Dynamic Estimation of the Auditory Temporal Response Function From MEG in Competing-Speaker Environments

Sahar Akram, Jonathan Z. Simon, and Behtash Babadi*, *Member, IEEE*

***Abstract*—*Objective*: A central problem in computational neuroscience is to characterize brain function using neural activity recorded from the brain in response to sensory inputs with statistical confidence. Most of existing estimation techniques, such as those based on reverse correlation, exhibit two main limitations: first, they are unable to produce dynamic estimates of the neural activity at a resolution comparable with that of the recorded data, and second, they often require heavy averaging across time as well as multiple trials in order to construct statistical confidence intervals for a precise interpretation of data. In this paper, we address the above-mentioned issues for estimating auditory temporal response function (TRF) as a parametric computational model for selective auditory attention in competing-speaker environments. *Methods:* The TRF is a sparse kernel which regresses auditory MEG data with respect to the envelopes of the speech streams. We develop an efficient estimation technique by exploiting the sparsity of the TRF and adopting an ℓ_1-regularized least squares estimator which is capable of producing dynamic TRF estimates as well as confidence intervals at sampling resolution from single-trial MEG data. *Results:* We evaluate the performance of our proposed estimator using evoked MEG responses from the human brain in an auditory attention experiment with two competing speakers. The TRFs are estimated dynamically over time using the proposed technique with multisecond resolution, which is a significant improvement over previous results with a temporal resolution of the order of a minute. *Conclusion:* Application of our method to MEG data reveals a precise characterization of the modulation of M50 and M100 evoked responses with respect to the attentional state of the subject at multisecond resolution. *Significance:* Our proposed estimation technique provides a high resolution real-time attention decoding framework in multispeaker environments with potential application in smart hearing aid technology.**

***Index Terms*—Adaptive filtering, attention, auditory processing, magnetoencephalography (MEG), speech segregation.**

Manuscript received September 9, 2016; revised October 28, 2016; accepted November 12, 2016. Date of publication November 15, 2016; date of current version July 15, 2017. *Asterisk indicates corresponding author.*

S. Akram is with the Starkey Hearing Technologies, Berkeley, CA 94704 USA (e-mail: sahar_akram@starkey.com).

J. Z. Simon is with the Department of Electrical and Computer Engineering, Institute for Systems Research, and the Department of Biology, University of Maryland, College Park, MD 20742 USA (e-mail: jzsimon@umd.edu).

*B. Babadi is with the the Department of Electrical and Computer Engineering and Institute for Systems Research, University of Maryland, College Park, MD 20742 USA (e-mail: behtash@umd.edu).

This paper conatins supplementary downloadable material of size 1.6 MB, available at http://ieeexplore.ieee.org (File size: 17.8 MB).

Digital Object Identifier 10.1109/TBME.2016.2628884

I. Introduction

DECODING the dynamics of brain activity underlying conscious behavior is one of the key questions in systems neuroscience. In order to quantify human's conscious experience, neuroimaging techniques such as electroencephalography (EEG) and magnetoencephalography (MEG) are widely used to record the neural activity from the brain with millisecond temporal resolution. From an estimation-theoretic perspective, a decoding framework must be able to reliably estimate the brain activity at a temporal resolution comparable with that of the EEG/MEG acquisition.

A large body of literature in neuroscience has revealed that sensory neurons, such as those in the auditory system, can undergo rapid and task-dependent changes in their response characteristics during attentive behavior, and thereby result in functional changes in the system over time [1]–[5]. In fact, task-based behavioral and neural plasiticy in the auditory cortex can occur within a time-frame of less than a second [6]. Therefore, a dynamic decoding framework on par with the sampling resolution of EEG/MEG is not only important from an estimation-theoretic view point, but is also crucial in order to better understand the neural correlates underlying sophisticated cognitive functions such as attention.

Most of the commonly used estimation methods for characterizing the neural response functions, however, provide static estimates of the spectrotemporal features over significantly longer periods of time compared to the sampling resolution of the neural data. Reverse correlation [7]–[9] and boosting [10]–[13] are two such widely-used techniques for characterizing the spectrotemporal receptive fields (STRFs) in the auditory system. In order to achieve a reliable estimate of the STRF, these methods require performing heavy averaging (i.e., integrating) over time, and in some cases over many trials. Hence these methods cannot measure changes in the STRF occurring on time scales faster than the order of the integration time, typically a minute [14]–[17]. As a result, they are not able to systematically track the aforementioned neural plasticity at a resolution of the order of a second. Moreover, in order to obtain a precise statistical interpretation of the data, it is crucial to compute statistical confidence intervals for the estimates. Confidence intervals are

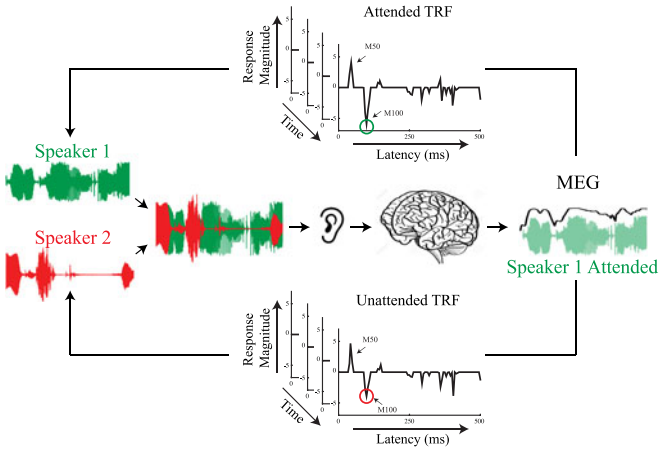

Fig. 1. Schematic depiction of dynamic TRF estimation using evoked MEG response and speech envelopes of the speakers. Here, the auditory scene consists of the mixture of two concurrent speech streams, in which the subject is attending to the first speaker. Earlier studies demonstrated that the significant TRF component corresponding to the M100 response is significantly larger for attended versus unattended speaker.

a key ingredient in assessing changes across different conditions as well as performing hypothesis testing. Current methods use bootstrap resampling over multiple trials [18] in order to compute statistical confidence intervals for the estimated parameters [15], [19]. Although the latter issue is not critical for batch-mode data analysis with multiple trials at hand, when combined with the low-temporal resolution of the estimates, forms a serious bottleneck in emerging real-time applications such as Brain–Computer Interfacing and neural prosthetics, in which only a single trial is available and statistically reliable decoding of neural activity at high temporal resolution is desired.

In this paper, we address these issues for the problem of estimating the auditory temporal response function (TRF) from MEG as a parametric statistical model for the auditory neural response in competing-speaker environments. The TRF can be described as a sparse kernel, relating the auditory neural response recorded via EEG/MEG to the envelopes of the speech streams in the auditory scene (see Fig. 1). The TRF provides an encoding model which generalizes the concept of event-related evoked responses: instead of averaging over multiple trials with the same stimulus to obtain the evoked response, the TRF kernel is obtained by averaging the effect of a diverse set of stimuli, presented as a continuous time-series, and hence results in a generalizable encoding model. It has been shown that in a competing-speaker environment, the slow modulations of the recorded neural response exhibit a higher correlation with the envelope of the attended speaker as opposed to the unattended speaker [6], [16]. Moreover, It has been shown that the amplitude of the well-known M100 auditory evoked response, a prominent robust peak appearing ∼100 ms following the stimulus presentation, is significantly modulated by the attentional state of the listener in experiments with pure tones [20], [21], wide-band noise [19], and speech [15] as the stimuli. In contrast, the amplitude of the M50 evoked response, the earliest auditory evoked response appearing ∼50 ms following the stimulus onset, seems to be task-independent and not modulated by attentional state of the listener [15], [19], [21]. Therefore, in this application obtaining a dynamic estimate of the TRF is equivalent to tracking a robust neural marker of selective auditory attention and thereby characterizing the attentional state of the listener at high temporal resolution.

To this end, we model the TRF over a Gaussian dictionary with time-varying coefficients, where the coefficients are assumed to be sparse. We then adopt an efficient real-time estimation technique, namely the $\ell 1$-regularized least squares (SPARLS) estimator [22], enabling us to compute a dynamic estimate of the TRF over time while enforcing the sparsity of the coefficients. In addition, we develop a novel filter for the recursive computation of the statistical confidence intervals based on recent results in high-dimensional sparse model estimation [23]. Both the estimator and the filter for computing the confidence intervals operate at the sampling resolution from single-trial MEG data. We evaluate the performance of our proposed estimator using evoked MEG responses from the human brain in an auditory attention experiment with two competing speakers. We examine TRF modulations as a function of attentional state and show that the attentional state can be reliably inferred via the estimated TRFs. Our results suggest that tracking the M100 component of the TRF (i.e., its response peak with ∼100 ms latency) in a dynamic fashion can be used as a robust marker of auditory attention in real-time applications.

