



Multivariate autoregressive models with exogenous inputs for intracerebral responses to direct electrical stimulation of the human brain

Jui-Yang Chang¹, Andrea Pigorini², Marcello Massimini², Giulio Tononi³, Lino Nobili⁴ and Barry D. Van Veen^{1*}

¹ Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI, USA

² Department of Clinical Sciences, University of Milan, Milan, Italy

³ Department of Psychiatry, University of Wisconsin, Madison, WI, USA

⁴ Centre of Epilepsy Surgery "C. Munari," Niguarda Hospital, Milan, Italy

Edited by:

Keiichi Kitajo, RIKEN Brain Science Institute, Japan

Reviewed by:

Stefan Haufe, Berlin Institute of Technology, Germany

Adam Barrett, University of Sussex, UK

*Correspondence:

Barry D. Van Veen, Department of Electrical and Computer Engineering, University of Wisconsin, 3611 Engineering Hall, 1415 Engineering Drive, Madison, WI 53706, USA.
e-mail: vanveen@engr.wisc.edu

A multivariate autoregressive (MVAR) model with exogenous inputs (MVARX) is developed for describing the cortical interactions excited by direct electrical current stimulation of the cortex. Current stimulation is challenging to model because it excites neurons in multiple locations both near and distant to the stimulation site. The approach presented here models these effects using an exogenous input that is passed through a bank of filters, one for each channel. The filtered input and a random input excite a MVAR system describing the interactions between cortical activity at the recording sites. The exogenous input filter coefficients, the autoregressive coefficients, and random input characteristics are estimated from the measured activity due to current stimulation. The effectiveness of the approach is demonstrated using intracranial recordings from three surgical epilepsy patients. We evaluate models for wakefulness and NREM sleep in these patients with two stimulation levels in one patient and two stimulation sites in another resulting in a total of 10 datasets. Excellent agreement between measured and model-predicted evoked responses is obtained across all datasets. Furthermore, one-step prediction is used to show that the model also describes dynamics in pre-stimulus and evoked recordings. We also compare integrated information—a measure of intracortical communication thought to reflect the capacity for consciousness—associated with the network model in wakefulness and sleep. As predicted, higher information integration is found in wakefulness than in sleep for all five cases.

Keywords: intracerebral EEG, evoked response, MVARX model, cross-validation, integrated information

1. INTRODUCTION

The remarkable cognitive abilities of the healthy human brain depend on an exquisite balance between functional specialization of local cortical circuits and their functional integration through long-range connections. Hence, there is considerable interest in characterizing long-range cause and effect or directional interactions in the human brain. Multivariate autoregressive (MVAR) models, sometimes referred to as vector autoregressive (VAR) models, have been widely applied to study directional cortical network properties from both intracranial data (e.g., Bernasconi and König, 1999; Brovelli et al., 2004; Winterhalder et al., 2005; Ding et al., 2006; Korzeniewska et al., 2008) and scalp EEG or MEG (e.g., Babiloni et al., 2005; Malekpour et al., 2012). An MVAR model describes each signal as a weighted combination of its own past values and the past values of other signals in the model—an autoregression—plus an error term. The weights relating the present of one signal to the past of another capture the causal or directed influence between signals. A variety of different metrics for summarizing the directed interactions in MVAR models have been proposed, including directed transfer functions (Kamiński and Blinowska, 1991), directed coherence

(Baccalá and Sameshima, 2001), conditional Granger causality (Geweke, 1984), and integrated information (Barrett and Seth, 2011).

MVAR models assume the data is stationary and of constant mean. While stationarity and constant mean may be reasonable assumptions for a relatively short duration of spontaneous data, evoked or event-related data appear to violate these assumptions. For example, the mean or average response to a stimulus varies with time. An MVAR model fit to data with a time-varying mean results in spurious interactions because the assumption of stationarity is violated. Adaptive or time-varying methods have been developed to relax stationarity assumptions (Ding et al., 2000; Möller et al., 2001; Astolfi et al., 2008). For example, a time-varying mean response is removed by subtracting the ensemble average (Ding et al., 2000) and the MVAR model parameters are allowed to vary with time. Adaptive models require specification of an adaptation rate parameter that effectively determines how much of the past data is used to estimate the present model parameters, or equivalently, how fast the model is changing. Models that use fast adaptation are able to track faster changes in the underlying data, but employ less data to estimate model

parameters and consequently possess more variability in the estimated model parameters (see Astolfi et al., 2008, for assessment of these issues).

