NeuroImage 124 (2016) 906–917

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling☆

Sahar Akram [a,b,*], Alessandro Presacco [c], Jonathan Z. Simon [a,b,d], Shihab A. Shamma [a,b], Behtash Babadi [a,b,*]

[a] Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA
[b] Institute for Systems Research, University of Maryland, College Park, MD 20742, USA
[c] Department of Hearing and Speech Science, University of Maryland, College Park, MD 20742, USA
[d] Department of Biology, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 17 July 2015
Accepted 21 September 2015
Available online 4 October 2015

Keywords:
Attention
MEG
Speech segregation
State-space models
Nonlinear filtering

ABSTRACT

The underlying mechanism of how the human brain solves the cocktail party problem is largely unknown. Recent neuroimaging studies, however, suggest salient temporal correlations between the auditory neural response and the attended auditory object. Using magnetoencephalography (MEG) recordings of the neural responses of human subjects, we propose a decoding approach for tracking the attentional state while subjects are selectively listening to one of the two speech streams embedded in a competing-speaker environment. We develop a biophysically-inspired state-space model to account for the modulation of the neural response with respect to the attentional state of the listener. The constructed decoder is based on a maximum *a posteriori* (MAP) estimate of the state parameters via the Expectation Maximization (EM) algorithm. Using only the envelope of the two speech streams as covariates, the proposed decoder enables us to track the attentional state of the listener with a temporal resolution of the order of seconds, together with statistical confidence intervals. We evaluate the performance of the proposed model using numerical simulations and experimentally measured evoked MEG responses from the human brain. Our analysis reveals considerable performance gains provided by the state-space model in terms of temporal resolution, computational complexity and decoding accuracy.

Introduction

One of the hallmarks of brain function is the ability to segregate and focus on an auditory object in a complex auditory scene. From a mathematical perspective, this is a highly ill-posed problem; however, our brain is able to solve this problem in a remarkably fast and accurate fashion. It has been hypothesized that after entering the auditory system, the complex auditory signal resulting from sound sources in a crowded environment is decomposed into acoustic features at different stages of the auditory pathway. Then, a rich representation of spectrotemporal features reaches the auditory cortex, where an appropriate binding of the relevant features and discounting of others leads to the perception of an auditory object (Bergman, 1994; Griffiths and Warren, 2004; Fishman and Steinschneider, 2010; Shamma et al., 2011). A compelling example is the Cocktail Party effect (Cherry, 1953; Brungart, 2001; McDermott, 2009), in which a listener is able to attend to an individual speaker in the presence of other competing speakers and to segregate the attended speech from all other sound sources in the environment.

The neural representation of speech as a distinct auditory object has been extensively studied using auditory scenes consisting of pairs of concurrent speech streams mixed into a single acoustic channel with no spatial cues provided. Any neural representation of a single stream of speech (considered as an auditory object) involves complex segregation and grouping processes (Ding and Simon, 2012a,b; Mesgarani and Chang, 2012; O'Sullivan et al., 2014), given the substantial overlaps in spectral and temporal domains. As reported by these studies, concurrent auditory objects — even those with highly overlapping spectrotemporal features — are neurally encoded as a distinct object in the auditory cortex and emerge as fundamental representational units for high-level cognitive processing. In the case of listening to speech, it has recently been demonstrated that the auditory response manifested in magnetoencephalographic recordings is strongly modulated by the spectrotemporal features of the speech (Ding and Simon, 2012b; Pasley et al., 2012). In the presence of two speakers, this modulation appears to be strongly phase-locked to the spectrotemporal features of the attended speaker as opposed to the unattended speaker (See Fig. 1) (Ding and Simon, 2012a; Mesgarani and Chang, 2012).

A widely-used mathematical approach for decoding these cortical modulations is reverse correlation, which can be used to reconstruct the stimulus from the response of the neural population, which then

☆ This work has been presented in part at the 2014 Neural Information Processing Systems (NIPS) Conference (Akram et al., 2014).

* Corresponding authors at: Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA.
E-mail addresses: sakram@umd.edu (S. Akram), behtash@umd.edu (B. Babadi).

http://dx.doi.org/10.1016/j.neuroimage.2015.09.048
1053-8119/

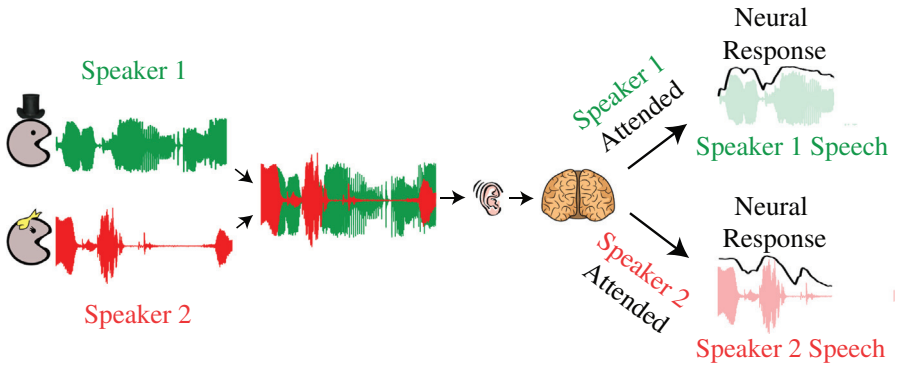

Fig. 1. Schematic depiction of auditory object encoding in the auditory cortex. Here, the auditory scene consists of the mixture of two concurrent speech streams. Recent studies show that cortical activity (black traces) is selectively phased-locked to the temporal envelope of the attended speaker as opposed to the unattended speaker's envelope.

can be compared with the original stimulus to reveal preserved or dismissed features in the population response (Bialek et al., 1991; Gielen et al., 1988; Hesselmans and Johannesma, 1989). Although useful for evaluating data from neural populations using electrocorticography (ECoG) (Mesgarani et al., 2009; Mesgarani and Chang, 2012), MEG (Ding and Simon, 2012a,b) and EEG (O'Sullivan et al., 2014; Mirkovic et al., 2015), this method has a number of limitations. The achievable temporal resolution of the current techniques is of the order of minutes. In a real-world scenario, attention of the listener can switch dynamically from one speaker to another; therefore, an appropriate decoder needs to have a dynamic estimation framework with high temporal resolution in order to capture attention switches in real-time, especially in light of the emergence and rapid growth of brain–computer interface systems. Moreover, these decoders often rely on ad-hoc assumptions and simplifications, which in turn overshadow a reliable statistical interpretation of the data.

In this paper, we overcome the aforementioned limitations by introducing a biophysically-inspired state-space model that accounts for the dynamicity of the attentional state as well as its correlation with MEG observations in a competing-speaker scenario. State-space models are widely used in control engineering for describing the dynamics of the systems under study (Hinrichsen and Pritchard, 2005). These models consist of two components: one relating the observations from a stochastic dynamical system to a set of unobserved state variables (forward model), and the other describing the time evolution of the unobserved states (state dynamics). By combining the forward model and state dynamics in a probabilistic framework, it is possible to obtain accurate estimates of the system parameters, perform prediction, and design control mechanisms. Here, we first utilize a forward model relating the auditory neural response activity to the envelopes of the two speech streams by employing the sparse structure of the auditory response. We then model the attentional state of the listener using a non-stationary Bernoulli process. Finally, we employ von Mises–Fisher circular statistics to form a robust inverse model that accounts for the correlation of the observed neural response activity with respect to the two speech streams. We use the Maximum *a posteriori* (MAP) estimation framework to infer the state-space parameters from the observed data. In particular, we devise a novel application of two nested Expectation–Maximization (EM) algorithms to efficiently solve the MAP problem.

Our proposed model has several advantages over existing methods. First, theoretically speaking, our state-space model is able to preserve dynamics as fast as the sampling resolution. Simulation studies as well as application to experimental data reveal that our model is indeed capable of predicting the attentional state of the listener with a temporal resolution of *seconds*, which is a significant improvement over the state-of-the-art temporal resolution of *minutes*. Second, we only require the envelopes of the two speech streams as covariates, which is a substantial reduction in the dimension of the spectrotemporal feature set used for decoding auditory attention. Finally, our state-space framework provides confidence bounds on the state parameters, which can in turn be used for precise statistical inference procedures such as hypothesis testing. We further provide simulation studies as well as applications of our method on experimentally acquired neural response data. Our analyses reveal the superior performance of the proposed decoder in tracking the attentional state of a listener in a competing-speaker environment, as compared to existing techniques.

Methods

Modeling

We divide our modeling framework into three stages: the forward problem of relating the neural response observations to the temporal features of the attended and unattended speech streams; the attention model which takes into account the dynamics of selective attention; and the inverse problem of decoding the attentional state of the listener given the neural response observations and the temporal features of the two speech streams.