II. Methods

A. Preliminaries and Motivation

Consider a task where the subject is listening to an acoustic stimulus consisting of two superimposed speech streams. Let the time series $y_1, y_2, \ldots, y_T$ denote the auditory component of the MEG observations (hereafter, it will be referred to as the neural response (see Section II-F). Let $e_n^{(j)}$ be the speech envelope of speaker j, for $j = 1, 2$ at time index n in the dB scale. We take the absolute value of the analytic extension (Hilbert Transform) followed by a low-pass filter with a cutoff frequency of 20 Hz as smoothed estimate of the envelope. In a linear model, the neural response at time index n is related to the envelope of speech as

$$y_n = \left(\boldsymbol{\tau}_n^{(1)}\right)^T \mathbf{e}_n^{(1)} + \left(\boldsymbol{\tau}_n^{(2)}\right)^T \mathbf{e}_n^{(2)} + v_n \qquad (1)$$

where $\boldsymbol{\tau}_n^{(j)}$ is a linear filter of length M denoted by the TRF of speaker j, $\mathbf{e}_n^{(j)} := [e_n^{(j)}, e_{n-1}^{(j)}, \ldots, e_{n-M+1}^{(j)}]^T$ is the covariate vector formed from the speech envelope of speaker j, for $j = 1, 2$, and v_n is a nuisance component accounting for stimulus-independent components manifested in the neural response. It is known that the auditory TRF is a sparse filter, with significant components corresponding to the M50 and M100 auditory responses [15], [16]. One of the commonly-used nonlinear techniques for estimating the TRF is known as Boosting [16], [24], where the components of the TRF are greedily selected to decrease the mean square error (MSE) of the fit to the neural response. The ℓ_1-regularized least squares method (LASSO) has also been recently used to estimate TRF with sparse components [25], [26]. A common shortcoming of these estimation techniques is their inability to track the parameters adaptively. Neural responses in the auditory system are known to be nonstationary (e.g., [1]–[5]), and hence adaptive estimation of their model parameters are crucial in many applications

of interest. In what follows, we will employ an adaptive algorithm for recursive estimation and tracking of the time-varying and sparse TRF based on MEG observations.

B. Adaptive TRF Estimation via the SPARLS Algorithm

Let

$$\boldsymbol{\tau}_i := \left[\left(\boldsymbol{\tau}_i^{(1)}\right)^T, \left(\boldsymbol{\tau}_i^{(2)}\right)^T\right]^T, \text{and } \mathbf{e}_i := \left[\left(\mathbf{e}_i^{(1)}\right)^T, \left(\mathbf{e}_i^{(2)}\right)^T\right]^T$$

be the concatenated TRF and envelope vector of the two speakers, respectively. Let $\widehat{\boldsymbol{\tau}}_i$ be an estimate of $\boldsymbol{\tau}_i$. The instantaneous error of the corresponding filter at time i is defined as

$$\varepsilon_i := y_i - \widehat{\boldsymbol{\tau}}_i^T \mathbf{e}_i. \tag{2}$$

The adaptive filtering operation at time n can then be stated as the following optimization problem:

$$\min_{\widehat{\boldsymbol{\tau}}_n} \quad f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \tag{3}$$

where $f \geq 0$ is a cost function. Suppose that noise in the linear model above is i.i.d. Gaussian, i.e., $v_n \sim \mathcal{N}(0, \sigma^2)$. We define the cost function to be

$$f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) := \frac{1}{2\sigma^2} \sum_{i=1}^{n} \lambda^{n-i} |\varepsilon_i|^2 \tag{4}$$

where $0 < \lambda \leq 1$ is a parameter often referred to as the forgetting factor. The forgetting factor λ gives more weight to the more recent filter errors in order to enforce adaptivity. In order to enforce smoothness, we consider a representation for $\boldsymbol{\tau}_n$ over a basis spanned by $\mathbf{G}$:

$$\boldsymbol{\tau}_n = \mathbf{G}\boldsymbol{\theta}_n.$$

Examples of $\mathbf{G}$ are the Gabor or Haar bases [27], [28]. In order to enforce sparsity, we estimate $\widehat{\boldsymbol{\theta}}_n$ by the following ℓ_1-regularized optimization [22]:

$$\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}_n}{\operatorname{argmin}} \frac{1}{2\sigma^2} \left\| \boldsymbol{\Lambda}_n^{1/2} \mathbf{y}_n - \boldsymbol{\Lambda}_n^{1/2} \mathbf{E}_n \mathbf{G} \boldsymbol{\theta}_n \right\|_2^2 + \eta \|\boldsymbol{\theta}_n\|_1 \tag{5}$$

where $\boldsymbol{\Lambda}_n := \operatorname{diag}(\lambda^{n-1}, \lambda^{n-2}, \ldots, 1)$, $\mathbf{y}_n = [y_1, y_2, \ldots, y_n]^T$, $\mathbf{E}_n = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]^T$, and η is a regularization parameter, representing a tradeoff between estimation error and sparsity of the TRF parameters. Note that the quadratic term is the same as $f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ expressed in vector form.

The SPARLS algorithm introduced in [22] provides a recursive solution to convex programs of the form in (5). Here, we briefly give an overview of the SPARLS algorithm (please refer to [22] for details). The idea of the SPARLS algorithm is to use the Expectation–Maximization (EM) algorithm to facilitate the optimization of the cost function. At the ℓth iteration of the EM algorithm, we have

$$\begin{cases} \text{E-step: } \mathbf{r}_n^{(\ell)} := \left(\mathbf{I} - \dfrac{\alpha^2}{\sigma^2} \mathbf{G}^T \mathbf{E}_n^T \boldsymbol{\Lambda}_n \mathbf{E}_n \mathbf{G}\right) \widehat{\boldsymbol{\theta}}_n^{(\ell)} + \dfrac{\alpha^2}{\sigma^2} \mathbf{G}^T \mathbf{E}_n^T \boldsymbol{\Lambda}_n \mathbf{y}_n \\ \text{M-step: } \widehat{\boldsymbol{\theta}}_n^{(\ell+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{ -\dfrac{1}{2\alpha^2} \left\| \mathbf{r}_n^{(\ell)} - \boldsymbol{\theta} \right\|_2^2 - \eta \|\boldsymbol{\theta}\|_1 \right\} \end{cases} \tag{6}$$

Algorithm 1: SPARLS$(b_n, \mathbf{x}_n, \mathbf{B}_{n-1}, \mathbf{u}_{n-1}, \widehat{\boldsymbol{\theta}}_{n-1}, \lambda, \eta, \alpha, \sigma)$.

Input : Observation b_n, covariate vector $\mathbf{x}_n$, $\mathbf{B}_{n-1}$, $\mathbf{u}_{n-1}$, $\widehat{\boldsymbol{\theta}}_{n-1}$, forgetting factor λ, regularization coefficient η, maximum number of iterations for EM algorithm K (default $K = 1$), α, and σ.

$\mathbf{B}_n = \lambda \mathbf{B}_{n-1} - (\frac{\alpha}{\sigma})^2 \mathbf{x}_n \mathbf{x}_n^T + (1-\lambda)\mathbf{I}$
$\mathbf{u}_n = \lambda \mathbf{u}_{n-1} + (\frac{\alpha}{\sigma})^2 b_n \mathbf{x}_n$
$\widehat{\boldsymbol{\theta}}_n^{(0)} = \widehat{\boldsymbol{\theta}}_{n-1}$
EM iterations:
for $\ell = 1, 2, \ldots, K$ **do**
  $\mathbf{r}_n^{(\ell-1)} = \mathbf{B}_n \widehat{\boldsymbol{\theta}}_n^{(\ell-1)} + \mathbf{u}_n$
  $\widehat{\boldsymbol{\theta}}_n^{(\ell)} = \mathcal{S}_{\eta\alpha^2}\left(\mathbf{r}_n^{(\ell-1)}\right)$
end
output: $\widehat{\boldsymbol{\theta}}_n \leftarrow \widehat{\boldsymbol{\theta}}_n^{(K)}$

where $\alpha^2 < \sigma^2$ is a step-size parameter. The M-step has a closed form solution $\widehat{\boldsymbol{\theta}}_n^{(\ell+1)} = \mathcal{S}_{\eta\alpha^2}(\mathbf{r}_n^{(\ell)})$, where the element-wise operator $\mathcal{S}_\tau(.) : \mathbb{R}^M \to \mathbb{R}^M$ is known as soft thresholding and whose ith component is defines as

$$(\mathcal{S}_\tau(\mathbf{x}))_i := \operatorname{sgn}(x_i) \max(|x_i| - \tau, 0) \tag{7}$$

for $i = 1, 2, \ldots, M$. By defining

$$\mathbf{B}_n := \mathbf{I} - \frac{\alpha^2}{\sigma^2} \mathbf{G}^T \mathbf{E}_n^T \boldsymbol{\Lambda}_n \mathbf{E}_n \mathbf{G}, \text{ and } \mathbf{u}_n := \frac{\alpha^2}{\sigma^2} \mathbf{G}^T \mathbf{E}_n^T \boldsymbol{\Lambda}_n \mathbf{y}_n \tag{8}$$

