened. In both cases, the attention is on speaker 1 during the $[0, 30)$ s interval and on speaker 2 during the $(30, 60]$ s interval. Also, we have considered a length of 0.4 s for the TRFs. Row B in Fig. S3 shows examples of the attended and the unattended TRFs for each of the two cases. In case 1, there is a large difference between the magnitude of the M100 components in the attended and the unattended TRFs, while in case 2, this difference is small compared to our estimation accuracy. We have also considered three higher latency components in the TRFs which are not modulated by the attentional state, similar to the M50 component. As shown in row A of Fig. S3, a zero-mean Gaussian i.i.d. noise is added to the TRF components as well. Note that similar to the EEG simulation, we have used a Gaussian kernel with the standard deviation of 10 ms to smooth the TRFs. This smoothness property is also observed in TRFs estimated from experimentally-recorded MEG signals (Ding and Simon, 2012a,b).

4.2 Parameter Selection

For the encoder estimation parameters in Algorithm 1, we have considered consecutive non-overlapping windows of length 0.25 s, i.e., $W = 50$, resulting in $K = 240$ instances, and we have assumed the same 0.4 s length for the TRFs, i.e., $L_e = 80$. We have chosen $\gamma = 0.005$ through cross-validation and $\lambda = 0.9167$, which results in an *effective* window length of 3 s for encoder estimation. Considering the smoothing Gaussian kernel used in the forward model, we have used the Gaussian dictionary matrix $\mathbf{G}_0 \in \mathbb{R}^{(L_e+1)\times(L_e+1)}$ for each speaker in the encoder estimation step to enforce smoothness in the TRFs. The dictionary columns consist of overlapping Gaussian kernels with the standard deviation of 10 ms, whose means cover the 0 s to 0.4 s lag with $T_s = 5$ ms increments. As a result, considering the simultaneous estimation of the two TRFs, the overall dictionary matrix would be $\mathbf{G} = \text{diag}\,(1, \mathbf{G_0}, \mathbf{G_0})$.

We have used the FASTA package (Goldstein et al., 2014) with Nesterov's acceleration method to implement the forward-backward splitting algorithm. All the prior distribution parameters of the state-space models are set similar to the EEG simulation in the paper, where $a_0 = 2.008$, $b_0 = 0.2016$, and the prior parameters for the attended and unattended distributions were tuned based on a separate 15 s sample trial. For the real-time state-space estimator, we have used a sliding window of length 15 s with a fixed forward-lag of 1.5 s, i.e., $K_A = \lfloor 15 f_s/W \rfloor$ and $K_F = \lfloor 1.5 f_s/W \rfloor$. The sample trial for tuning the distribution parameters can be thought of as an initialization step for the estimator prior to its real-time application.

4.3 Estimation Results

Figure S4 shows the results of our estimation framework. Row A contains the estimated TRFs for the encoding model. The major components of the TRFs are retrieved in the estimates while the ℓ_1-norm penalty in Eq. (S1) has significantly denoised these components as compared with the original noisy versions in row A of Fig. S3. Row B in Fig. S4 displays the extracted magnitudes of the M100 components from the estimated TRFs at each instance. The attention marker in this case is defined as the magnitude of the M100 component, where the M100 component is calculated as the minimum value of the TRF estimate around the 100 ms lag. Notice that there is a significant statistical difference between the extracted M100 components for the attended and unattended speakers in case 1, while the estimated M100 components are highly variable in case 2 and do not show a strong attention modulation effect.