The forward problem: Estimating the temporal response function

Consider a task where the subject is passively listening to a speech stream. Let the discrete-time neural response observation at time t, sensor j, and trial r be denoted by $x_{t,j,r}$, for $t = 1, 2, ..., T, j = 1, 2, ..., M$ and $r = 1, 2, ..., R$. Let the time series $y_{1,r}, y_{2,r}, ..., y_{T,r}$ denote an auditory component of the MEG observations. This component can be obtained through source localization techniques or sensor-space source separation algorithms, and will be referred to hereafter as the neural response (See Section MEG processing and neural source localization). Also, let E_t be the speech envelope of the speaker at time t in dB scale. In a linear model, the neural response is linearly related to the envelope of speech as:

$$y_{t,r} = \tau_t * E_t + v_{t,r}, \tag{1}$$

where τ_t is a linear filter of length L denoted by the temporal response function (TRF), $*$ denotes the convolution operator, and $v_{t,r}$ is a nuisance component accounting for trial-dependent and stimulus-independent components manifested in the neural response. It is known that the TRF is a sparse filter, with significant components analogous to the M50 and M100 auditory responses (Ding and Simon, 2012a,b). A commonly-used technique for estimating the TRF is known as Boosting (David et al., 2007; Ding and Simon, 2012b), where the components of the TRF are greedily selected to decrease the mean square error (MSE) of the fit to the neural response. We employ an alternative estimation

framework based on ℓ_1-regularization. Let $\boldsymbol{\tau} := [\tau_L, \tau_{L-1}, ..., \tau_1]'$ be the time-reversed version of the TRF filter in vector form, and let $\mathbf{E}_t := [E_t, E_{t-1}, ..., E_{t-L+1}]'$. In order to obtain a sparse estimate of the TRF, we seek the ℓ_1-regularized estimate:

$$\hat{\boldsymbol{\tau}} = \underset{\tau}{\operatorname{argmin}} \sum_{r,t=1}^{R,T} \|y_{t,r} - \boldsymbol{\tau}'\mathbf{E}_t\|_2^2 + \gamma\|\boldsymbol{\tau}\|_1, \tag{2}$$

where γ is the regularization parameter. The above problem can be solved using standard optimization software. We use a fast solver based on iteratively re-weighted least squares (Ba et al., 2014). The parameter γ is chosen by two-fold cross-validation, where the first half of the data is used for estimating τ and the second half is used to evaluate the goodness-of-fit in the MSE sense. In a competing-speaker environment, where the subjects are only attending to one of the two speakers, the linear model takes the form:

$$y_{t,r} = \tau_t^a * E_t^a + \tau_t^u * E_t^u + v_{t,r}, \tag{3}$$

with τ_t^a, E_t^a, τ_t^u, and E_t^u, denoting the TRF and envelope of the attended and unattended speakers, respectively. The above estimation framework can be generalized to the two-speaker case by replacing the regressor $\tau' E_t$ with $\boldsymbol{\tau}^{a\prime}\mathbf{E}_t^a + \boldsymbol{\tau}^{u\prime}\mathbf{E}_t^u$, where $\boldsymbol{\tau}^a$, $\mathbf{E}_t^a$, $\boldsymbol{\tau}^u$, and $\mathbf{E}_t^u$ are defined in a fashion similar to the single-speaker case. Similarly, the regularization $\gamma\|\boldsymbol{\tau}\|_1$ is replaced by $\gamma^a\|\boldsymbol{\tau}^a\|_1 + \gamma^u\|\boldsymbol{\tau}^u\|_1$.

Selective attention: A non-stationary Bernoulli process

Suppose that at each window of observation, the subject is attending to either of the two speakers. Let $n_{k,r}$ be a binary variable denoting the attention state of the subject at window k and trial r:

$$n_{k,r} = \begin{cases} 1 & \text{attending to speaker 1} \\ 0 & \text{attending to speaker 2} \end{cases} \tag{4}$$

The subjective experience of attending to a specific speech stream among a number of competing speeches reveals that the attention may switch to a competing speaker, although not intended so by the listener. Therefore, we model the statistics of $n_{k,r}$ by a Bernoulli process with a success probability of p_k:

$$P(n_{k,r}|p_k) = p_k^{n_{k,r}}(1-p_k)^{1-n_{k,r}}. \tag{5}$$

A value of p_k close to 1 (respectively 0) implies attention to speaker 1 (respectively 2). The process $\{p_k\}_{k=1}^K$ is assumed to be common among different trials. In order to model the dynamics of p_k, we perform a change of variables by defining z_k such that

$$p_k = \text{logit}^{-1}(z_k) := \frac{\exp(z_k)}{1+\exp(z_k)}. \tag{6}$$

Note that z_k and p_k have a one-to-one monotonic relation, i.e., when z_k varies from $-\infty$ to ∞, p_k monotonically varies from 0 to 1. Hence, instead of working with p_k with a restricted range, we impose dynamics on z_k which admits a larger class of widely-used linear dynamic models. To this end, we employ a first-order autoregressive model of the form:

$$z_k = z_{k-1} + w_k, \tag{7}$$

where w_k is an uncertainty parameter. The autoregressive model in Eq. (7) implies that the parameter z_k at time k is equal to z_{k-1} at time $k-1$ up to some uncertainty which is modeled by a random variable w_k. Since the range of z_k is symmetric around zero, we assume that the uncertainty parameters $\{w_k\}_{k=1,2,\cdots}$ follow centered independent Gaussian distributions with unknown variances $\{\eta_k\}_{k=1,2,\cdots}$. If need be, higher order autoregressive processes can be used to model the dynamics of z_k as well as non-Gaussian distributions to capture the uncertainty w_k. However, our simulation studies as well as the analysis of real data suggest that it is not necessary to go beyond the first-order model and Gaussian uncertainty parameters for the problem at hand. We further assume that η_k are distributed according to the conjugate prior given by the inverse-Gamma distribution with hyper-parameters α (shape) and β (scale).

The inverse problem: Decoding attentional modulation

Let $y_{1,r}, y_{2,r}, ..., y_{T,r}$ denote the neural response time series at trial r, for $r = 1, 2, ..., R$ during an observation period of length T. For a window length W, let

$$\mathbf{y}_{k,r} := \left[y_{(k-1)W+1,r}, y_{(k-1)W+2,r}, \cdots, y_{kW,r}\right], \tag{8}$$

for $k = 1, 2, ..., K := \lfloor T/W \rfloor$. Also, let $E_{i,t}$ be the speech envelope of speaker i at time t in dB scale, $i = 1, 2$. We extract the envelope of the speech signal by taking the absolute value of its analytic extension (Hilbert Transform) and low-pass filter with a cut-off frequency of 20 Hz to obtain a smoothed envelope. Let τ_t^a and τ_t^u denote the TRFs of the attended and unattended speakers, respectively. The neural response predictors in the linear model are given by:

$$\begin{cases} e_{1,t} := \tau_t^a * E_{1,t} + \tau_t^u * E_{2,t}, & \text{attending to speaker 1} \\ e_{2,t} := \tau_t^a * E_{2,t} + \tau_t^u * E_{1,t}, & \text{attending to speaker 2} \end{cases}, \quad t = 1, 2, \cdots, T. \tag{9}$$

Let

$$\mathbf{e}_{i,k} := \left[e_{i,(k-1)W+1}, e_{i,(k-1)W+2}, \ldots, e_{i,kW}\right], \tag{10}$$

for $i = 1, 2$ and $k = 1, 2, ..., K$. Recent work by (Ding and Simon, 2012a) suggests that the neural response y_k is more correlated with the predictor $\mathbf{e}_{i,k}$ when the subject is attending to the ith speaker at window k. Let

$$\theta_{i,k,r} := \arccos\left(\left\langle \frac{\mathbf{y}_{k,r}}{\|\mathbf{y}_{k,r}\|_2}, \frac{\mathbf{e}_{i,k}}{\|\mathbf{e}_{i,k}\|_2} \right\rangle\right) \tag{11}$$

denote the empirical correlation between the observed neural response and the model prediction when attending to speaker i at window k and trial r. When $\theta_{i,k,r}$ is close to 0 (respectively π), the neural response and its predicted value are highly (respectively poorly) correlated. Inspired by the findings of (Ding and Simon, 2012a), we model the statistics of $\theta_{i,k,r}$ by the von Mises–Fisher distribution (Fisher, 1993) with density:

$$p(\theta_{i,k,r}) = \frac{2\kappa_i^{W/2-1}}{(2\pi)^{W/2} I_{W/2-1}(\kappa_i)} \exp(\kappa_i \cos(\theta_{i,k,r})), \quad \theta_{i,k,r} \in [0,\pi], \quad i = 1, 2 \tag{12}$$

where $I_W(\cdot)$ is the W^{th} order modified Bessel function of the first kind, and κ_i denotes the spread parameter of the von Mises–Fisher distribution for $i = 1, 2$. Note that the extra normalization factor of 2 in the numerator is due to the restriction of $\theta_{i,k,r}$ to $[0, \pi]$. The von Mises–Fisher distribution gives more (respectively less) weight to higher (respectively lower) values of correlation between the neural response and its predictor. The spread parameter κ_i accounts for the concentration of $\theta_{i,k,r}$ around 0. Fig. 2 shows a schematic depiction of the von Mises–Fisher statistics in modeling the correlation of the neural response with its predictors based on speech envelopes. We assume a conjugate prior of the form $p(\kappa_i) \propto \kappa_i^{d(W/2-1)} \frac{\exp(c_0 d\kappa_i)}{I_{W/2-1}(\kappa_i)^d}$ over κ_i, for some hyper-parameters c_0 and d.