it is easy to see that $\mathbf{r}_n^{(\ell)} = \mathbf{B}_n \widehat{\boldsymbol{\theta}}_n^{(\ell)} + \mathbf{u}_n$. In turn, $\mathbf{B}_n$ and $\mathbf{u}_n$ can be recursively updated by

$$\begin{cases} \mathbf{B}_n = \lambda \mathbf{B}_{n-1} - \frac{\alpha^2}{\sigma^2} \mathbf{G}^T \mathbf{e}_n \mathbf{e}_n^T \mathbf{G} + (1-\lambda)\mathbf{I} \\ \mathbf{u}_n = \lambda \mathbf{u}_{n-1} + \frac{\alpha^2}{\sigma^2} y_n \mathbf{e}_n \mathbf{G} \end{cases} \tag{9}$$

which results in simple recursive rules to adaptively carry out the EM algorithm. Algorithm 1 summarizes the SPARLS algorithm. Finally, given the estimate $\widehat{\boldsymbol{\theta}}_n$, the TRF estimate is defined as $\widehat{\boldsymbol{\tau}}_n := \mathbf{G}\widehat{\boldsymbol{\theta}}_n$, for all n.

C. Adaptive Estimation of Confidence Intervals

Obtaining statistical confidence intervals for the estimated TRF values is of utmost importance for inference purposes. Most of the commonly used estimation methods require averaging over multiple trials or bootstrap resampling from a limited number of observations to produce confidence regions [18]. In real-time applications, however, statistical confidence intervals must be computed from a single observation stream. In our setting, where a listener is attending to a speaker in a competing speaker environment, it is also desired to compute the confidence intervals for the estimated TRF in real time, in order to test for the reliability and precision of the inferred TRF dynamics.

To this end, we need to overcome two main challenges: first, the ℓ_1-regularized least squares estimate is biased, and hence the commonly-used asymptotic normality assumptions cannot be used for obtaining confidence intervals. Second, the confidence intervals need to be computed recursively in accordance with the real time requirements. In order to address these issues, we take advantage of recent results in high-dimensional statistics for characterizing the confidence bounds for ℓ_1-regularized maximum likelihood (ML) problems [23]. In [23], a procedure called "de-sparsifying" is introduced to account for the bias in ℓ_1-regularized ML estimates, by inverting the Karush–Kuhn–Tucker (KKT) conditions. In our setting, the de-sparsified estimate of $\boldsymbol{\theta}_n$ can be expressed as

$$\widehat{\boldsymbol{\theta}}_n^u := \widehat{\boldsymbol{\theta}}_n + \boldsymbol{\Theta}_n \mathbf{G}^T \mathbf{E}_n^T \boldsymbol{\Lambda}_n \left(\mathbf{y}_n - \mathbf{E}_n \mathbf{G} \widehat{\boldsymbol{\theta}}_n \right) \tag{10}$$

where $\boldsymbol{\Theta}_n$ is a relaxed approximation to the inverse of $\boldsymbol{\Sigma}_n := \mathbf{G}^T \mathbf{E}_n^T \boldsymbol{\Lambda}_n \mathbf{E}_n \mathbf{G}$, obtained through a procedure known as node-wise regression [29]. Then, an asymptotic point-wise confidence interval for the TRF estimate $(\widehat{\boldsymbol{\tau}}_n)_i = (\mathbf{G}\widehat{\boldsymbol{\theta}}_n)_i$ at a significance level ν can be computed as

$$CI_{i,n} := \pm \Phi^{-1}(1-\nu/2)\sigma \sqrt{\left(\mathbf{G}\boldsymbol{\Theta}_n \widetilde{\boldsymbol{\Sigma}}_n \boldsymbol{\Theta}_n^T \mathbf{G}^T \right)_{i,i}} \tag{11}$$

for $i = 1, \ldots, M$, where $\widetilde{\boldsymbol{\Sigma}}_n := \mathbf{G}^T \mathbf{E}_n^T \boldsymbol{\Lambda}_n^2 \mathbf{E}_n \mathbf{G}$ and $\Phi(.)$ denotes the CDF of the normal distribution. A similar treatment to that used for obtaining the SPARLS recursions, results in a fully recursive algorithm to adaptively estimate the confidence intervals. The corresponding algorithm is summarized in Algorithm 2. In short, the estimation procedure outlined in Algorithm 2, consists of $M+1$ runs of the SPARLS algorithm per time index n. In the first run, $\widehat{\boldsymbol{\theta}}$ is estimated at each time point n. Then, the estimate $\widehat{\boldsymbol{\theta}}_n$ is de-sparsified and confidence intervals are computed via an adaptive version of the node-wise regression procedure in the remaining M runs of the SPARLS algorithm. The details of the derivation of this adaptive algorithm are given in Appendix A, and the choices of the parameters are discussed in Section II-G. A MATLAB implementation of our algorithm applied to a simulated example is archived on the GitHub repository [30].

Algorithm 2: Adaptive estimation of the TRF with confidence intervals.

Input : Neural response $\{\mathbf{y}_n\}_{n=1}^N$, Speech envelope $\{\mathbf{E}_n\}_{n=1}^N$, forgetting factor λ, regularization coefficient η, maximum number of iterations for EM algorithm K, α, σ, $\widetilde{\alpha}$, $\widetilde{\sigma}$, ϵ, and confidence level ν.

Initialization: Initial guess of $\widehat{\boldsymbol{\tau}}_1$,
$\mathbf{B}_1 = \mathbf{I} - (\frac{\alpha}{\sigma})^2 \mathbf{e}_1 \mathbf{G}\mathbf{G}^T \mathbf{e}_1^T$, $\mathbf{u}_1 = (\frac{\alpha}{\sigma})^2 y_1 \mathbf{e}_1 \mathbf{G}$, $\boldsymbol{\Sigma}_1$, $\widetilde{\boldsymbol{\Sigma}}_1$, $\mathbf{d}_1$, and initial guess of $\boldsymbol{\gamma}_{i,1}$, $\boldsymbol{\sigma}_{i,1}$,
$\widetilde{\mathbf{B}}_{i,1} = \mathbf{I} - (\frac{\widetilde{\alpha}}{\widetilde{\sigma}})^2 (\mathbf{G}^T \mathbf{e}_1)_{\setminus i} (\mathbf{G}^T \mathbf{e}_1)_{\setminus i}^T$ and
$\widetilde{\mathbf{u}}_{i,n} = (\frac{\widetilde{\alpha}}{\widetilde{\sigma}})^2 (\mathbf{G}^T \mathbf{e}_1)_{\setminus i} (\mathbf{G}^T \mathbf{e}_1)_i$, for $i = 1, \ldots, M$.

for all $n \geq 2$ **do**

$\mathsf{SPARLS}\big(y_n, \mathbf{e}_n\mathbf{G}, \mathbf{B}_{n-1}, \mathbf{u}_{n-1}, \boldsymbol{\tau}_{n-1}, \lambda, \eta, \alpha, \sigma\big)$
$\boldsymbol{\Sigma}_n = \lambda \boldsymbol{\Sigma}_{n-1} + \mathbf{G}^T \mathbf{e}_n \mathbf{e}_n^T \mathbf{G}$
$\mathbf{d}_n = \lambda \mathbf{d}_{n-1} + y_n \mathbf{e}_n \mathbf{G}$

Node-wise Regression

for *i = 1, 2, ..., M* **do**

$\mathsf{SPARLS}\big((\mathbf{G}^T\mathbf{e}_n)_i, (\mathbf{G}^T\mathbf{e}_n)_{\setminus i}, \widetilde{\mathbf{B}}_{i,n-1}, \widetilde{\mathbf{u}}_{i,n-1}, \gamma_{i,n-1}, \lambda, \eta, \widetilde{\alpha}, \widetilde{\sigma}\big)$
$\boldsymbol{\sigma}_{i,n} = \lambda \boldsymbol{\sigma}_{i,n-1} + (\mathbf{G}^T \mathbf{e}_n)_{\setminus i} (\mathbf{G}^T \mathbf{e}_n)_i$
$\omega_{i,n} = (\boldsymbol{\Sigma}_n)_{i,i} - \boldsymbol{\sigma}_{i,n}^T \boldsymbol{\gamma}_{i,n}$

end

Construct $\mathbf{C}_n$ based on $\boldsymbol{\gamma}_{1,n}, \boldsymbol{\gamma}_{2,n}, \ldots, \boldsymbol{\gamma}_{M,n}$. (A.10)
Construct $\boldsymbol{\Theta}_n$ based on $\mathbf{C}_n$ & $\omega_{1,n}, \omega_{2,n}, \ldots, \omega_{M,n}$. (A.11, A.12 & A.13)
$\widehat{\boldsymbol{\tau}}_n = \mathbf{G}\widehat{\boldsymbol{\theta}}_n$
$\widehat{\boldsymbol{\tau}}_n^{\mathsf{u}} = \widehat{\boldsymbol{\tau}}_n + \boldsymbol{\Theta}_n (\mathbf{d}_n - \boldsymbol{\Sigma}_n \widehat{\boldsymbol{\tau}}_n)$
$\widetilde{\boldsymbol{\Sigma}}_n = \lambda^2 \widetilde{\boldsymbol{\Sigma}}_{n-1} + \mathbf{G}^T \mathbf{e}_n \mathbf{e}_n^T \mathbf{G}$
$\mathsf{CI}_{i,n} = \pm \Phi^{-1}(1-\nu/2)\ \epsilon \sqrt{\left(\mathbf{G}\boldsymbol{\Theta}_n \widetilde{\boldsymbol{\Sigma}}_n \boldsymbol{\Theta}_n^T \mathbf{G}^T \right)_{i,i}}$

end

output: $\widehat{\boldsymbol{\tau}}_n$, $\widehat{\boldsymbol{\tau}}_n^{\mathsf{u}}$, $\{\mathsf{CI}_{i,n}\}_{i=1}^M$ for all $n \geq 2$.