During the pre-surgical evaluation of drug-resistant epileptic patients, direct electrical stimulation of the brain is systematically performed for diagnostic purposes to identify the epileptogenic zone (Munari et al., 1994). Electrical stimulation generates a time-varying response at the recording sites. In this paper we propose describing the response of the brain using stationary MVAR models with an exogenous input (MVARX) derived from the stimulus characteristics. MVARX models are commonly used in econometric time series analysis (Lütkepohl, 2006). The advantage of the MVARX model is that it does not require subtraction of the mean and consequent reduction in signal-to-noise ratio (SNR) or the complication of time-varying models to capture the response evoked by direct electrical stimulation. The model captures both the mean evoked response and the background activity present during the recordings. We demonstrate the effectiveness of the MVARX model using intracerebral recordings from epilepsy patients.

Direct electrical stimulation of the brain presents several modeling challenges. Although the timing and location of the stimulus is known precisely, the response of the brain in the near vicinity of the stimulus cannot be measured due to electrical artifacts and the propagation of the stimulus to more distant sites depends on the topology of axons in the vicinity of the stimulation site (Ranck, 1975). Electrical stimulation creates action potentials in neurons whose axons pass near the stimulus site. These neurons synapse both near and distant to the stimulation site, so the stimulus actually activates multiple, *a priori* unknown areas. The MVARX model explicitly accounts for this effect with a bank of finite impulse response (FIR) filters that capture the impact of the exogenous input, i.e., stimulus, on all recording sites. The exogenous input filter coefficients and the MVAR model parameters are simultaneously estimated from the recordings and knowledge of the stimulation times using a least squares procedure. The exogenous input filter coefficients describe the conduction paths from the stimulus site to each recording site, while the MVAR model parameters capture the causal interactions between recording sites.

The MVARX model is applied to 10 datasets collected from three subjects in wakefulness and NREM sleep. Two stimulation levels are studied in one subject, and two stimulation sites in another. The data consists of the intracranial response to 30 current impulses separated by 1 s. A cross-validation (CV) procedure is introduced for choosing the memory in the MVARX model. We demonstrate that a stationary MVARX model accurately describes the activity evoked by direct electrical stimulation. Comparison to a series of univariate autoregressive models with exogenous inputs (ARX) reveals that causal interactions must be modeled to accurately describe the measured activity. The series of ARX models result in much larger modeling error than the MVARX model. One-step prediction performance is used to demonstrate that the MVARX model also captures spontaneous fluctuations in the recorded data. The MVARX model errors pass a whiteness test while the univariate ARX models do not, further supporting the applicability of the MVARX model.

The MVARX models are employed to contrast integrated information in wakefulness and sleep. Integrated information is a measure of the extent to which the information generated by the causal interactions in the model cannot be partitioned into independent subparts of the system. Hence, integrated information measures the balance between functional specialization and integration represented by the model. Theoretical considerations (Tononi, 2004; Laureys, 2005; Dehaene et al., 2006; Seth et al., 2008) indicate that integrated information should be less in sleep than in wakefulness. This prediction is confirmed in all 10 datasets using our MVARX model.

This paper is organized as follows. Section 2 describes the data and preprocessing procedures. Section 3 defines the MVARX model, introduces the method for estimating the model parameters, including our CV approach for selecting model memory, and presents the residual whiteness test. Section 4 demonstrates the effectiveness of the proposed model using the 10 datasets described above and section 5 applies the MVARX models to contrast integrated information in wakefulness and sleep. This paper concludes with a discussion in section 6. For notation, boldface lower and upper case symbols represent vectors and matrices, respectively, while superscript T denotes matrix transpose and superscript -1 denotes matrix inverse. The trace of a matrix \mathbf{A} is $\text{tr}[\mathbf{A}]$ and the determinant is $\det(\mathbf{A})$. $E\{a\}$ denotes the expectation of a random variable a . The Euclidean norm of a vector \mathbf{x} is $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T \mathbf{x}}$. The number of elements in a set S is $|S|$. $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means that the vector \mathbf{x} is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