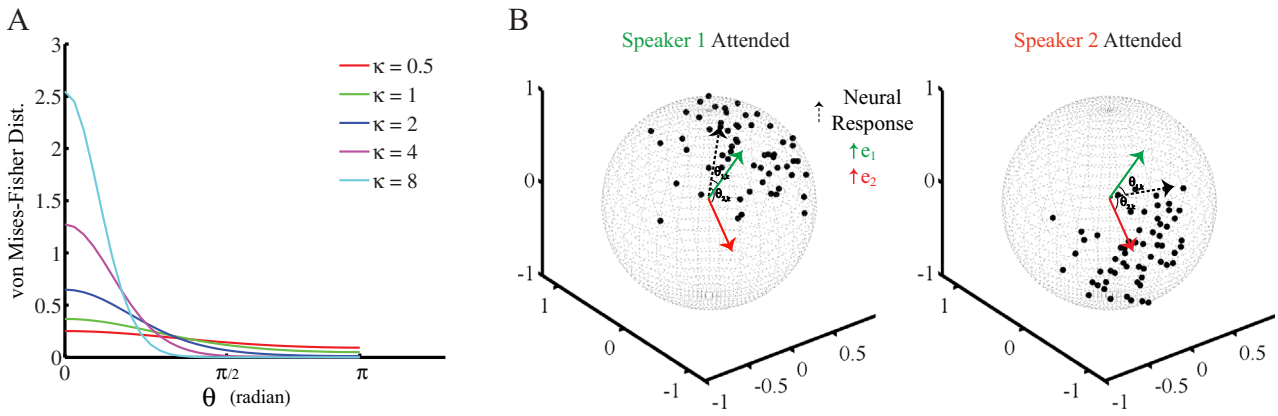

Fig. 2. A) von Mises–Fisher probability density for different κ parameters. B) Schematic view of von Mises–Fisher statistics on a three dimensional sphere: normalized neural response data points are shown by black dots on the unit sphere. Red and green arrows indicate the vectors of predicted neural response based on attending to speaker 1 or speaker 2, respectively. The angles between the neural response at window k and each of the predictors are shown as $\theta_{1,k}$ and $\theta_{2,k}$, for the case of attending to speaker 1 (left plot) and speaker 2 (right plot), respectively. The point cloud formed by the neural response is aligned with the direction of the predictor vector corresponding to the attention state.

Parameter estimation: A novel em-based decoder

Let

$$\Omega := \left\{\kappa_1, \kappa_2, \{z_k\}_{k=1}^{K}, \{\eta_k\}_{k=1}^{K}\right\} \tag{13}$$

be the set of state-space parameters. In principle, these parameters can be estimated through maximum *a posteriori* (MAP) estimation. However, due to the involved functional form of the log-likelihood and particularly temporal coupling of the state parameters, direct maximization of the log-posterior requires solving a high dimensional convex optimization problem. Instead, we use a novel form of the Expectation–Maximization (EM) algorithm to efficiently estimate the state parameters (Dempster et al., 1977). Taking $\{n_{k,r}\}_{k=1,r=1}^{K,R}$ as the unobserved data, the complete data log-posterior can lead to a feasible MAP estimate of the parameters, due to its tractable functional form for optimization purposes.

The overall estimation procedure consists of two nested EM algorithms and is outlined in Algorithm 1. At the ℓth iteration of the outer EM, the E-step involves computing $\mathbb{E}\{n_{k,r}^{(\ell+1)}|\Omega^{(\ell)}, \{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}\}$, using Bayes' rule, and the M-step updates $\kappa_1^{(\ell+1)}$, $\kappa_2^{(\ell+1)}$, $\{\eta_k^{(\ell+1)}\}_{k=1}^{K}$ and $\{z_k^{(\ell+1)}\}_{k=1}^{K}$. As for the last two sets of parameters, the maximization in the M-step itself is computed using the inner EM algorithm. In the inner EM algorithm, the E-step corresponds to a Bernoulli smoothing algorithm (Smith and Brown, 2003; Smith et al., 2004) and the M-step updates the state variance sequence (Shumway and Stoffer, 1982). The detailed derivations of the estimation procedure are provided in Appendices A and B. Confidence intervals for the estimated values of p_k can be obtained by mapping the confidence intervals of the posterior estimates of the Gaussian variables z_k via the inverse logit mapping (See the output of Algorithm 1). In summary, the decoder inputs the neural response observations and the envelopes of the two speech streams, and outputs the Bernoulli success probability sequence corresponding to attending to speaker 1. The choice of the hyperparameters will be discussed in Section Decoding auditory attention from MEG: a simulation study. We will refer to the estimator outlined in Algorithm 1 as the attention decoder in the remainder of the paper.

Subjects, stimuli, and procedures

Eleven normal-hearing, right-handed young adults (ages between 20 and 31) participated in this study, consisting of two experiments: constant-attention experiment (eight subjects, three female) and attention-switch (seven subjects, four female). Four subjects (three female) participated in both experiments. All subjects were compensated for their participation. The experimental procedures were approved by the University of Maryland Institutional Review Board. Written, informed consent was obtained from each subject before the experiment.

The stimuli consist of segments from the book *A Child's History of England* by Charles Dickens, narrated by two different readers (of opposite genders). Four speech segments (one target and one masker segment for each speaker) were used to generate three speech mixtures. Each speech mixture was constructed by mixing two speech segments digitally in a single channel with duration of 1 min, as described next. The first mixture was generated using the male target segment and the female masker segment, whereas the second mixture was generated using the female target segment and the male masker segment. The third mixture was generated using male and female target segments. Periods of silence longer than 300 ms were shortened to 300 ms to keep the speech streams flowing continuously. All stimuli were low-pass-filtered below 4 kHz and delivered diotically at both ears using tube phones plugged into the ear canals. In all trials, the stimuli were mixtures with equal root-mean-square values of sound amplitude, presented roughly at a 65 dB sound pressure level (SPL).

In the constant-attention experiment, subjects were asked to focus on one speaker (speaker 1, male; speaker 2, female) through the entire trial. In the attention-switch experiment, subjects were instructed to focus on one speaker in the first 28 s of the trial, switch their attention to the other speaker after hearing a 2 second pause (28th to 30th seconds), and maintain their focus on the latter speaker through the end of that trial. Consequently, there were four conditions: 1) attending to speaker 1 for the entire trial duration, 2) attending to speaker 2 for the entire trial duration, 3) attending to speaker 1 and switching to speaker 2 halfway through the trial, and 4) attending to speaker 2 and switching to speaker 1 halfway through the trial. The first mixture was used as the stimulus for condition 1, second mixture for condition 2 and third mixture for conditions 3 and 4. Each mixture was repeated three times during each experimental condition. The first second of each section was replaced by the clean recording from the target speaker to help the listener attend to the target speaker. After each condition was presented, subjects answered comprehensive questions related to the passage on which they focused, as a way to keep them motivated

in attending to the target speaker. Eighty percent of the questions were correctly answered on average. The order of presentation for the constant-attention experiment (conditions 1 and 2), and the attention switch (conditions 3 and 4) was counterbalanced across subjects participating in that experiment.

Algorithm 1. Estimation of the state-space parameters

input : Neural response $\{\mathbf{y}_{k,r}\}_{k,r=1}^{K,R}$, tolerance $\mathsf{tol} \in (0, 0.001)$, significance level α, and maximum number of iterations for outer and inner EM algorithms $L_{\max}$ and $M_{\max} \in N^+$, respectively.

Initialization: initial guess of state variables $z_k^{(0)}$ and state-noise variances $\eta_k^{(0)}$ for $k = 1, 2, \cdots, K$, initial conditions $z_{0|0}$ and $\sigma_{0|0}$, Initial values for von Mises-Fisher distribution parameters $\kappa_1^{(0)}$ and $\kappa_2^{(0)}$. Initialize iteration numbers to $l = 1$ *and* $m = 1$;

Outer EM iteration:

while $l \leq L_{\max}$ *or relative change in log-posterior* $\geq \mathsf{tol}$ **do**

E-step: Compute $\mathbb{E}^{(\ell)}\{n_{k,r}\} := \mathbb{E}\left\{n_{k,r} \middle| \{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}, \Omega^{(\ell)}\right\}$, *for all* $k = 1, 2, \cdots, K$ *and* $r = 1, 2, \cdots, R$. (A.2).

M-step: Update $\kappa_1^{(\ell+1)}$ *and* $\kappa_2^{(\ell+1)}$ (A.3).

Inner EM iteration:

while $m \leq M_{\max}$ *or relative change in log-posterior* $\geq \mathsf{tol}$ **do**

E-step: Compute $\bar{z}_{k|K}^{(\ell+1,m)} := \mathbb{E}\left\{z_k \middle| \{\mathbb{E}^{(\ell)}\{n_{k,r}\}\}_{k,r=1}^{K,R}\right\}$ *and* $\sigma_{k|K}^{(\ell+1,m)} := \mathsf{Var}\left\{z_k \middle| \{\mathbb{E}^{(\ell)}\{n_{k,r}\}\}_{k,r=1}^{K,R}\right\}$ *for all* $k = 1, 2, \cdots, K$ *using Bernoulli smoothing* (A.5, A.6).