D. Subjects, Stimuli, and Procedures

Seven normal-hearing, right-handed young adults (ages between 20 and 31) participated in this study, consisting of two experiments: constant-attention experiment (five subjects, three female) and attention-switch experiment (five subjects, two female). Three subjects (two female) participated in both experiments. Subjects were all 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. This data were used in an earlier study by the authors [26].

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 dichotically 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 s pause (28th to 30th s), 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. Overall, subjects were presented with 12 trials (2 constant-attention/attention-switch $\times$ 2 male/female target speakers $\times$ 3 repetitions) each 1 min long. The time intervals between the trials were randomly chosen between 3 and 5 s. After each condition was presented, subjects answered comprehensive questions related to the passage on which they focused, as a way to keep them motivated on attending to the target speaker. Ninety 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.

E. Data Recording

MEG signals were recorded in a dimly lit magnetically shielded room (Yokogawa Electric Corporation, Tokyo, Japan) 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 $\sim$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/Hz or better in the white noise region.

The presentation software package from Neurobehavioral Systems was used to present stimuli to the subjects. The sounds (approximately 65 dB SPL) were delivered to the participants 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 4000 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 [31]. 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.

F. MEG Processing

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 [32]. DSS is a blind source separation technique that 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. In other words, this algorithm decomposes the data into temporally uncorrelated components by removing inconsistent temporal components that are not phased-locked to the stimulus. The recorded neural response during each trial was band-pass filtered between 1 and 8 Hz and down sampled to 200 Hz before submission to the DSS analysis. We found that only the first DSS component contains a significant amount of stimulus information, so analysis was restricted to this component, which we denote by the auditory MEG component.

G. Parameter Selection

We consider a length of 500 ms for the TRFs at a temporal resolution of 5 ms, resulting in a TRF vector of length $M = 100$. In order to enforce smoothness, we use a dictionary $\mathbf{G}$ consisting of overlapping Gaussian kernels sampled at $\Delta = 5$ ms intervals with means covering 0–500 ms, with 5 ms spacing across TRF length. The standard deviation of the kernels is chosen to be 20 ms, consistent with the average full width at half maximum (FWHM) of an auditory MEG evoked responses (M50 and M100), empirically obtained from MEG studies. The FWHM is given by $2\sqrt{2\ln(2)}\nu \approx 2.355\nu$, where ν is the standard deviation of the gaussian kernel, so choosing $\nu \approx 8.5$, results in a FWHM of order $\approx$20 ms. The corresponding matrix $\mathbf{G}$ is an $M \times M$ matrix of the form $[\mathbf{g}_1^T, \mathbf{g}_2^T, \ldots, \mathbf{g}_M^T]^T$, where $\mathbf{g}_k := [g_\nu(-k\Delta), g_\nu(\Delta - k\Delta), \ldots, g_\nu(M\Delta - k\Delta)]^T$ and $g_\nu(t)$ is the pdf of a zero-mean Gaussian distribution with variance ν^2.

The speech envelopes for speaker 1 and 2 are normalized to have zero mean and variances of $1/M$. To estimate the variance of the observation noise in the linear model, we used the mean squared error from the LASSO estimate of the TRF using the MEG signal and the speech envelope of either speakers. In SPARLS routines used for the node-wise regression, $\widetilde{\sigma}$ is set to $1/\sqrt{2}$ to make $\frac{1}{2\widetilde{\sigma}^2}$ equal to 1, resulting in the ℓ_1-regularized ML optimization problem in A.10.

The parameter α in SPARLS should be chosen such that $\alpha^2 \leq \sigma^2/s$, where s is the largest eigenvalue of $\mathbf{\Lambda}_n^{1/2}\mathbf{E}_n\mathbf{G}\mathbf{G}^T\mathbf{E}_n^T\mathbf{\Lambda}_n^{1/2}$ [22]. Therefore, through an offline tuning, α in the first instance of the SPARLS is chosen to be equal to $\sigma/80$, and $\widetilde{\alpha}$ for the rest of M instances is set to $\widetilde{\sigma}/85$. A choice of $K = 3$ EM iterations is used for all instances of SPARLS algorithm. The forgetting factor parameter, λ, can be fine-tuned according to the time-variation rate of the true TRF in the range of $(0, 1)$. Note that $\lambda = 1$ is used for RLS algorithm, when the true estimate is expected to be constant over time. On the other hand, for smaller values of λ the contribution of the earlier data points in the current estimate of the parameters diminishes and results in wider confidence regions for the estimate. It can be proven that $\left\lceil \frac{1}{1-\lambda} \right\rceil$ samples is the effective data length required for stable estimation of the parameters in LASSO problems with exponentially weighted log-likelihood [33]. We chose $\lambda = 0.999$ throughout the analysis, which corresponds to a data length of $\frac{1}{(1-\lambda)f_s} = 5$ s. Therefore, in order to eliminate the transient effects in estimating the TRFs,

the first and last 5 s of the estimates are discarded. The parameter η, which controls the tradeoff between the sparsity of the estimate and the MSE, was chosen by two-fold cross validation for each subject individually.

III. Results

A. Application to Experimental MEG Data

In order to evaluate the performance of the proposed TRF tracking algorithm, we collected MEG data from multiple subjects, while they were listening to a competing speaker environment. Subjects were required to attend to either of the speakers according to the experimental conditions. We denote by $\widehat{\boldsymbol{\tau}}_n^{\text{att}}$ and $\widehat{\boldsymbol{\tau}}_n^{\text{unatt}}$ the TRFs corresponding to the attended and unattended speakers at time index n, respectively. The TRFs are estimated dynamically for each trial individually, using the first DSS component of the recorded MEG data and the speech envelope of the attended and unattended speakers as the covariates. The field map of the first DSS component is shown in Fig. 2(a) for a sample subject, which shows a dominantly auditory localization. The estimated TRFs have significant peaks at approximately 40–80 ms latency, corresponding to the M50 auditory evoked response. We denote the magnitude of this peak at time n by $\widehat{\tau}_{M50,n}$. This is the earliest response in the auditory cortex that is known to play an important role in the investigation of primary auditory cortex [34], and early auditory system maturation in humans [35], [36]. The M50 response is usually followed by a deflection at about 100 ms latency with respect to the onset of the trial, known as the M100 evoked response. The M100 responses are also detected saliently in the estimated TRFs using the proposed estimation algorithm, and are denoted by $\widehat{\tau}_{M100,n}$.

We investigated the effect of attention on the $\widehat{\tau}_{M50,n}$ and $\widehat{\tau}_{M100,n}$ components in multiple experimental conditions. According to the previous studies, the M100 evoked response is known to be modulated by attention, whereas the M50 evoked response is not attention modulated [19]–[21]. These results were obtained by analyzing grand averaged evoked fields over hundreds of trials and multiple subjects [19], [21], with confidence intervals computed through bootstrap resampling. In these studies, the TRF and the corresponding $\widehat{\tau}_{M50,n}$ and $\widehat{\tau}_{M100,n}$ components were estimated using a boosting algorithm [15], [16], from which an averaged TRF is obtained for a trial of 1 min duration.

In order to validate our proposed TRF estimation technique, we will first confirm the previously obtained results on the effect of attention on M50 and M100 TRF components, and then use attention-modulated component of the TRF as a proxy for decoding the attentional state of the listener in real-time. In the constant-attention experiments (conditions 1 & 2), the $\widehat{\tau}_{M50,n}$ and $\widehat{\tau}_{M100,n}$ responses are extracted from the estimated TRFs, over the time period of 5 to 55 s of each trial. In order to evaluate the effect of attention on these two auditory components at the subject level, the amplitude differences between the attended and unattended conditions, respectively, are computed across time per subject and per trial and are shown in Fig. 2(b1) and (b2). On average over the trial duration and at a confidence level of 95%, the $\widehat{\tau}_{M50,n}$ component of the attended TRF shows a significant difference compared to that of the unattended TRF in only 4 out of the 10 trials (3 decreases and 1 increase), whereas the $|\widehat{\tau}_{M100,n}|$ component is significantly increased in the attended condition in 8 out of the 10 trials.