2. DATA

2.1. SUBJECTS AND EXPERIMENTAL PROTOCOL

Three subjects with long-standing drug-resistant focal epilepsy participated in this study. All patients were candidates for surgical removal of the epileptic focus. During pre-surgical evaluation the patients underwent individual investigation with stereotactically implanted intracerebral multilead electrodes for precise localization of the epileptogenic areas (Cossu et al., 2005). All patients gave written informed consent before intracerebral electrode implantation as approved by the local Ethical Committee. Confirmation of the hypothesized seizure focus and localization of epileptogenic tissue in relation to essential cortex was achieved by simultaneous scalp and intracerebral electrode recording, as well as intracerebral stimulation during wakefulness and sleep to further investigate connectivity of epileptogenic and healthy tissue (Valentín et al., 2002, 2005). The decision on implantation site, duration of implantation and stimulation site(s) was made entirely on clinical needs. Stereoelectroencephalography (SEEG) activity was recorded from platinum-iridium semiflexible multilead intracerebral electrodes, with a diameter of 0.8 mm, a contact length of 2 mm, an intercontact distance of 1.5 mm and a maximal contact number of 18 (Dixi Medical, Besançon, France) (Cossu et al., 2005). The individual placement of electrodes was ascertained by post-implantation tomographic imaging (CT) scans. Scalp EEG activity was recorded from two platinum needle electrodes placed during surgery at “10–20” positions Fz and Cz on the scalp. Electroocular activity was registered at the outer canthi of both eyes, and submental electromyographic activity

was acquired with electrodes attached to the chin. EEG and SEEG signals were recorded using a 192-channel recording system (Nihon-Kohden Neurofax-110) with a sampling rate of 1000 Hz. Data was recorded and exported in EEG Nihon-Kohden format (Nobili et al., 2011, 2012). The data for each channel is obtained using bipolar referencing to a neighboring contact located entirely in the white matter. Intracerebral stimulations were started on the third day after electrode implantation. In eight out of ten cases we discuss, stimulation of strength 5 mA were performed, while for the other two cases stimulation of 1 mA were applied. At each stimulation session, the stimulation is applied at a single channel and SEEG recordings were obtained from all other channels. A single stimulation session consisted of a 30 impulse stimulation train at intervals of 1 s. Each impulse is of 0.2-ms duration. The channels that were stimulated were chosen based on clinical requirements. All patients included in this study were stimulated during wakefulness and stage 4 of NREM sleep. Sleep staging was performed using standard criteria (Rechtschaffen and Kales, 1968). Stimulations which elicited muscle twitches, sensations or cognitive symptoms, were excluded from this study, in order to prevent possible awareness of stimulation or alteration of sleep depth.

In our analysis, we consider a subset of 8–12 recording channels of all channels for each subject, as illustrated in **Figure 1**. The 8–12 channels were selected based on approximately maximizing the distance between the subset of channels that are both artifact free and near the surface of the cortex.