M-step: Update $\eta_k^{(\ell+1,m)}$, *for all* $k = 1, 2, \cdots, K$. (A.7).

end

$z_k^{(\ell+1)} := \bar{z}_{k|K}^{(\ell+1,m)}$, $\sigma_k^{(\ell+1)} := \sigma_{k|K}^{(\ell+1,m)}$, *and* $\eta_k^{(\ell+1)} := \eta_k^{(\ell+1,m)}$, *for all* $k = 1, 2, \cdots, K$.

end

output: For $L \leq L_{\max}$ denoting the final counter value of the outer EM, output $\widehat{\kappa}_1 := \kappa_1^{(L+1)}$, $\widehat{\kappa}_2 := \kappa_2^{(L+1)}$, and for all $k = 1, 2, \cdots, K$ output $\widehat{\eta}_k := \eta_k^{(L+1)}$, $\widehat{p}_k := \mathsf{logit}^{-1}\left(z_k^{(L+1)}\right)$, and the confidence interval $\mathcal{CI}_k$ at a level of $1 - \alpha$ given by:

$$\mathcal{CI}_k = \left[\mathsf{logit}^{-1}\left(z_k^{(L+1)} - \Phi^{-1}(1-\tfrac{\alpha}{2})\sqrt{\sigma_k^{(L+1)}}\right), \quad \mathsf{logit}^{-1}\left(z_k^{(L+1)} + \Phi^{-1}(1-\tfrac{\alpha}{2})\sqrt{\sigma_k^{(L+1)}}\right)\right].$$

A pilot study from subjects listening to single speakers was performed prior to the main study. In this experiment, 6 trials (3 repetitions each of speaker 1 and speaker 2 target segments) were presented to the subjects and recordings were used for estimating the Temporal Response Functions (TRFs) in the forward model.

Data recording

MEG signals were recorded in a dimly lit magnetically shielded room (Yokogawa Electric Corporation) using a 160-channel whole-head system (Kanazawa Institute of Technology, Kanazawa, Japan), and with a sampling rate of 1 kHz. Detection coils were arranged in a uniform array on a helmet-shaped surface on the bottom of the dewar, with 25 mm between the centers of two adjacent 15.5-mm-diameter coils. Sensors are configured as first-order axial gradiometers with a baseline of 50 mm; their field sensitivities are 5 fT/$\sqrt{\text{Hz}}$ or better in the white noise region.

Stimuli were presented using the software package Presentation (Neurobehavioral Systems, Inc., Berkeley, CA, USA). The sounds (approximately 65 dB SPL) were delivered to the participants Õ ears with 50 Ω sound tubing (E-A-RTONE 3A; Etymotic Research), attached to E-A-RLINK foam plugs inserted into the ear canal. The entire acoustic delivery system was equalized to give an approximately flat transfer function from 40 to 3000 Hz, thereby encompassing the range of the presently delivered stimuli.

A 200 Hz low-pass filter and a notch filter at 60 Hz were applied to the magnetic signal online. Three of the 160 channels were magnetometers separated from the others and used as reference channels in measuring and canceling environmental noise (de Cheveigné and Simon, 2007). Five electromagnetic coils were used to measure each subject's head position inside the MEG machine. The head position was measured twice during the experiment, once before and once after to quantify the head movement.

MEG processing and neural source localization

Recorded MEG signals contained both stimulus-driven responses and stimulus-irrelevant background neural activity. In order to extract components that were phase-locked to the stimulus and consistent over trials, as opposed to the random irrelevant activities, we employed the Denoising Source Separation (DSS) algorithm (Särelä and Valpola, 2005; de Cheveigné and Simon, 2008). This algorithm is a blind source separation method that decomposes the data into temporally uncorrelated components by removing inconsistent temporal components not phased-locked to the stimulus. In other words, DSS suppresses the components of the data that are noise-like and enhances those that are consistent across trials, with no knowledge of the stimulus or the timing of the task. The recorded neural response during each 60 s was band-pass filtered between 1–8 Hz and down sampled to 200 Hz before submission to the DSS analysis. We found that the first DSS component alone was sufficient, so analysis was restricted to this component, which we denote by the auditory neural response throughout this paper. The spatial magnetic field distribution pattern of the auditory neural response was used for neural source localization. In all subjects, the magnetic field corresponding to the auditory neural response showed a stereotypical bilateral dipolar pattern (See Fig. 3-A).

Results

In order to evaluate the performance of the state-space model in decoding the attentional state of listeners and to illustrate the effectiveness of this model in various stimulus conditions, a number of realistic simulations and experimental data sets were employed. We first present our results on the robust estimation of the TRF, which forms the basis of the forward models used in both simulations and experimental data analysis. We will then present simulation results which highlight the capability of our proposed estimation framework in tracking the attentional state under a wide range of SNR values as well as dynamics. Finally, we will apply the proposed attentional decoding framework to experimental MEG data from several subjects which chimes in accordance to our simulation studies.

TRF estimation

TRFs corresponding to the attended speaker were estimated from the pilot conditions, where only single speech streams were presented to the subjects. Separate TRFs were obtained for speakers one and two, using 3 repeated trials for each and the TRF with smaller normalized least square error was chosen and used throughout the rest of our analysis. The TRF corresponding to the unattended speaker was approximated by truncating the attended TRF beyond a lag of 90 ms, on the grounds of the recent findings of Ding and Simon (2012a), which show that the components of the unattended TRF are significantly suppressed beyond the M50 evoked field. An example of an estimated TRF using the auditory neural response for a sample subject is shown in Fig. 3-B. The spatial magnetic field distribution pattern of the auditory neural response (Fig. 3-A) demonstrates a stereotypical bilateral dipolar pattern, as expected for auditory evoked field.

Decoding auditory attention from MEG: A simulation Study

In order to simulate neural response modulated by attention, first a binary sequence $\{n_{k,r}\}_{k=1,r=1}^{240,3}$ was generated as realizations of a Bernoulli process with success probability $p_k = 0.95$ or 0.05, corresponding to attention to the first or second speakers, respectively. The total observation time was 60 s with a sampling rate of $F_s = 200$ Hz ($T = 12{,}000$ samples) and the processing window length was chosen to be 250 ms ($W = 50$ samples). Using a TRF template of length 0.5 s

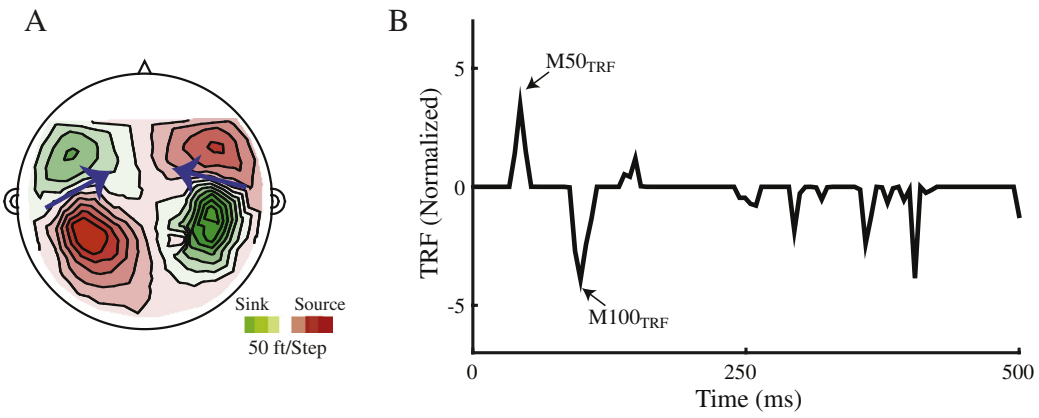

Fig. 3. A) MEG magnetic field distribution for the first DSS component of a sample subject shows a stereotypical pattern of neural activity, originating separately in the left and right auditory cortices. Red and green contours represent the magnetic field strength. Blue arrows schematically represent the locations of the dipole currents, generating the measured magnetic field. B) Estimated TRF for the sample subject has significant components analogous to the well-known M50 and M100 auditory responses, as well as later responses, as demonstrated in the figure.

estimated from experimental data (See Section TRF estimation), we generated 3 trials for various SNR values and with multiple attention switches throughout each trial.

Figs. 4-A and -B show the simulated neural signal (black traces) and predictors of attending to speaker one and two (red traces) at an SNR of 10 dB. Regions indicated by arrows in panels A and B demonstrate the time intervals, in which listeners are supposed to attend to either of the two speakers.

The hyper-parameters for the von Mises–Fisher distribution were chosen as $d = 100KR/2$ and $c_0 = 0.01$, consistent with the observed correlation values between the simulated neural response and the model prediction. The choice of $d = 100KR/2$ gives more weight to the prior than the empirical estimate of κ_i. The hyper-parameters α and β for the inverse-Gamma prior on the state variance were chosen as $\alpha = 2.01$ and $\beta = 0.5$. This choice of α close to 2 results in a non-informative prior, as the variance of the prior is given by $\beta^2/[(\alpha - 1)^2(\alpha - 2)] \approx 245$, while the mean is given by $\beta/(\alpha - 1) \approx 0.5$.

Estimated values of $\{p_k\}_{k=1}^{240}$ (green trace) and the corresponding confidence intervals (green hull) are shown in Fig. 4-C. The estimated p_k values reliably track the attentional state, and the transitions are captured with high accuracy. MEG data recorded from the brain is usually contaminated with environmental noise as well as nuisance sources of neural activity, which can considerably decrease the SNR of the measured signal. In order to test the robustness of the decoder with respect to observation noise, we repeated the above simulation with SNR values ranging from −20 to 10 dB. As demonstrated in Fig. 4-D, the confidence intervals and the estimated transition width widen gracefully as the SNR decreases. The dynamic denoising feature of the proposed state-space model results in a desirable decoding performance for SNR values above −15 dB (Fig. 4-E).