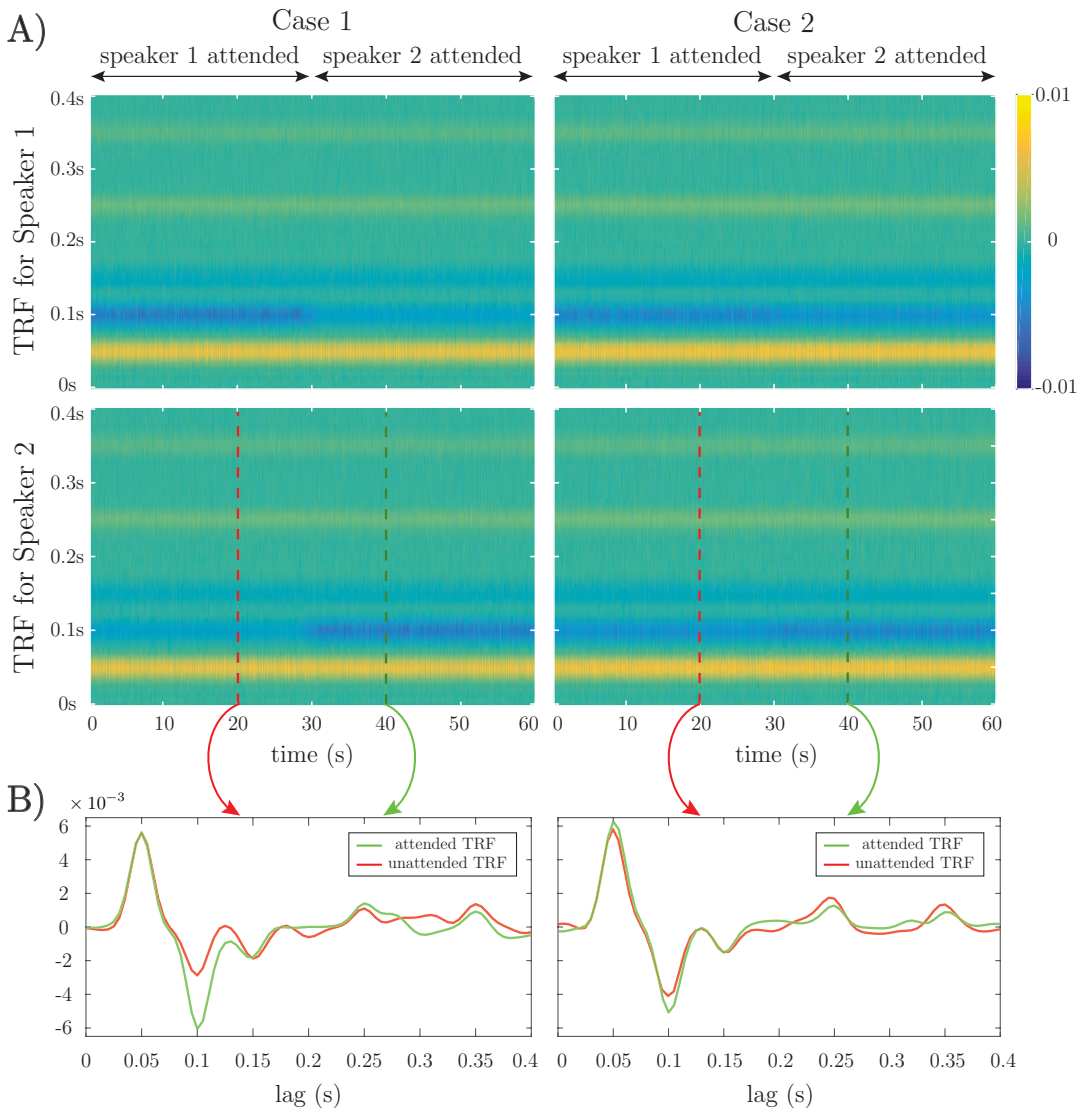

Figure S3. The TRFs $\tau_t^{(1)}$ and $\tau_t^{(2)}$ used for the simulation model in Eq. (S26). A) TRFs for case 1 (strong modulation in M100 components) and case 2 (weak modulation in M100 components). B) Snapshots of the attended and unattended TRFs for the two cases.