2.2. PREPROCESSING

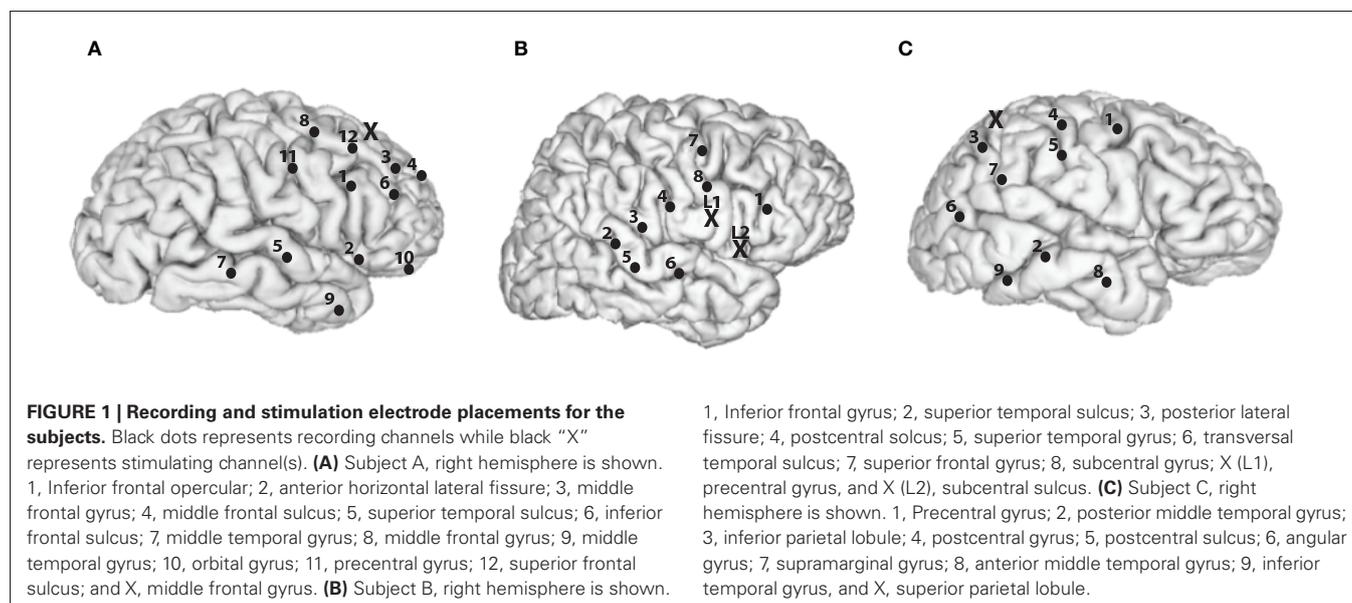
During each stimulation session, a raw trigger signal that indicates the occurrence of current stimulation with 1 and the absence of stimulation with 0 is collected at a sampling rate of 1000 Hz in addition to the SEEG recordings. We use a Tukey-windowed median filter to remove volume conduction artifacts within 39 ms of each stimulus. First, a median filter of order 19 is applied to the raw data channel by channel. Next, the

raw data within a 39-ms window centered at each stimulus is replaced with a weighted average of the raw data and the median filtered data to eliminate the artifact. The weights for the median filtered data take the form of a Tukey window (Bloomfield, 2000, p. 69) and are zero for ± 20 ms away from the stimulus, a cosine rising from 0 to 1 beginning at 19 ms prior to the stimulus and ending at 10 ms prior to the stimulus, unity until 10 ms post-stimulus, and then a cosine decreasing from 1 to 0 ending at 19 ms post-stimulus. The weighting applied to the raw data are one minus those applied to the median filtered data. **Figure 2** illustrates the results of this process. The cleaned data is then lowpass filtered by an FIR filter with passband-edge of 48 Hz and stopband-edge of 49.9 Hz to eliminate 50 Hz powerline contamination, and the lowpass filtered data is downsampled by a factor of 10 to a sampling frequency of 100 Hz. The portion of the downsampled data containing responses to stimulation are further segmented into 30 epochs of data $\mathbf{y}_n^{(j)}$, each of which contains 100 samples. Here superscript (j) denotes epoch index while subscript n denotes time index. The start of each epoch is from 12 samples (0.12 s) before the occurrence of a stimulus and the end is 87 samples (0.87 s) post-stimulus. Similarly, the raw trigger signal is lowpass filtered, downsampled by 10, and partitioned into 100-sample epochs $x_n^{(j)}$.

In principle, filtering the signal may have an impact on model estimation and causality inference (Barnett and Seth, 2011). We minimize the potential impact of filtering by specifying the stopband edge of the lowpass filter close to the post-downsampling Nyquist frequency.

2.3. IDENTIFICATION OF OUTLYING EPOCHS

An automated procedure is employed to exclude epochs that markedly deviate from the majority of epochs due to non-stationary brain activity or other factors. Let $\mathbf{y}_n^{(j)} = [y_{1,n}^{(j)}, y_{2,n}^{(j)}, \dots, y_{d,n}^{(j)}]^T$ represent the d channels of recordings at



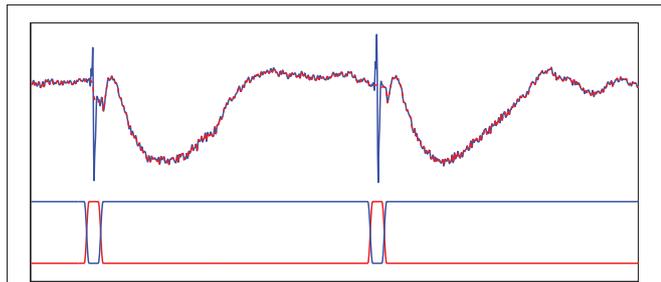


FIGURE 2 | Tukey-windowed median filtering for eliminating volume conduction artifacts. The upper trace depicts an example of raw data (blue solid line) and the Tukey-windowed median filter output (red dashed line). The lower trace depicts the weighting applied to the raw data (blue solid line) and the median filtered data (red solid line) to eliminate the volume conduction artifact.