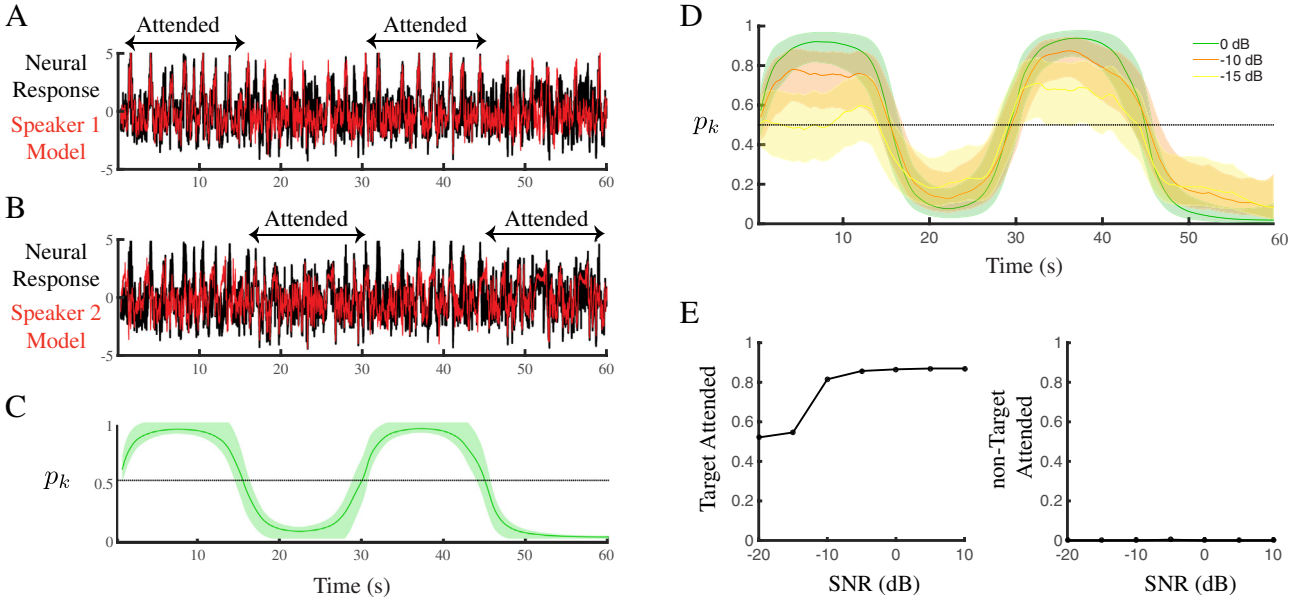

Fig. 4. Simulated neural response (black traces) and model prediction (red traces) of A) speaker one and B) speaker two at SNR = 10 dB. Black arrows indicate the instructed attentional state of the subjects. The MEG units are in *pT/m*. C) Estimated values of $\{p_k\}$ with 95% confidence intervals. D) Estimated values of $\{p_k\}$ from simulated neural response vs. SNR = 0, −10 and −15 dB. Error hulls indicate 95% confidence intervals. E) Behavioral results of the simulated neural response vs. SNR values ranging from −20 to 10 dB. The time fraction for which the estimated attentional state follows the target speaker (the opposite speaker) as a function of different SNRs is shown in the left panel (right panel).

Decoding auditory attention from MEG: Application to experimental MEG data

We assessed our proposed state-space model and decoder on experimental MEG data recorded from 11 human subjects who listened to one of the two competing speakers in constant-attention and attention-switch experiments (see Methods). All hyper-parameters in the model were chosen similar to those of the simulation studies in the previous section, except for the prior parameter c_0 for the von Mises–Fisher distribution which was conservatively chosen as $c_0 = 0.01$, consistent with the observed correlation values between the simulated neural response and the model prediction.

The predicted p_k values resulted from single and multi-trial analysis are shown in Fig. 5 for three sample subjects. For multi-trial analysis (3rd panel of each plot) 90% confidence intervals are shown by the shaded hulls around the estimated values. In the first and second conditions subjects were instructed to maintain their attention through the entire experiment to the speaker one and speaker two, respectively (Figs. 5-A and -B). The decoding results demonstrate the decoder's reliable recovery of the attention modulation by estimating $\{p_k\}$ close to 1 for the first condition and values close to 0 for the second condition. For the third and fourth conditions, subjects were instructed to switch their attention after hearing a 2 s pause, in the middle of each trial, from the speaker one to the speaker two (Fig. 5–C) and from the speaker two to the speaker one (Fig. 5-D). Using multiple-trial analysis, the decoder was able to capture the attentional switch occurring roughly halfway through the trial. The decoding of individual trials in the fourth panel of Fig. 5-C & -D suggest that the exact switching times were not consistent across different trials, as the attentional switch might have occurred slightly earlier or later than the presented cue.

The performance of individual subjects were evaluated by computing time fractions in which the target speaker or the alternative speaker were followed according to the estimated results from the state-space decoder. All computations were done within the confidence interval of

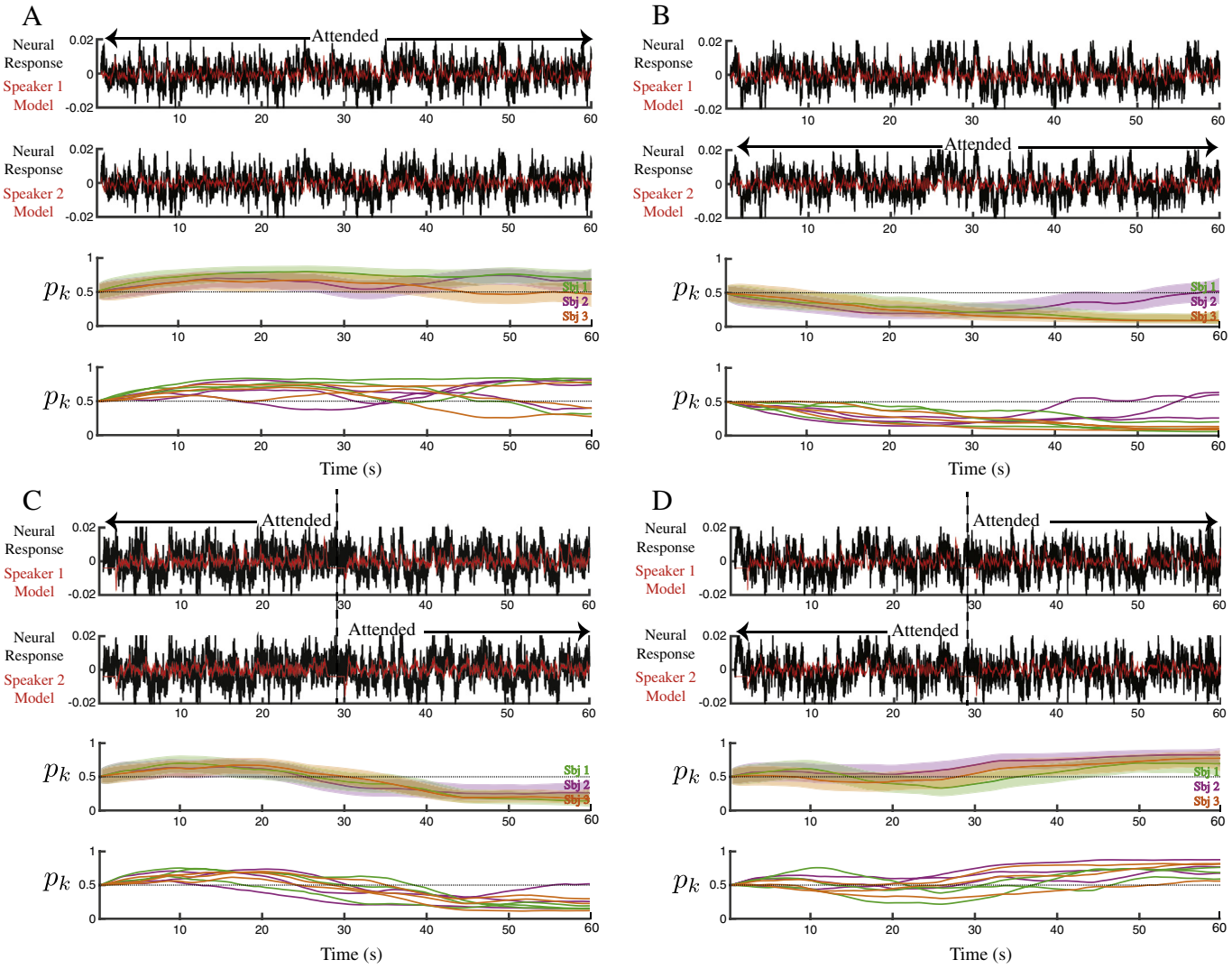

Fig. 5. Decoding auditory attentional modulation in experimental MEG data. In each subplot, the neural response (black traces) and the model prediction (red traces) for attending to speaker one and speaker two are shown in the first and second panels, respectively, for one sample subject. The third panel shows the estimated values of $\{p_k\}$ and the corresponding confidence intervals using multi-trial analysis for three sample subjects. The fourth panel shows the estimated $\{p_k\}$ values for single trials. A) Condition 1: attending to the speaker one through the entire trial. B) Condition 2: attending to the speaker two through the entire trial. C) Condition 3: attending to the speaker one until $t = 28$ *s* and switching attention to the speaker two after the 2 s pause. D) Condition 4: attending to the speaker two until $t = 28$ *s* and switching attention to the speaker one after the 2 s pause. Dashed lines in subplots C and D indicate the start of the 2 s silence cue for attentional switch. Error hulls indicate 90% confidence intervals. The MEG units are in *pT/m*.