Fig. 2(c1) and (c2) shows the time course of the M50 and M100 differences, respectively, for all the trials. In order to assess the statistical changes of the M50 and M100 components at the population level across trials, for each trial the time fractions in which the differences are significantly larger, smaller or no different from 0 are computed based on the 95% confidence bounds obtained by our algorithm. Trials are then labeled according to the maximum time fraction computed for each. The amplitude difference of the $\widehat{\tau}_{M50,n}^{\text{att}}$ and $\widehat{\tau}_{M50,n}^{\text{unatt}}$ component, corresponding to the attended and unattended TRFs, respectively, is not significantly different from zero in 81% of the trials. In 9% of the trials $\left|\widehat{\tau}_{M50,n}^{\text{att}}\right|$ is larger than $\left|\widehat{\tau}_{M50,n}^{\text{unatt}}\right|$, and for the remaining 10%, $\left|\widehat{\tau}_{M50,n}^{\text{unatt}}\right|$ is larger than $\left|\widehat{\tau}_{M50,n}^{\text{att}}\right|$. For the $\widehat{\tau}_{M100,n}$ component, the difference between the attended and unattended conditions are significantly positive in 76% of the trials. In 24% of the trials the differences are not significantly different from 0.

Moreover, the average time fractions in which the $\widehat{\tau}_{M100,n}^{\text{att}}$ amplitude is significantly larger than $\widehat{\tau}_{M100,n}^{\text{unatt}}$ is computed in percentage for each trial and for all subjects in both constant-attention and attention-switch experiments. The results are shown in Fig. 3(a1) and (a2). For the $\widehat{\tau}_{M100,n}$ component, the amplitude differences are above the 45° line for 86% of the trials, demonstrating a saliently strong attention modulation in $\widehat{\tau}_{M100,n}$ amplitudes; however, for the $\widehat{\tau}_{M50,n}$ response, the results are more symmetrically scattered above (42%) and below (58%) the 45° line, indicating no significant selectivity to attention. In order to assess the performance gain of our proposed algorithm, we have computed $\widehat{\tau}_{M50,n}$ and $\widehat{\tau}_{M100,n}$ response amplitudes via cross correlation using sliding time windows of length 5 s (the same as the effective time window for our proposed method) with 4.5 s overlaps. The corresponding results are shown in Fig. 3(b1) and (b2). The results obtained by cross correlation show no modulation pattern with respect to the attentional state of the listeners for either the $\widehat{\tau}_{M100,n}$ or $\widehat{\tau}_{M50,n}$ responses. This analysis confirms the performance gain obtained by our sparse adaptive algorithm which in accurately isolating the contributions of the M50 and M100 responses in time and capturing the attentional modulation in the $\widehat{\tau}_{M100,n}$ component.

In order to further highlight the performance gain of our sparse adaptive algorithm, we will next make a performance comparison with the widely used linear least squares estimate of the TRF. To this end, we have estimated the TRFs from the first 30 s of each trial and computed the correlation values between the MEG signal and the model predictions based on speakers 1 and 2 using the remaining 30 s of each trial, via sliding time windows of length 5 s with 4.5 s overlap. The scatter plot of the resulting correlation differences between the models with attended and unattended speakers are shown in Fig. 4. As it can be observed from Fig. 4, the correlation differences are not indicative of the attentional state of the listeners in either the constant-attention or the attention-switch experiments, whereas in Fig. 3(a2), the difference of the $\widehat{\tau}_{M100,n}$ components strongly encodes the attentional modulation. This analysis highlights the advantage of the adaptive estimation of the TRF, as opposed

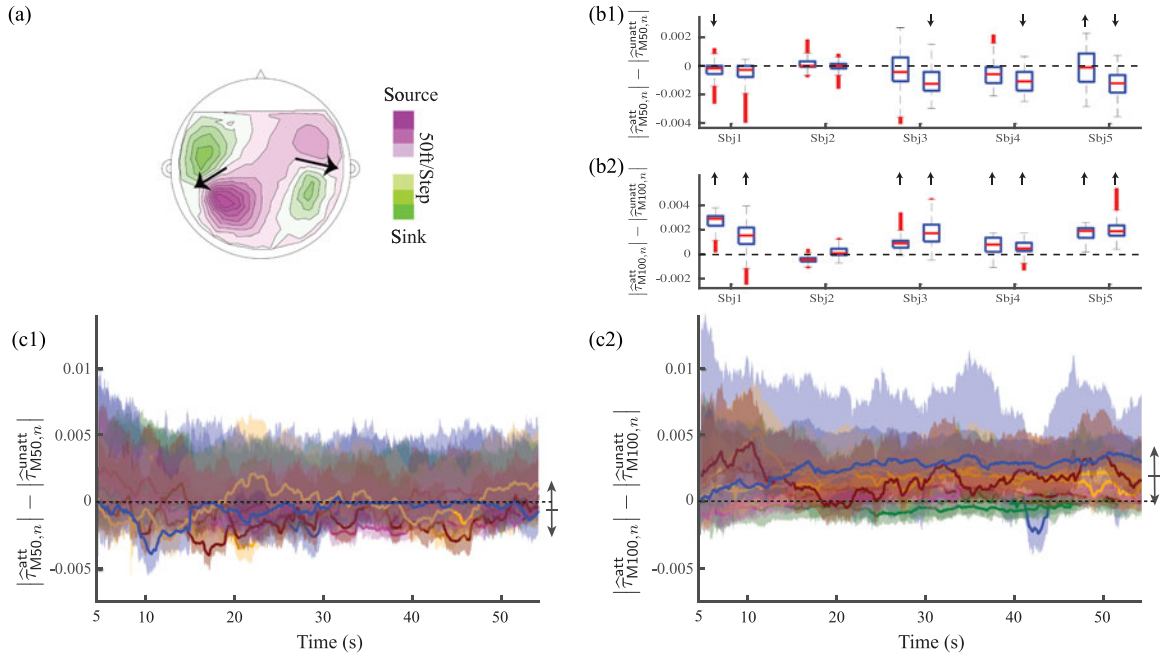

Fig. 2. (a) MEG magnetic field map for the first DSS component of a sample subject. A stereotypical pattern of neural activity, originating separately in the left and right auditory cortices is observed. Black arrows schematically represent the locations of the equivalent dipole currents, generating the measured magnetic field. The $\widehat{\tau}_{M50,n}$ (b1) and $\widehat{\tau}_{M100,n}$ (b2) response amplitude differences for the attended (with superscript att) versus unattended (with superscript $unatt$) conditions are computed for each of the two trials from five participants. Each box plot indicates the statistics of $\widehat{\tau}_{M50,n}$ and $\widehat{\tau}_{M100,n}$ response differences in the constant-attention trials, for all the time points which significantly differ from zero at a confidence level of 95%. On average over the trial duration and at a confidence level of 95%, the $\widehat{\tau}_{M50,n}$ component of the attended TRF shows a significant change compared to its unattended counterpart in only 4 out of 10 trials (3 decreases and 1 increase), whereas $\left|\widehat{\tau}_{M100,n}\right|$ is significantly increased in 8 out of 10 trials. The upward (respectively downward) arrows indicate the trials for which a significant increase (resprectively decrease) in the TRF component differences is observed. Response differences for $\widehat{\tau}_{M50,n}$ (c1) and $\widehat{\tau}_{M100,n}$ (c2) are also plotted over time to demonstrate trackability of response differences with high temporal resolution, along with confidence intervals around each estimation point, using the proposed algorithm. Different colors indicate results from different subjects, where each trace is the averaged response difference over all three trials of each attended condition (speaker 1 or 2), in the constant-attention experiment. The $\widehat{\tau}^{\text{att}}_{M50,n}$ and $\widehat{\tau}^{\text{unatt}}_{M50,n}$ amplitude differences are not significantly different from zero over 81% of the trials; for the remaining 19% they were split almost equally between $\left|\widehat{\tau}^{\text{unatt}}_{M50,n}\right| < \left|\widehat{\tau}^{\text{att}}_{M50,n}\right|$ (9%) and $\left|\widehat{\tau}^{\text{unatt}}_{M50,n}\right| > \left|\widehat{\tau}^{\text{att}}_{M50,n}\right|$ (10%). The $\widehat{\tau}_{M100,n}$ amplitude differences are significantly positive in 76% of the trials. For the remaining 24%, the $\widehat{\tau}^{\text{att}}_{M100,n}$ and $\widehat{\tau}^{\text{unatt}}_{M100,n}$ amplitude differences are not significantly different from 0. In summary, the amplitude differences of the $\widehat{\tau}_{M50}$ responses for both the attended and unattended TRFs are not significantly different from zero, whereas for the $\widehat{\tau}_{M100}$ responses, there are significantly positive amplitude differences between the attended and unattended TRFs. The vertical double-headed arrows indicate the extent of the average amplitude differences between the attended and unattended TRF components at the end of the trial. Error hulls indicate 95% confidence intervals around the estimated parameters.

to the commonly-used static TRF estimation, in capturing the nonstationarity of the underlying neural processes modulated by attention.