time $n = 1, 2, \dots, N_j$ from epochs $j = 1, 2, \dots, J$. For epoch m , we compute the time-varying mean $\mu_{\mathbf{y}}^{-m}(n)$ and time-varying covariance matrix $\Sigma_{\mathbf{y}}^{-m}(n)$ by excluding the m -th epoch of data. That is,

$$\mu_{\mathbf{y}}^{-m}(n) = \frac{1}{J-1} \sum_{j=1, j \neq m}^J \mathbf{y}_n^{(j)} \quad (1)$$

$$\Sigma_{\mathbf{y}}^{-m}(n) = \frac{1}{J-2} \times \sum_{j=1, j \neq m}^J \left(\mathbf{y}_n^{(j)} - \mu_{\mathbf{y}}^{-m}(n) \right) \left(\mathbf{y}_n^{(j)} - \mu_{\mathbf{y}}^{-m}(n) \right)^T, \quad (2)$$

for $n = 1, \dots, 100$. Here $m = 1$ to J and J is 30 for all data sets considered. Then the squared Mahalanobis distance (Penny, 1996) between the epoch m and the other epochs is computed as

$$D^2(m) = \sum_{n=1}^{100} \left(\mathbf{y}_n^{(m)} - \mu_{\mathbf{y}}^{-m}(n) \right)^T \times \left(\Sigma_{\mathbf{y}}^{-m}(n) \right)^{-1} \left(\mathbf{y}_n^{(m)} - \mu_{\mathbf{y}}^{-m}(n) \right). \quad (3)$$

Epochs with $D^2(m)$ exceeding

$$100 \cdot d + 60\sqrt{2 \cdot 100 \cdot d} \quad (4)$$

are declared as outliers and removed from subsequent analysis. Intuitively, if the data is Gaussian, then $D^2(m)$ is Chi-squared distributed with $100 \cdot d$ degrees of freedom. This implies that the threshold rules out an epoch m if $D^2(m)$ exceeds its mean plus 60 standard deviations. Thus this threshold only excludes epochs that have a large deviation from the temporal average of the other epochs. The number of epochs retained for analysis are given in **Table 1**.

Table 1 | Number of non-outlying epochs used in analysis.

Dataset	Wakefulness epochs	Sleep epochs
Subject A, 1 mA	29	25
Subject A, 5 mA	28	22
Subject B, L1	30	24
Subject B, L2	30	29
Subject C	30	29

3. METHODS

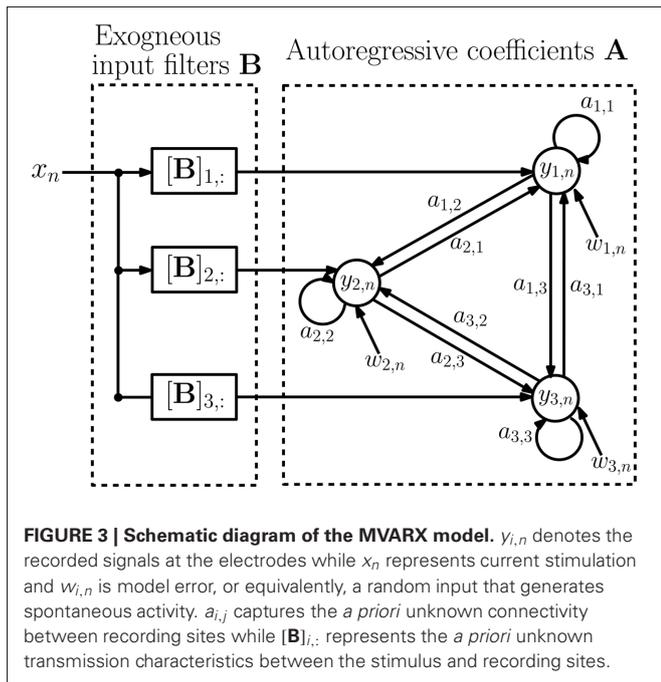
3.1. MVARX MODEL

The MVARX model of order (p, ℓ) describes the data as follows (Lütkepohl, 2006):