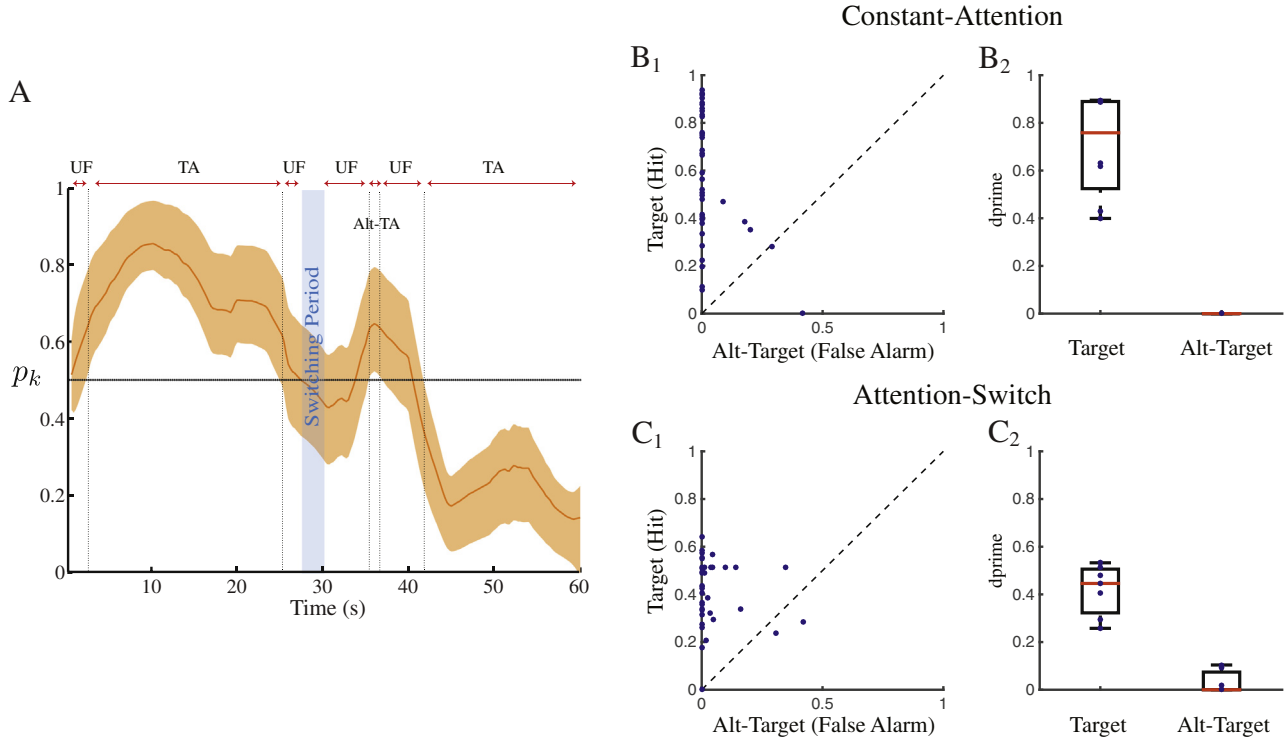

Fig. 6. Schematic illustration of attentional states and behavioral analysis. A) The estimated attentional condition at each time point can take one of the followings states: Target Attended (TA), Alternative Target Attended (Alt-TA), and the Unfollowed state (UF). Examples of the attentional states for a sample subject are depicted in panel A, for a sample trial from condition 3. B1, C1) Target speaker attended time fractions are plotted with respect to the Alt-target attended time fractions for individual subjects in constant-attention and attention-switch experiments, respectively. B2, C2) Target and Alt-target attended time fractions are computed via multi-trial analysis. Box plots indicate the median and quartile percentages of subjects' behavioral performances in attending to the target and non-target speakers (first and second boxes in each plot, respectively). Individual subject performances, shown in blue markers, are plotted on top of each box plot.

90% for multi-trial and 70% for single-trial analysis. An illustrative example of the time intervals in which a sample subject is in target, alternative target (Alt-target) or unfollowed attentional sate is shown in Fig. 6-A, for a sample trial in speaker one-speaker two attention-switch condition (condition 3). The evaluated target and Alt-target attentional time fractions for single trials are plotted in Figs. 6-B1 and -B2, for the constant-attention and the attention-switch experiments, respectively. As shown in these figures, most of the data points fall above the identity line, indicating larger time fractions in which the target speakers were attended vs. the alternative targets. The behavioral results from multi-trial analysis were significantly improved compared to the single-trial estimations (one way ANOVA, $P < 0.01$). This is indeed expected from the state-space formulation, as the variance of the state variable z_k is inversely proportional to the number of trials R (See Eq. (A.5)). The results of multi-trial estimations are shown in Figs. 6-C1 & -C2 for each individual subject and two experimental conditions. The median, 25% and 75% quartile values are shown in separate box plots for target and Alt-target attended time fractions and for each individual experiment. In addition, individual subject performances averaged over condition pairs within constant-attention experiment (conditions one & two) and attention-switch experiment (conditions three & four) are plotted in blue on top of the corresponding box plots. Evaluated performances for the decoded attentional states show that time fractions in which the target speakers were attended to, were significantly larger than the Alt-target attended time fractions (one way ANOVA, $P < 0.001$), highlighting the successful decoding of the attentional states via the state-space model.

Discussion

In this study, we developed a biophysically-inspired state-space model that provides an estimation framework for decoding the attentional state of a listener in a competing-speaker environment. The proposed algorithm takes advantage of the temporal continuity in the attentional state, resulting in a decoding performance, which is highly accurate and resolved in time. Parameter estimation of this model is carried out using the EM algorithm, which is tied to the efficient computation of the Bernoulli process smoothing, resulting in a very low overall computational complexity. The output of the state-space model at each EM iteration is plotted in Fig. 7 for a sample subject and all four experimental conditions. These plots illustrate the convergence path of the EM iterations in estimating the attention probability values p_k, starting from values at chance level (0.5) and converging to values near 0 or 1 depending on the targeted speaker.

The novel state space model proposed in this study is supported by performance evaluation of the model on realistic simulated data, as well as evoked neural activity from the auditory cortex of humans, recorded via MEG. These studies divulge two main advantages in the current model over the state of the art methods such as the reverse correlation technique (Bialek et al., 1991; Gielen et al., 1988; Hesselmans and Johannesma, 1989).

First, in this proposed model, temporal resolution of the estimated state of attention is in the order of a few seconds rather than a minute. This resolution is comparable to empirically estimated speed of attention switching in humans; therefore the proposed model provides a dynamic framework for tracking the attentional state of a listener in real world scenarios. This is a considerable improvement over the commonly used methods based on reverse correlation, in which the recovery of the stimulus paradigm from the corresponding neural response results is a poor reconstruction of the stimulus using short processing time windows, and therefore fails in tracking the attentional state in a precise fashion (Ding and Simon, 2012a; Mesgarani and Chang, 2012).

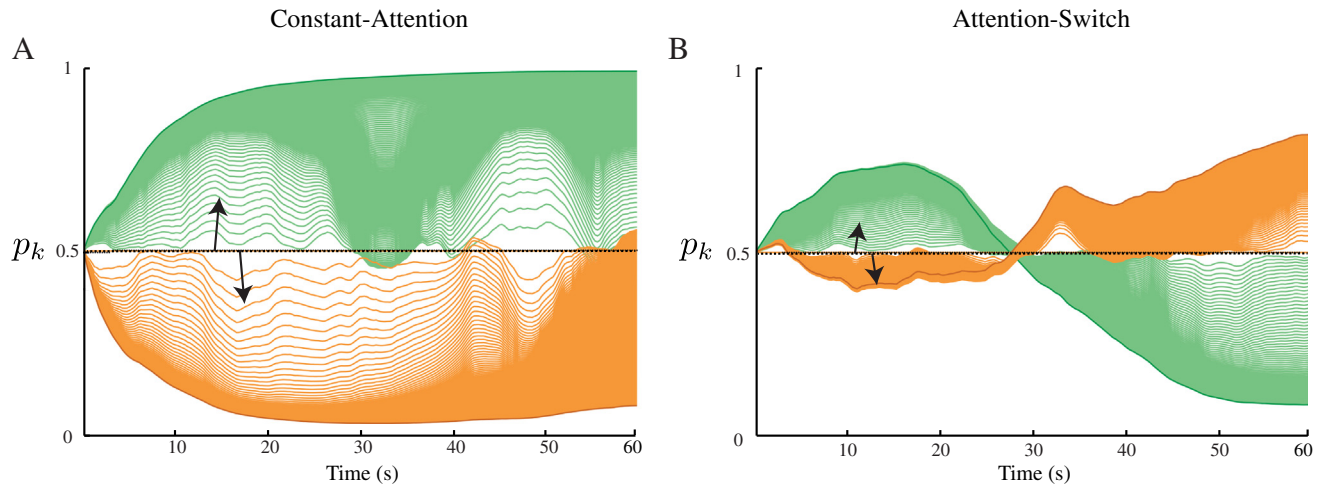

Fig. 7. A step-wise illustration of the EM convergence. A) The output of the state-space decoder is plotted after each EM iteration for sample trials of attending to speaker 1 (green curves), and attending to speaker 2 (orange curves), in the Constant-Attention experiment. B) EM iterations are plotted for sample trials of the Attention-Switch experiment and for attention switches from speaker 1 to speaker 2 (green curves), and from speaker 2 to speaker 1 (orange curves).