The above results confirm that our proposed method improves the earlier approaches in multiple ways: first, the estimated TRF is recursively updated in time, resulting in a multisecond temporal resolution as opposed to an averaged static estimate from minutes of neural data. Second, all the analysis is done on a single trial from a single subject with no averaging over multiple trials or subjects. Last, in the proposed algorithm, confidence intervals are systematically calculated for the estimated parameters and on par with the sampling resolution of the recorded neural data in a single trial, as opposed to the bootstrap resampling technique which requires multiple realizations of the parameters for computing the confidence bounds.

B. Decoding Auditory Attention From TRF Dynamics

We are further interested in employing the attention-modulation characteristic of the $\widehat{\tau}_{M100,n}$ response as an indicator of the attentional state of the listener. Consider the attention-switch experiment in which subjects were asked to attend to one of the speakers for the first half of the trial and then switch their attention to the opposite speaker for the remaining time of that trial. We first use the proposed algorithm to compute the TRFs for both speaker 1 and speaker 2, as outlined in Section III-A. We can then monitor the relative amplitude of the $\widehat{\tau}_{M100,n}$ component in the estimated TRFs and decode which speaker the listener was attending to. This is specifically important for real-time applications, such as the next generation of intelligent hearing aids, in which a real-time and reliable decoding framework for the attentional state of listeners is required, as the neural data recorded via a commercialized EEG device is streaming to the processor of the hearing aid device to improve amplification selectivity.

To better illustrate the dynamic tracking of the $\widehat{\tau}_{M100,n}$ response, two videos from two different subjects consisting of the estimated TRFs for speakers 1 and 2 from a single trial are provided in an attention-switching condition (`TRFVideo1.mov`, and `TRFVideo2.mov`, respectively). In these videos, the

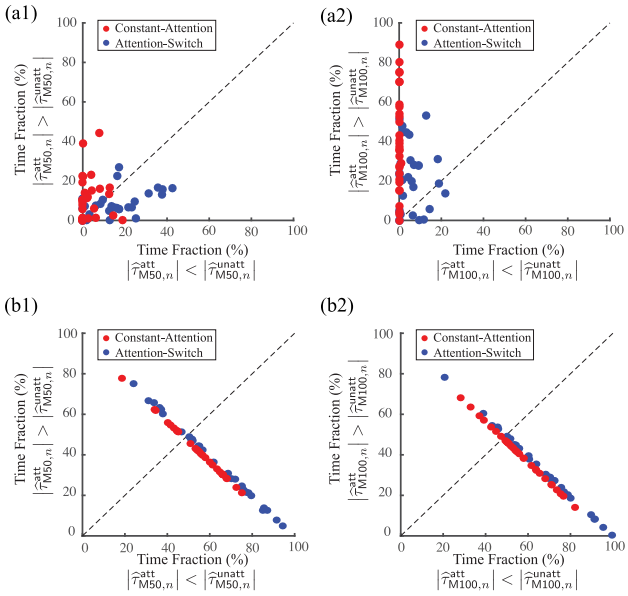

Fig. 3. Amplitude comparison for (a1) $\widehat{\tau}_{M50,n}$ and (a2) $\widehat{\tau}_{M100,n}$ components during attended and unattended conditions. Each circle corresponds to a single trial, and the constant-attention and attention-switch conditions are color coded by red and blue circles, respectively. For each trial, the time fractions in which the amplitude of the auditory component is significantly larger (*y*-axis) or smaller (*x*-axis) in attended versus unattended TRFs are computed in percentage. The dashed line (45° line) corresponds to the condition that the TRF component is not modulated by selective attention. For the $\widehat{\tau}_{M100,n}$ component, the amplitude differences are above the 45° line for 86% of the trials, demonstrating a saliently strong attention modulation in $\widehat{\tau}_{M100,n}$ amplitudes; however, for the $\widehat{\tau}_{M50,n}$ response, the results are more symmetrically scattered above (42%) and below (58%) the 45° line, implying no significant selectivity to attention. Scatter plots showing the corresponding results obtained via the cross-correlation method are presented for the $\widehat{\tau}_{M50,n}$ (b1) and $\widehat{\tau}_{M100,n}$ (b2) components. Cross correlation was performed on each trial using a sliding time window of length 5 with 4.5 s overlap between the successive windows. No attention modulation pattern is detected in either the $\widehat{\tau}_{M50,n}$ or the $\widehat{\tau}_{M100,n}$ responses.

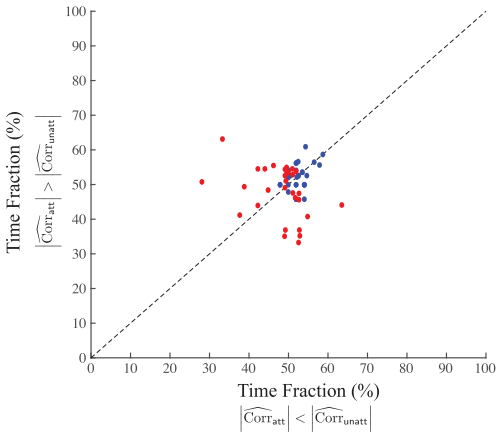

Fig. 4. Scatter plot of the correlations between the MEG signal and the model predictions of the attended and unattended speakers using the commonly-used static TRF estimates obtained by the least squares technique. Each circle corresponds to a single trial, and the constant-attention and attention-switch conditions are color coded by red and blue circles, respectively. For each trial, the time fractions in which the correlation values are significantly larger (*y*-axis) or smaller (*x*-axis) for the attended versus unattended speakers are computed in percentage. The dashed line (45° line) corresponds to the condition that the correlations are not modulated by selective attention.

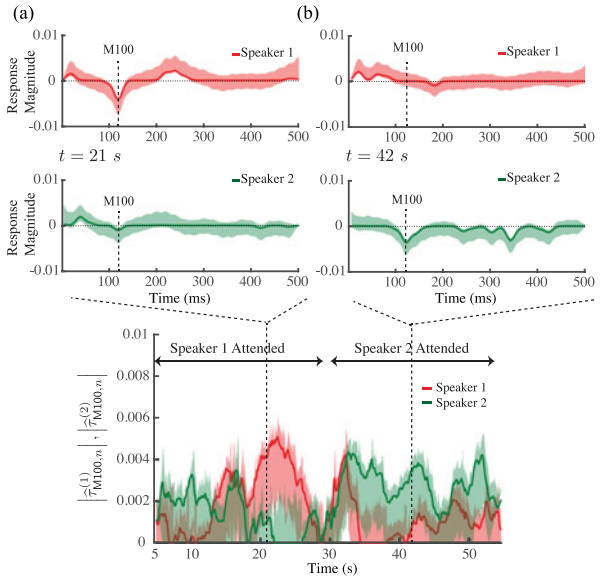

Fig. 5. Tracking the attentional state through the estimated $\widehat{\tau}_{M100,n}$ amplitudes. Results are shown for a sample subject. Bottom panel: the amplitude of the $\widehat{\tau}_{M100,n}$ response for the estimated TRFs from speaker 1 and speaker 2 are plotted as a function of time (5 to 55 ms) in red and green, respectively. According to the $\widehat{\tau}_{M100,n}$ amplitude comparisons, the attention switch occurs at around 15 s after the onset of the trial. Top panel: The TRF estimates for both speakers at times 21 and 42 s are shown in the insets A and B, respectively. The putative temporal location of the $\widehat{\tau}_{M100,n}$ components are indicated via the dash lines in each subplot. Error hulls indicate 95% confidence intervals around the estimated parameters.

output of the algorithm for each of the estimated TRFs is plotted as a function of time. Subjects are required to stay attended to the speaker 1 for the first 28 s of the trial and switch to the second speaker for the second half (the target speaker is specified with "Instructed to Attend" in green). The $\widehat{\tau}_{M100,n}$ responses for each of the TRFs are circled in blue as they appear significantly different from the base line. The middle panel shows the instantaneous $\widehat{\tau}_{M100,n}$ responses, corresponding to the attended and unattended speakers and reveals the abrupt change of the relative $\widehat{\tau}_{M100,n}$ amplitudes occurring around the switching time.