$$\mathbf{y}_n^{(j)} = \sum_{i=1}^p \mathbf{A}_i \mathbf{y}_{n-i}^{(j)} + \sum_{i=0}^{\ell} \mathbf{b}_i x_{n-i}^{(j)} + \mathbf{w}_n^{(j)}, \quad (5)$$

where $x_n^{(j)}$ denotes the input at time n and epoch j . The $d \times d$ matrices $\mathbf{A}_i = \{a_{m,n}(i)\}$ contain autoregressive coefficients describing the influence of channel n on channel m at lag i , and the $d \times 1$ vectors $\mathbf{b}_i = \{b_m(i)\}$ contain filter coefficients from the stimulus to channel m at lag i . The vectors $\mathbf{w}_n^{(j)}$ are $d \times 1$ zero-mean noise vectors with covariance matrix \mathbf{Q} and are assumed to satisfy $E\{\mathbf{w}_n^{(j)} (\mathbf{w}_s^{(j)})^T\} = 0$, for either $i \neq j$ or $n \neq s$. We assume that the epochs are of varying lengths N_j and are possibly disconnected in time to accommodate rejection of outlying epochs. **Figure 3** depicts a schematic diagram of an example MVARX model. The diagram assumes there are three recording electrodes corresponding to the recordings $y_{1,n}$, $y_{2,n}$, and $y_{3,n}$ (the epoch index j is omitted in the figure for simplicity). The intracranial EEG signals recorded at the electrodes contain contributions due to the current stimulus response and background brain activity. The exogenous input x_n represents the current stimulation. If $\mathbf{B} = [\mathbf{b}_0, \dots, \mathbf{b}_\ell]$ is a $d \times (\ell + 1)$ matrix of exogenous input coefficients, then the i -th row of \mathbf{B} , $[\mathbf{B}]_{i,:}$, is the impulse response of the filter representing the unknown transmission characteristics between the current stimulus and the i -th recording channel. The autoregressive coefficients $\mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_p]$ indicate how past values of the recorded signals affect present values. The autoregressive order p determines the time extent of the past that affect the present values and may be regarded as the memory of the system. The signals $w_{1,n}$, $w_{2,n}$, and $w_{3,n}$ can be interpreted as modeling errors or alternatively as a process that generates spontaneous activity.

Electrodes can be used either as stimulating or recording electrodes but cannot be used simultaneously for recording and stimulation. Moreover, the electrodes closest to the stimulation site are affected by huge electrical artifacts and they cannot be used because of consequent low SNR. Hence the recorded data $\mathbf{y}_n^{(j)}$ contains recordings of the effect of the stimulation at distant sites, not the stimulation itself. Stimulation depolarizes the membranes of neurons passing through the neighborhood of the stimulating electrode, possibly creating action potentials



in neurons that synapse near the stimulation site and at distant locations (Ranck, 1975), a phenomenon termed fibers of passage. Thus, stimulation generates an “input” that is conveyed to potentially all recording sites in a manner that depends on the axonal topology in the vicinity of the stimulation site. This topology and consequent stimulation effects are usually unknown and described in our MVARX model by the exogenous input filters **B**. In our model we assume the exogenous input is given by the trigger signal associated with delivery of a current pulse, so **B** captures both the shape of the delivered stimulus and the unknown direct propagation of the input to each recording site.

Denote $\mathbf{y}_{n,s}^{(j)}$ and $\mathbf{y}_{n,e}^{(j)}$ as the spontaneous activity and stimulus response to the exogenous input, respectively, at time n from epoch j . Equation (5) can be alternatively expressed as

$$\mathbf{y}_n^{(j)} = \mathbf{y}_{n,s}^{(j)} + \mathbf{y}_{n,e}^{(j)} \tag{6}$$

$$\mathbf{y}_{n,s}^{(j)} = \sum_{i=1}^p \mathbf{A}_i \mathbf{y}_{n-i,s}^{(j)} + \mathbf{w}_n^{(j)} \tag{7}$$

$$\mathbf{y}_{n,e}^{(j)} = \sum_{i=1}^p \mathbf{A}_i \mathbf{y}_{n-i,e}^{(j)} + \sum_{i=0}^{\ell} \mathbf{b}_i x_{n-i}^{(j)} \tag{8}$$