Second, the principled statistical framework used in constructing the decoder allows us to obtain confidence bounds on the estimated attentional state. This feature is crucial to obtaining a statistically principled framework for assessing the validity of the algorithm output. Moreover, the proposed approach benefits from the inherent model-based dynamic denoising of the underlying state-space model, and is able to reliably decode the attentional state under very low SNR conditions. A comparison of our method with a correlation-based classifier (without the state-space mechanism) was presented in our earlier work (Akram et al., 2014), which confirmed the latter observation and revealed a significant performance gap.

A potential application of this analysis framework is to be used as a real-time cocktail party analyzer, tracking the attentional state of a listener in a complex auditory environment. The state-space model provides estimation of the probability of attending to either one of the speakers at each time point t based on the recorded neural data at all other time points before (via non-linear filtering) and after (via backward smoothing) t. Assuming that the cognitive state of attention is a continuous process in time, this continuity is appropriately accounted for in the proposed model; however, for real-time Brain–Computer Interface (BCI) applications, the smoothing step can be omitted and estimation of the attentional state can be causally carried out via the proposed non-linear filter.

Future work includes generalization of the proposed model to more realistic and complex auditory environments with more diverse sources such as mixtures of speech, music and structured background noise. Nevertheless, the promising performance of the proposed algorithm for MEG recordings makes it an appealing candidate for EEG applications.

Acknowledgment

This work was supported by the National Institutes of Health (NIH), 1R01AG036424.

Appendix A. Parameter estimation of the inverse problem

Let

$$\Omega := \left\{\kappa_1, \kappa_2, \{z_k\}_{k=1}^{K}, \{\eta_k\}_{k=1}^{K}\right\} \tag{A.1}$$

be the set of parameters.

The log-posterior of the parameter set Ω given the observed data $\{\theta_{i,k,r}\}_{i,k,r=1}^{2,T,R}$ is given by:

$$\begin{aligned}\log p\left(\Omega\middle|\{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}\right) = \sum_{r,k=1}^{R,K} \log\left[\frac{2\kappa_1^{W/2-1}p_k}{(2\pi)^{W/2}I_{W/2-1}(\kappa_1)}\exp\left(\kappa_1\cos\left(\theta_{1,k,r}\right)\right)\right.\\ \left.+\frac{2\kappa_2^{W/2-1}(1-p_k)}{(2\pi)^{W/2}I_{W/2-1}(\kappa_2)}\exp\left(\kappa_2\cos\left(\theta_{2,k,r}\right)\right)\right] + \left[(\kappa_1+\kappa_2)c_0 d + d(W/2-1)\left(\log\kappa_1+\log\kappa_2\right) - d\left(\log I_{W/2}(\kappa_1)+\log I_{W/2}(\kappa_2)\right)\right]\\ -\sum_{r,k=1}^{R,K}\left\{\frac{1}{2\eta_k}(z_k-z_{k-1})^2+\frac{1}{2}\log\eta_k+(\alpha+1)\log\eta_k+\frac{\beta}{\eta_k}\right\}+\text{cst.}\end{aligned}$$

where cst. denotes terms that are not functions of Ω. The MAP estimate of the parameters is difficult to obtain given the involved functional form of the log-posterior. However, the complete data log-posterior, where the unobservable sequence $\{n_{k,r}\}_{k=1,r=1}^{K,R}$ is given, takes the form:

$$\begin{aligned}\log p\left(\Omega\middle|\{\theta_{i,k,r},n_{k,r}\}_{i,k,r=1}^{2,K,R}\right) = &\sum_{r,k=1}^{R,K} n_{k,r}[(W/2-1)\log(\kappa_1)+\kappa_1\cos(\theta_{1,k,r})-\log I_{W/2-1}(\kappa_1)]\\ &+\sum_{r,k=1}^{R,K}(1-n_{k,r})[(W/2-1)\log(\kappa_2)+\kappa_2\cos(\theta_{2,k,r})\log I_{W/2-1}(\kappa_2)]\\ &+[(\kappa_1+\kappa_2)c_0 d+d(W/2-1)(\log\kappa_1+\log\kappa_2)-d(\log I_{W/2}(\kappa_1)+\log I_{W/2}(\kappa_2))]\\ &+\sum_{r,k=1}^{R,K}\left[n_{k,r}\log p_k+(1-n_{k,r})\log(1-p_k)\right]\\ &-\sum_{r,k=1}^{R,K}\left\{\frac{1}{2\eta_k}(z_k-z_{k-1})^2+\frac{1}{2}\log\eta_k+(\alpha+1)\log\eta_k+\frac{\beta}{\eta_k}\right\}+\text{cst.}\end{aligned}$$

The log-posterior of the parameters given the complete data has a tractable functional form for optimization purposes. Therefore, by taking $\{n_{k,r}\}_{k=1,r=1}^{K,R}$ as the unobserved data, we can estimate Ω via the EM algorithm (Dempster et al., 1977). Using Bayes' rule, the expectation of $n_{k,r}$, given $\{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}$ and current estimates of the parameters $\Omega^{(\ell)} := \{\kappa_1^{(\ell)}, \kappa_2^{(\ell)}, \{z_k^{(\ell)}\}_{k=1}^K, \{\eta_k^{(\ell)}\}_{k=1}^K\}$ is given by:

$$\mathbb{E}\left\{n_{k,r}\middle|\{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R},\Omega^{(\ell)}\right\} = \frac{\dfrac{2\kappa_1^{(\ell)W/2-1}p_k^{(\ell)}}{(2\pi)^{W/2}I_{W/2-1}\left(\kappa_1^{(\ell)}\right)}\exp\left(\kappa_1^{(\ell)}\cos(\theta_{1,k,r})\right)}{\dfrac{2\kappa_1^{(\ell)W/2-1}p_k^{(\ell)}}{(2\pi)^{W/2}I_{W/2-1}\left(\kappa_1^{(\ell)}\right)}\exp\left(\kappa_1^{(\ell)}\cos(\theta_{1,k,r})\right)+\dfrac{2\kappa_2^{(\ell)W/2-1}\left(1-p_k^{(\ell)}\right)}{(2\pi)^{W/2}I_{W/2-1}\left(\kappa_2^{(\ell)}\right)}\exp\left(\kappa_2^{(\ell)}\cos(\theta_{2,k,r})\right)}. \tag{A.2}$$

Denoting the above expectation by the shorthand $\mathbb{E}^{(\ell)}\{n_{k,r}\}$, the M-step of the EM algorithm for $\kappa_1^{(\ell+1)}$ and $\kappa_2^{(\ell+1)}$ gives:

$$\kappa_i^{(\ell+1)} = A^{-1}\left(\frac{\sum_{r,k=1}^{R,K}\varepsilon_{i,k,r}^{(\ell)}\cos(\theta_{i,k,r})+c_0 d}{d+\sum_{r,k=1}^{R,K}\varepsilon_{i,k,r}^{(\ell)}}\right),\quad \varepsilon_{i,k,r}^{(\ell)} = \begin{cases}\mathbb{E}^{(\ell)}\{n_{k,r}\} & i=1\\ 1-\mathbb{E}^{(\ell)}\{n_{k,r}\} & i=2\end{cases} \tag{A.3}$$

where $A(x) := -\frac{W/2-1}{x}+\frac{0.5(I_{W/2-2}(x)+I_{W/2}(x))}{I_{W/2-1}(x)}$, with $I_W(\cdot)$ denoting the W^{th} order modified Bessel function of the first kind. Inversion of $A(\cdot)$ can be carried out numerically in order to find $\kappa_1^{(\ell+1)}$ and $\kappa_2^{(\ell+1)}$. The M-step for $\{\eta_k\}_{k=1}^K$ and $\{z_k\}_{k=1}^K$ corresponds to the following maximization problem:

$$\underset{\{z_k,\eta_k\}_{k=1}^K}{\operatorname{argmax}}\sum_{r,k=1}^{R,K}\left[\mathbb{E}^{(\ell)}\{n_{k,r}\}z_k-\log(1+\exp(z_k))-\frac{1}{2\eta_k}\left[(z_k-z_{k-1})^2+2\beta\right]-\frac{1+2(\alpha+1)}{2}\log\eta_k\right]. \tag{A.4}$$

An efficient approximate solution to this maximization problem is given by another EM algorithm, where the E-step is the point process smoothing algorithm (Smith and Brown, 2003; Smith et al., 2004) and the M-step updates the state variance sequence (Shumway and Stoffer, 1982). At iteration m, given an estimate of $\eta_k^{(\ell+1)}$, denoted by $\eta_k^{(\ell+1,m)}$, the forward pass of the E-step for $k = 1, 2, ..., K$ is given by:

$$\begin{cases}\bar{z}_{k|k-1}^{(\ell+1,m)} = \bar{z}_{k-1|k-1}^{(\ell+1,m)}\\ \sigma_{k|k-1}^{(\ell+1,m)} = \sigma_{k-1|k-1}^{(\ell+1,m)}+\dfrac{\eta_k^{(\ell+1,m)}}{R}\\ \bar{z}_{k|k}^{(\ell+1,m)} = \bar{z}_{k|k-1}^{(\ell+1,m)}+\sigma_{k|k-1}^{(\ell+1,m)}\left[\sum_{r=1}^{R}\mathbb{E}^{(\ell)}\{n_{k,r}\}-R\dfrac{\exp\left(\bar{z}_{k|k}^{(\ell+1,m)}\right)}{1+\exp\left(\bar{z}_{k|k}^{(\ell+1,m)}\right)}\right]\\ \sigma_{k|k}^{(\ell+1,m)} = \left[\dfrac{1}{\sigma_{k|k-1}^{(\ell+1,m)}}+R\dfrac{\exp\left(\bar{z}_{k|k}^{(\ell+1,m)}\right)}{\left(1+\exp\left(\bar{z}_{k|k}^{(\ell+1,m)}\right)\right)^2}\right]^{-1}\end{cases}. \tag{A.5}$$