Fig. 5 shows the time course of the estimated $\widehat{\tau}_{M100,n}$ components corresponding to the two speakers, as well as the snapshots of the TRFs at $t = 21$ s and $t = 42$ s for one of the two subjects (corresponding to `TRFVideo1.mov`). In this sample trial, the estimated $\widehat{\tau}_{M100,n}$ for speaker 1 appears to be significantly larger compared to the estimated $\widehat{\tau}_{M100,n}$ component for speaker 2 in the time range of $t = 13$ to 27 s, whereas the opposite is true for the time range of $t = 33$ to 55 s. Note that there is no objective behavioral measure to capture the true switching moment for each trial. Therefore, subjects might perform the switching a little earlier or later than the presentation of the cue (a 2 s pause at 28 s).

IV. Conclusion

The goal of this study is to develop a dynamic estimation framework for computing auditory TRF from noninvasive neural recordings, with the capability of tracking the ongoing changes in the brain activity in an attention-modulated auditory task, which is shown to be correlated with cognitive behavioral and

perceptual changes in human listeners. In a competing speaker environment with a male and female speaker, subjects were instructed to attend to one of the speakers while ignoring the other in multiple experimental conditions. The TRFs estimated from the MEG data revealed a strong attentional modulation in the component with ∼100 ms lag. This component is considered to be analogous to the M100 evoked auditory response that is known to be modulated by attention [15], [16], [19].

The main achievement of this study is developing an adaptive TRF estimation technique which outperforms commonly used batch-mode estimators—such as those based on reverse correlation, boosting and LASSO—in its ability to track the attentional modulation of the estimated TRF, on par with the sampling resolution of the recorded neural data. To this end, we have used a recursive ℓ_1-regularized least squares (SPARLS) approach, based on an EM-type algorithm that provides a considerable performance gain over the conventional linear estimation techniques.

Our TRF estimation technique is complemented with a novel adaptive filter for computing statistical confidence intervals of the estimated parameters that are recursively updated as the data are incoming. In contrast, for commonly-used estimation techniques, the confidence intervals are computed through heavy averaging of the neural data over time and multiple trials. Therefore, these techniques are not suitable for real-time applications in which statistically reliable estimates of the response with high temporal resolution are required from a single trial. Our proposed filter builds up on recent advances in theoretical statistics on constructing confidence intervals for regularized estimates of high-dimensional linear models. A MATLAB implementation of our algorithm is archived on the GitHub repository [30].

Application of the proposed estimation technique on experimentally acquired MEG data suggests that this technique is a strong candidate for an attention decoder in multispeaker environments and can reliably identify the attended speaker with multisecond resolution in time. The promising performance of the proposed algorithm on MEG recordings makes it an appealing candidate for EEG applications, which forms the future direction of this research.

Appendix A
Recursive Estimation of Confidence Intervals Using Node-Wise Regression

Recall the optimization problem from the linear model

$$\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}_n}{\operatorname{argmin}} \frac{1}{2\sigma^2} \left\| \boldsymbol{\Lambda}_n^{1/2}\mathbf{y}_n - \boldsymbol{\Lambda}_n^{1/2}\mathbf{E}_n\mathbf{G}\boldsymbol{\theta}_n \right\|_2^2 + \eta\|\boldsymbol{\theta}_n\|_1. \tag{A.1}$$

The minimizer $\widehat{\boldsymbol{\theta}}_n$ satisfies the KKT conditions given by

$$-\mathbf{G}^T\mathbf{E}_n^T\boldsymbol{\Lambda}_n^{1/2}\left(\boldsymbol{\Lambda}_n^{1/2}\mathbf{y}_n - \boldsymbol{\Lambda}_n^{1/2}\mathbf{E}_n\mathbf{G}\widehat{\boldsymbol{\theta}}_n\right) + \sigma^2\eta\mathbf{g}_n = 0 \tag{A.2}$$

where $\mathbf{g}_n$ is a subgradiant of the ℓ_1-norm at $\widehat{\boldsymbol{\theta}}_n$. Substituting (1) in the above equation, we get

$$\boldsymbol{\Sigma}_n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) + \sigma^2\eta\mathbf{g}_n = \mathbf{E}_n^T\boldsymbol{\Lambda}_n\mathbf{v}_n \tag{A.3}$$

where $\boldsymbol{\tau}_n$ indicates the true TRF and $\boldsymbol{\Sigma}_n := \mathbf{G}^T\mathbf{E}_n^T\boldsymbol{\Lambda}_n\mathbf{E}_n\mathbf{G}$. Let $\boldsymbol{\Theta}_n$ be an approximation to the inverse of $\boldsymbol{\Sigma}_n$, then,

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n + \sigma^2\eta\boldsymbol{\Theta}_n\mathbf{g}_n = \boldsymbol{\Theta}_n\mathbf{E}_n^T\boldsymbol{\Lambda}_n\mathbf{v}_n - \boldsymbol{\Delta}_n \tag{A.4}$$

where

$$\boldsymbol{\Delta}_n := \boldsymbol{\Lambda}_n^{1/2}(\boldsymbol{\Theta}_n\boldsymbol{\Sigma}_n - \mathbf{I})(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n). \tag{A.5}$$

In [23], it is shown that $\boldsymbol{\Delta}_n$ is asymptotically negligible under suitable sparsity assumptions. So, we can have the following unbiased estimator:

$$\widehat{\boldsymbol{\theta}}_n^u = \widehat{\boldsymbol{\theta}}_n + \boldsymbol{\Theta}_n\mathbf{G}^T\mathbf{E}_n^T\boldsymbol{\Lambda}_n(\mathbf{y}_n - \mathbf{E}_n\mathbf{G}\widehat{\boldsymbol{\theta}}_n). \tag{A.6}$$

Therefore, asymptotically we have

$$\widehat{\boldsymbol{\theta}}_n^u - \boldsymbol{\theta}_n = \boldsymbol{\Theta}_n\mathbf{G}^T\mathbf{E}_n^T\boldsymbol{\Lambda}_n\mathbf{v}_n \tag{A.7}$$

which implies that $\widehat{\boldsymbol{\theta}}_n^u - \boldsymbol{\theta}_n$ converges to a Gaussian distribution with mean zero and covariance $\sigma^2\left(\boldsymbol{\Theta}_n\widetilde{\boldsymbol{\Sigma}}_n\boldsymbol{\Theta}_n^T\right)$, where $\widetilde{\boldsymbol{\Sigma}}_n := \mathbf{G}^T\mathbf{E}_n^T\boldsymbol{\Lambda}_n^2\mathbf{E}_n\mathbf{G}$. Noting that $\widehat{\boldsymbol{\tau}}_n = \mathbf{G}\widehat{\boldsymbol{\theta}}_n$, the asymptotic pointwise confidence interval for $(\widehat{\boldsymbol{\tau}}_n^u)_i$ is given by

$$\pm\Phi^{-1}(1-\nu/2)\,\sigma\sqrt{\left(\mathbf{G}\boldsymbol{\Theta}_n\widetilde{\boldsymbol{\Sigma}}_n\boldsymbol{\Theta}_n^T\mathbf{G}^T\right)_{i,i}} \tag{A.8}$$

where $\Phi(.)$ denotes the CDF of $\mathcal{N}(0,1)$.

Next, we will derive a recursive formulation for computing the approximate inverse $\boldsymbol{\Theta}_n$. Let $(\mathbf{GE}_n)_j$ versus $(\mathbf{GE}_n)_{\setminus j}$, denote to the jth column and the submatrix of $\mathbf{GE}_n$ with the jth column removed, respectively. The node-wise regression for the jth column of $\mathbf{GE}_n$ corresponds to computing [29]:

$$\widehat{\boldsymbol{\gamma}}_{j,n} := \underset{\boldsymbol{\gamma}\in\mathbb{R}^{M-1}}{\operatorname{argmin}} \left\{ \|\boldsymbol{\Lambda}_n^{1/2}(\mathbf{GE}_n)_j - (\mathbf{GE}_n)_{\setminus j}\boldsymbol{\Lambda}_n^{1/2}\boldsymbol{\gamma}\|_2^2 + \eta\|\boldsymbol{\gamma}\|_1 \right\} \tag{A.9}$$

where $\widehat{\boldsymbol{\gamma}}_{j,n}$ is a vector of coefficients for a sparse representation of the jth column of $\mathbf{GE}_n$ in terms of the rest of the columns. Note that the computation of $\widehat{\boldsymbol{\gamma}}_{j,n}$ can be carried out recursively using the SPARLS algorithm, for all $j = 1, 2, \ldots, M$ in parallel. The coefficients $\widehat{\boldsymbol{\gamma}}_{j,n}$ are then used to form the following matrix:

$$\mathbf{C}_n := \begin{bmatrix} 1 & -(\widehat{\boldsymbol{\gamma}}_{2,n})_1 & \cdots & -(\widehat{\boldsymbol{\gamma}}_{M,n})_1 \\ -(\widehat{\boldsymbol{\gamma}}_{1,n})_2 & 1 & \cdots & -(\widehat{\boldsymbol{\gamma}}_{M,n})_2 \\ \vdots & \vdots & \ddots & \vdots \\ -(\widehat{\boldsymbol{\gamma}}_{1,n})_M & -(\widehat{\boldsymbol{\gamma}}_{M,n})_M & \cdots & 1 \end{bmatrix}. \tag{A.10}$$