Rows C and D of Fig. S4 show the output of the batch-mode and real-time state-space estimators, respectively. In case 1, both the batch-mode and real-time estimators perform well in tracking the attentional state. Note that the sharp drop of the attention probability near $\sim 30\,\mathrm{s}$ in Row D is due to the fact that at each instance the real-time estimator does not observe the attention markers beyond the $1.5\,\mathrm{s}$ forward lag, whereas the batch-mode estimator estimates the probabilities given the entire trial. In case 2, the batch-mode estimator performs well even though the M100 components are not visually indicative of the attentional state. However, the classification confidence decreases considerably specially in the $(30, 60]$ s interval. The real-time estimator in case 2 closely follows the batch-mode estimator, but is more sensitive to the fluctuations of the extracted M100 components. Thus, its performance undergoes further degradation going from case 1 to 2, as compared with that of the batch-mode estimator. The red arrows in rows C and D of case 2 in Fig. S4 mark instances where the less robustness of real-time estimator resulted in misclassifications, while the batch-mode estimator classified the attended speaker correctly.

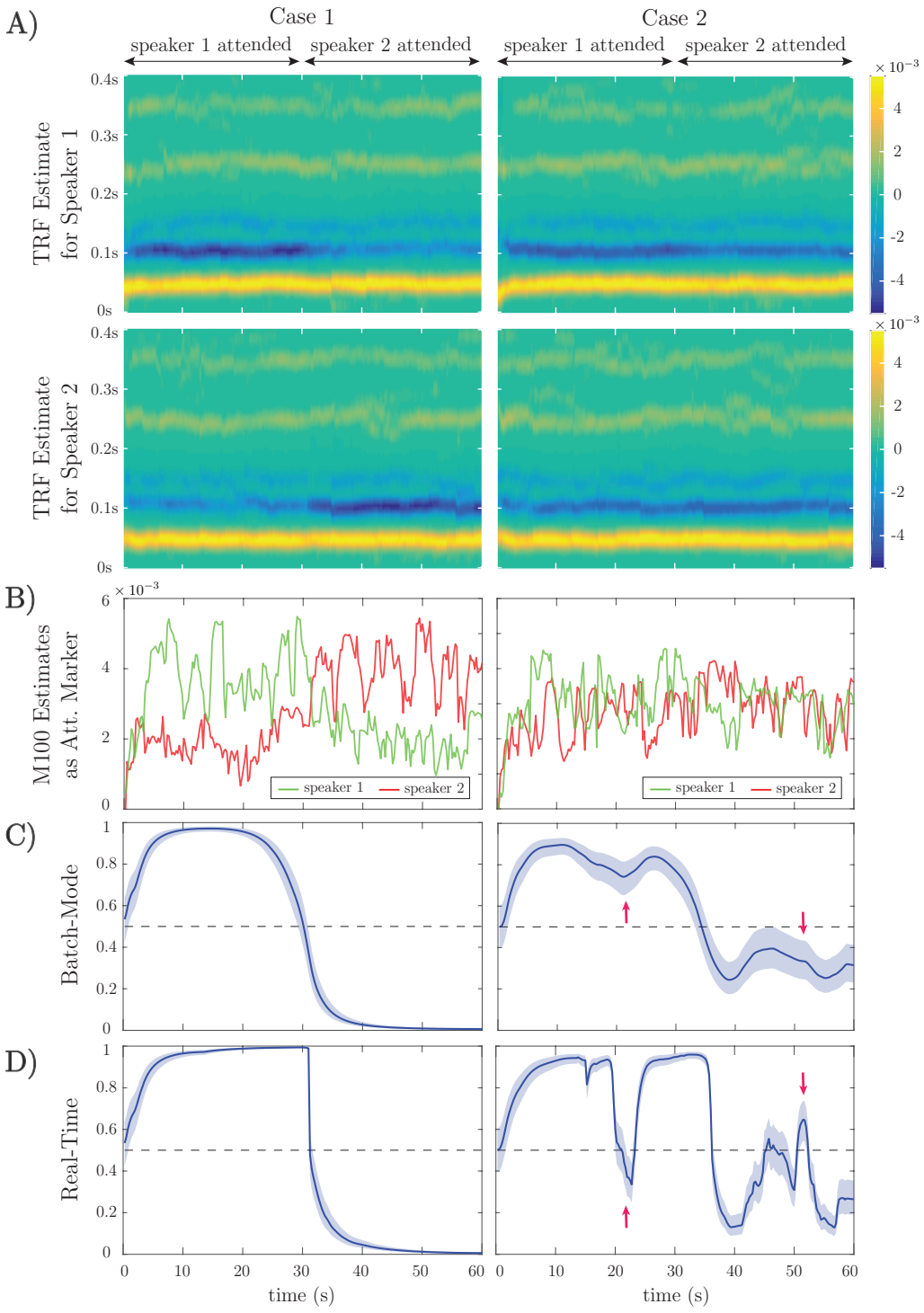

Figure S4. Estimation results of application to simulated MEG data: A) Estimated TRFs for case 1 (strong modulation in M100 components) and case 2 (weak modulation in M100 components). B) Estimated M100 magnitudes as the attention markers. C) Outputs of the batch-mode estimator as the estimated probability of attending to speaker 1. D) Outputs of the real-time estimator as the estimated probability of attending to speaker 1. The real-time estimator is less robust to the statistical fluctuations in the extracted M100 components, which can result in misclassifications as shown for two example instances marker by red arrows. However, it follows the general trend of the batch-mode estimator closely despite its online access to data.

It is worth noting that as we are using an encoding model in this case, the overall delay in estimating the attentional state is the forward-lag window, i.e., 1.5 s, and unlike the case of using the decoding model, the encoder lag does not contribute to the delay. Our analysis of the effect of K_F on the MSE of the real-time estimator with respect to the batch-mode was nearly identical to that presented for the EEG simulation, and is thus omitted for brevity.