Note that the third equation in the forward filter is non-linear in $\bar{z}_{k|k}^{(\ell+1,m)}$, and can be solved using standard techniques (e.g., Newton's method). More details on derivation of the non-linear forward filter can be found in Appendix B. For $k = K-1, K-2, ..., 1$, the backward pass of the E-step is given by:

$$\begin{cases}s_k^{(\ell+1,m)} = \sigma_{k|k}^{(\ell+1,m)}/\sigma_{k+1|k}^{(\ell+1,m)}\\ \bar{z}_{k|K}^{(\ell+1,m)} = \bar{z}_{k|k}^{(\ell+1,m)}+s_k^{(\ell+1,m)}\left(\bar{z}_{k+1|K}^{(\ell+1,m)}-\bar{z}_{k+1|k}^{(\ell+1,m)}\right)\\ \sigma_{k|K}^{(\ell+1,m)} = \sigma_{k|k}^{(\ell+1,m)}+s_k^{(\ell+1,m)}\left(\sigma_{k+1|K}^{(\ell+1,m)}-\sigma_{k+1|k}^{(\ell+1,m)}\right)s_k^{(\ell+1,m)}\end{cases}. \tag{A.6}$$

The M-step gives the updated value of $\eta_k^{(\ell+1,m+1)}$ as:

$$\begin{aligned}\eta_k^{(\ell+1,m+1)} &= \frac{\mathbb{E}\left(z_k^2\middle|\Omega^{(\ell)},\{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}\right)+\mathbb{E}\left(z_{k-1}^2\middle|\Omega^{(\ell)},\{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}\right)-2\mathbb{E}\left(z_k,z_{k-1}\middle|\Omega^{(\ell)},\{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}\right)+2\beta}{1+2(\alpha+1)}\\ &= \frac{\left(\bar{z}_{k|K}^{(\ell+1,m)}-\bar{z}_{k-1|K}^{(\ell+1,m)}\right)^2+\sigma_{k|K}^{(\ell+1,m)}+\sigma_{k-1|K}^{(\ell+1,m)}-2\sigma_{k|K}^{(\ell+1,m)}s_{k-1}^{(\ell+1,m)}+2\beta}{1+2(\alpha+1)}.\end{aligned} \tag{A.7}$$

Appendix B. Derivation of the recursive nonlinear filtering algorithm

Assume that at time $(k-1)$, $z_{k-1|k-1}$ and $\sigma^2_{k|k-1}$ are given. The distribution of z_k given all the data up to time k is $\mathcal{N}(z_{k-1|k-1},\sigma^2_{k|k-1})$, where $\sigma^2_{k|k-1}=\eta_k+\sigma^2_{k-1|k-1}$. To derive the non-linear recursive filter, we keep track of the parameters of the posterior distribution $p(z_k|\Omega)$:

$$\log(p(z_k|\Omega))=\left\{-\frac{(z_k-z_{k-1|k-1})^2}{\sigma^2_{k|k-1}}\right\}\left\{\mathbb{E}^{(\ell)}\{n_k\}z_k-\log(1+\exp(z_k))-\frac{\beta}{\eta_k}+\frac{1+2(\alpha+1)}{2}\log\eta_k\right\}. \tag{B.1}$$

To find the mode of $p(z_k|\Omega)$, we apply Gaussian approximation to the posterior density. The approximation is based on recursively computing the posterior mode $z_{k|k}$ and computing its variance $\sigma^2_{k|k}$ as the negative inverse Hessian of the log-posterior probability density (Tanner, 1993). Differentiating Eq. (A.1) w.r.t. z_k gives

$$-\frac{z_k-z_{k-1|k-1}}{\sigma^2_{k|k-1}}+\mathbb{E}^{(\ell)}\{n_k\}-\frac{\exp(z_k)}{1+\exp(z_k)}=0 \tag{B.2}$$

and solving for z yields

$$z_k=z_{k-1|k-1}+\sigma^2_{k|k-1}\left\{\mathbb{E}^{(\ell)}\{n_k\}-\frac{\exp(z_k)}{1+\exp(z_k)}\right\}. \tag{B.3}$$

This equation is non-linear w.r.t. z_k and can be solved using the Newton's method. The Hessian of Eq. (B.1) is given by

$$\frac{-1}{\sigma^2_{k|k-1}}-\frac{\exp(z_k)(1+\exp(z_k))-\exp^2(z_k)}{(1+\exp(z_k))^2} \tag{B.4}$$

and hence the variance of z_k, under the Gaussian approximation is given by:

$$\sigma^2_{k|k}=\left(\frac{1}{\sigma^2_{k|k-1}}+\frac{\exp(z_{k|k})}{(1+\exp(z_{k|k}))^2}\right)^{-1}. \tag{B.5}$$

Appendix C. Covariance smoothing

The lagged covariance $\sigma^2_{k,l|K}$ can be computed from the state-space covariance smoothing algorithm (De Jong and Mackinnon, 1988) given by the following equation:

$$\sigma^2_{k,l|K}=\sigma^2_{k|k}\left(\sigma^2_{k+1|k}\right)^{-1}\sigma^2_{k+1,l|K} \tag{C.1}$$

for $1\le k\le l\le K$. Hence, the lagged covariance term appearing in our E-step is given by:

$$\mathrm{Cov}\left\{z_{k+1},z_k\middle|\Omega^{(\ell)},\{\theta_{i,k,r}\}_{i,k,r=1}^{2,K,R}\right\}=\sigma^2_{k,k+1|K}=\sigma^2_{k|k}\left(\sigma^2_{k+1|k}\right)^{-1}\sigma^2_{k+1|K} \tag{C.2}$$

which is easily computable using the smoothed state variances.

References

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

Ba, D., Babadi, B., Purdon, P.L., Brown, E.N., 2014. Convergence and stability of iteratively re-weighted least squares algorithms. IEEE Trans. Signal Process. 62, 183–195.

Bergman, A.S., 1994. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, Cambridge, MA.

Bialek, W., Rieke, F., Van Steveninck, R.d.R., Warland, D., 1991. Reading a neural code. Science 252, 1854–1857.

Brungart, D.S., 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Am. 109, 1101–1109.

Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 25, 975–979.

David, S.V., Mesgarani, N., Shamma, S.A., 2007. Selective cortical representation of attended speaker in multi-talker speech perception. Netw. Comput. Neural Syst. 18, 191–221.

de Cheveigné, A., Simon, J.Z., 2007. Denoising based on time-shift PCA. J. Neurosci. Methods 165, 297–305.

de Cheveigné, A., Simon, J.Z., 2008. Denoising based on spatial filtering. J. Neurosci. Methods 171, 331–339.

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

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. 39, 1–38.

Ding, N., Simon, J.Z., 2012a. Emergence of neural encoding of auditory objects while listening to competing speakers. PNAS 109, 11854–11859.
Ding, N., Simon, J.Z., 2012b. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. 107, 78–89.
Fisher, N.I., 1993. Statistical Analysis of Spherical Data. Cambridge University Press.
Fishman, Y.I., Steinschneider, M., 2010. Neural correlates of auditory scene analysis based on inharmonicity in monkey primary auditory cortex. J. Neurosci. 30, 12480–12494.
Gielen, C., Hesselmans, G., Johannesma, P., 1988. Sensory interpretation of neural activity patterns. Math. Biosci. 88, 15–35.
Griffiths, T.D., Warren, J.D., 2004. What is an auditory object? Nat. Rev. Neurosci. 5, 887–892.
Hesselmans, G.H., Johannesma, P.I., 1989. Spectro-temporal interpretation of activity patterns of auditory neurons. Math. Biosci. 93, 31–51.
Hinrichsen, D., Pritchard, A.J., 2005. Mathematical Systems Theory I: Modelling, State Space Analysis, Stability and Robustness vol. 1. Springer Science & Business Media.
McDermott, J.H., 2009. The cocktail party problem. Curr. Biol. 19, R1024–R1027.
Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236.
Mesgarani, N., David, S.V., Fritz, J.B., Shamma, S.A., 2009. Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J. Neurophysiol. 102, 3329.
Mirkovic, B., Debener, S., Jaeger, M., De Vos, M., 2015. Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. J. Neural Eng. 12, 046007.
O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., Lalor, E.C., 2014. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. Cereb. Cortex 25 (7), 1697–1706.
Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., Chang, E.F., 2012. Reconstructing speech from human auditory cortex. PLoS Biol. 10, 175.
Särelä, J., Valpola, H., 2005. Denoising source separation. J. Mach. Learn. Res. 6, 233–272.
Shamma, S.A., Elhilali, M., Micheyl, C., 2011. Temporal coherence and attention in auditory scene analysis. Trends Neurosci. 34, 114–123.
Shumway, R.H., Stoffer, D.S., 1982. An approach to time series smoothing and forecasting using the EM algorithm. J. Time Ser. Anal. 3, 253–264.
Smith, A.C., Brown, E., 2003. Estimating a state-space model from point process observations. Neural Comput. 15, 965–991.
Smith, A.C., Frank, L.M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., Graybiel, A.M., Suzuki, W.A., Brown, E.N., 2004. Dynamic analysis of learning in behavioral experiments. J. Neurosci. 24, 447–461.
Tanner, M.A., 1993. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. Springer-Verlag.