For $j = 1, 2, \ldots, M$, letting

$$\omega_{j,n}^2 := \left(\boldsymbol{\Lambda}_n^{1/2}(\mathbf{GE}_n)_j - \boldsymbol{\Lambda}_n^{1/2}(\mathbf{GE}_n)_{\setminus j}\widehat{\boldsymbol{\gamma}}_{j,n}\right)^T\boldsymbol{\Lambda}_n^{1/2}(\mathbf{GE}_n)_j, \tag{A.11}$$

and

$$\mathbf{T}_n^2 := \operatorname{diag}(\omega_{1,n}^2, \omega_{2,n}^2, \ldots, \omega_{M,n}^2) \tag{A.12}$$

the approximate inverse matrix $\boldsymbol{\Theta}_n$ is defined as

$$\boldsymbol{\Theta}_n := \mathbf{T}_n^{-2}\mathbf{C}_n. \tag{A.13}$$

Finally, using the recursions

$$\boldsymbol{\Sigma}_n = \lambda \boldsymbol{\Sigma}_{n-1} + \mathbf{G}^T \mathbf{e}_n \mathbf{e}_n^T \mathbf{G}$$

$$\widetilde{\boldsymbol{\Sigma}}_n = \lambda^2 \widetilde{\boldsymbol{\Sigma}}_{n-1} + \mathbf{G}^T \mathbf{e}_n \mathbf{e}_n^T \mathbf{G}$$

$$\boldsymbol{\sigma}_{i,n} = \lambda \boldsymbol{\sigma}_{i,n-1} + (\mathbf{G}^T \mathbf{e}_n)_{\setminus i} (\mathbf{G}^T \mathbf{e}_n)_i$$

and noting that $\omega_{i,n} = (\boldsymbol{\Sigma}_n)_{i,i} - \boldsymbol{\sigma}_{i,n}^T \boldsymbol{\gamma}_{i,n}$, the recursive steps given in Algorithm 2 follow. Please refer to [23] for the technical details regarding the above-mentioned asymptotic result.

ACKNOWLEDGMENT

This work was supported in part by the National Institutes of Health Award No. 1R01AG036424 and the National Science Foundation Award No. 1552946. We would like to thank Alessandro Presacco for providing us with a subset of the data used in this paper.

REFERENCES

[1] J. Fritz *et al.*, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," *Nature Neurosci.*, vol. 6, no. 11, pp. 1216–1223, 2003.
[2] J. Fritz, M. Elhilali, and S. Shamma, "Active listening: Task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex," *Hearing Res.*, vol. 206, no. 1, pp. 159–176, 2005.
[3] C. E. Schreiner and J. A. Winer, "Auditory cortex mapmaking: Principles, projections, and plasticity," *Neuron*, vol. 56, no. 2, pp. 356–365, 2007.
[4] S. Atiani, M. Elhilali, S. V. David, J. B. Fritz, and S. A. Shamma, "Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields," *Neuron*, vol. 61, no. 3, pp. 467–480, 2009.
[5] J. Ahveninen *et al.*, "Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise," *Proc. Nat. Academy Sci.*, vol. 108, no. 10, pp. 4182–4187, 2011.
[6] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.
[7] F. E. Theunissen *et al.*, "Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli," *Netw. Comput. Neural Syst.*, vol. 12, no. 3, pp. 289–316, 2001.
[8] D. Smyth *et al.*, "The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation," *J. Neurosci.*, vol. 23, no. 11, pp. 4746–4759, 2003.
[9] C. K. Machens, M. S. Wehr, and A. M. Zador, "Linearity of cortical receptive fields measured with natural sounds," *J. Neurosci.*, vol. 24, no. 5, pp. 1089–1100, 2004.
[10] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors)," *Ann. Statist.*, vol. 28, no. 2, pp. 337–407, 2000.
[11] J. Friedman *et al.*, "Discussion of boosting papers," *Ann. Statist.*, vol. 32, pp. 102–107, 2004.
[12] T. Zhang and B. Yu, "Boosting with early stopping: Convergence and consistency," *Ann. Statist.*, vol. 33, pp. 1538–1579, 2005.
[13] S. V. David, N. Mesgarani, and S. A. Shamma, "Estimating sparse spectro-temporal receptive fields with natural stimuli," *Netw. Comput. Neural Syst.*, vol. 18, no. 3, pp. 191–212, 2007.
[14] B. N. Pasley *et al.*, "Reconstructing speech from human auditory cortex," *PLoS Biol.*, vol. 10, no. 1, 2012, Art. no. e1001251.
[15] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proc. Nat. Academy Sci.*, vol. 109, no. 29, pp. 11 854–11 859, 2012.
[16] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *J. Neuriphysiol.*, vol. 107, no. 1, pp. 78–89, 2012.
[17] J. A. O'Sullivan *et al.*, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.
[18] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL, USA: CRC Press, 1994.
[19] M. Chait, J. Z. Simon, and D. Poeppel, "Auditory M50 and M100 responses to broadband noise: Functional implications," *NeuroReport*, vol. 15, no. 16, pp. 2455–2458, 2004.
[20] K. Lange, F. Rösler, and B. Röder, "Early processing stages are modulated when auditory stimuli are presented at an attended moment in time: An event-related potential study," *Psychophysiology*, vol. 40, no. 5, pp. 806–817, 2003.
[21] M. Chait *et al.*, "Neural dynamics of attending and ignoring in human auditory cortex," *Neuropsychologia*, vol. 48, no. 11, pp. 3262–3271, 2010.
[22] B. Babadi, N. Kalouptsidis, and V. Tarokh, "Sparls: The sparse RLS algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4013–4025, Aug. 2010.
[23] S. Van de Geer *et al.*, "On asymptotically optimal confidence regions and tests for high-dimensional models," *Ann. Statist.*, vol. 42, no. 3, pp. 1166–1202, 2014.
[24] S. V. David, N. Mesgarani, and S. A. Shamma, "Selective cortical representation of attended speaker in multi-talker speech perception," *Netw. Comput. Neural Syst.*, vol. 18, no. 3, pp. 191–221, 2007.
[25] S. Akram *et al.*, "A state-space model for decoding auditory attentional modulation from meg in a competing-speaker environment," in *Proc. Conf. Adv. Neural Inf. Process. Syst. 27*, 2014, pp. 460–468. [Online]. Available: http://papers.nips.cc/paper/
[26] S. Akram *et al.*, "Robust decoding of selective auditory attention from meg in a competing-speaker environment via state-space modeling," *NeuroImage*, vol. 124, pp. 906–917, 2016.
[27] H. G. Feichtinger and T. Strohmer, *Gabor Analysis and Algorithms: Theory and Applications*. New York, NY, USA: Springer, 2012.
[28] A. Haar, "Zur theorie der orthogonalen funktionensysteme," *Mathematische Annalen*, vol. 69, no. 3, pp. 331–371, 1910.
[29] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, 2006.
[30] *Real Time Estimation of Auditory Temporal Response Function MATLAB Code*. 2016. [Online]. Available on GitHub Repository: https://github.com/saharakram/Real-Time-Estimation-of-Auditory-Temporal-Response-Function
[31] A. de Cheveigné and J. Z. Simon, "Denoising based on time-shift PCA," *J. Neurosci. Methods*, vol. 165, no. 2, pp. 297–305, 2007.
[32] A. de Cheveigné and J. Z. Simon, "Denoising based on spatial filtering," *J. Neurosci. Methods*, vol. 171, no. 2, pp. 331–339, 2008.
[33] A. Sheikhattar *et al.*, "Recursive sparse point process regression with application to spectrotemporal receptive field plasticity analysis," *IEEE Trans. Signal Process.*, vol. 64, no. 8, pp. 2026–2039, Apr. 2016.
[34] A. Rupp *et al.*, "The representation of peripheral neural activity in the middle-latency evoked field of primary auditory cortex in humans," *Hearing Res.*, vol. 174, no. 1, pp. 19–31, 2002.
[35] J. E. O. Cardy *et al.*, "Prominence of M50 auditory evoked response over M100 in childhood and autism," *NeuroReport*, vol. 15, no. 12, pp. 1867–1870, 2004.
[36] D. S. Beal *et al.*, "Speech-induced suppression of evoked auditory fields in children who stutter," *NeuroImage*, vol. 54, no. 4, pp. 2994–3003, 2011.

Sahar Akram's photograph and biography not available at the time of publication.

Jonathan Z. Simon's photograph and biography not available at the time of publication.

Behtash Babadi's photograph and biography not available at the time of publication.