REFERENCES

Akram, S., Presacco, A., Simon, J. Z., Shamma, S. A., and Babadi, B. (2016). Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *NeuroImage* 124, 906–917

Akram, S., Simon, J. Z., and Babadi, B. (2017). Dynamic estimation of the auditory temporal response function from MEG in competing-speaker environments. *IEEE Transactions on Biomedical Engineering* 64, 1896–1905

Babadi, B., Kalouptsidis, N., and Tarokh, V. (2010). SPARLS: The sparse RLS algorithm. *IEEE Transactions on Signal Processing* 58, 4013–4025

Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering* (Springer New York). 185–212

De Jong, P. and Mackinnon, M. J. (1988). Covariances for smoothed estimates in state space models. *Biometrika* 75, 601–602

Ding, N. and Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences* 109, 11854–11859

Ding, N. and Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology* 107, 78–89

Duchi, J. and Singer, Y. (2009). Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research* 10, 2899–2934

Figueiredo, M. A., Nowak, R. D., and Wright, S. J. (2007). Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing* 1, 586–597

Goldstein, T., Studer, C., and Baraniuk, R. (2014). A field guide to forward-backward splitting with a FASTA implementation. *arXiv eprint* abs/1411.3406

Goldstein, T., Studer, C., and Baraniuk, R. (2015). FASTA: A generalized implementation of forward-backward splitting. `http://arxiv.org/abs/1501.04979`

Jenatton, R., Mairal, J., Bach, F. R., and Obozinski, G. R. (2010). Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 487–494

Sheikhattar, A., Fritz, J. B., Shamma, S. A., and Babadi, B. (2015). Recursive sparse point process regression with application to spectrotemporal receptive field plasticity analysis. *IEEE Transactions on Signal Processing* 64, 2026–2039

Smith, A. C. and Brown, E. N. (2003). Estimating a state-space model from point process observations. *Neural Computation* 15, 965–991

Smith, A. C., Frank, L. M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., et al. (2004). Dynamic analysis of learning in behavioral experiments. *Journal of Neuroscience* 24, 447–461

Tanner, M. A. (1991). *Tools for Statistical Inference*, vol. 3 (Springer)