Note that in practice $\mathbf{y}_{n,s}^{(j)}$ and $\mathbf{y}_{n,e}^{(j)}$ are not directly observed and cannot be separated from $\mathbf{y}_n^{(j)}$ without knowledge of the MVARX model parameters. The stimulus response component $\mathbf{y}_{n,e}^{(j)}$ is a deterministic term that depends entirely on the stimulus and the model. Given the model parameters $\Theta = [\mathbf{A}, \mathbf{B}]$, we can generate $\mathbf{y}_{n,e}^{(j)}$ by applying the stimulus sequence $\mathbf{x}_n^{(j)}$ to Equation (8) with zero initial conditions. Recall that $\mathbf{w}_n^{(j)}$ is assumed to be zero mean,

so $\mathbf{y}_{n,s}^{(j)}$ is a zero mean random process reflecting the spontaneous component of the recordings. It is common in MVAR modeling to subtract the mean prior to estimating MVAR model parameters (Ding et al., 2000). This corresponds to removing the stimulus response $\mathbf{y}_{n,e}^{(j)}$ and is unnecessary with the MVARX model. We shall assume that the stimulus is repeated multiple times such that averaging $\mathbf{y}_{n,e}^{(j)}$ with respect to the stimulus onset times produces the evoked response of the system. This is not required by the model in Equation (5) but is consistent with conventional electrophysiology practice.

The autoregressive parameters **A** model the inherent neural connectivity between sites—how activity at one site propagates to another site. This is evident in Equations (5–8) by the fact that the \mathbf{A}_i are applied to $\mathbf{y}_{n-i}^{(j)}$. If the spontaneous activity $\mathbf{y}_{n,s}^{(j)}$ is very weak relative to $\mathbf{y}_{n,e}^{(j)}$ then the response is described entirely by Equation (8) and the measured data $\mathbf{y}_n^{(j)} \approx \mathbf{y}_{n,e}^{(j)}$. In this case there is a potential modeling ambiguity as there are many different combinations of \mathbf{A}_i and \mathbf{b}_i that could be used to describe $\mathbf{y}_{n,e}^{(j)}$ over a finite duration. For example, $\mathbf{y}_{n,e}^{(j)}$ can be described on $1 \leq n \leq \ell + 1$ by setting $\mathbf{A}_i = 0$ and only using \mathbf{b}_i . We control potential ambiguities associated with relatively weak spontaneous activity by limiting ℓ to a value commensurate with the expected duration of stimulus propagation through fibers of passage. This ensures that **B** is not able to capture long duration interactions associated with feed forward and feedback connectivity between sites. Based on previous experimental evidence (Matsumoto et al., 2004), we set $\ell = 10$ to accommodate a 100 ms duration of propagation through fibers of passage. We will discuss this choice more thoroughly in section 6.

3.2. ESTIMATION OF MVARX MODEL PARAMETERS

Suppose that we have the recordings and inputs $\{(\mathbf{y}_n^{(j)}, \mathbf{x}_n^{(j)}) : j = 1, 2, \dots, J, n = 1, 2, \dots, N_j\}$ for J epochs of N_j samples each. Denote $n_0 = \max(p, \ell)$, and suppose that $N_j \geq n_0 + 1$, for all j . Using the first n_0 samples as the initial values, the model in Equation (5) can be rewritten in a simplified form:

$$\mathbf{y}_n^{(j)} = \Theta \mathbf{z}_{n-1}^{(j)} + \mathbf{w}_n^{(j)}, \tag{9}$$

for $j = 1, \dots, J, n = n_0 + 1, \dots, N_j$, where the $d \times (dp + \ell + 1)$ matrix $\Theta = [\mathbf{A}, \mathbf{B}]$ and the vector of dimension $dp + \ell + 1$, $\mathbf{z}_{n-1}^{(j)} = [(\mathbf{y}_{n-1}^{(j)})^T, (\mathbf{y}_{n-2}^{(j)})^T, \dots, (\mathbf{y}_{n-p}^{(j)})^T, x_n^{(j)}, x_{n-1}^{(j)}, \dots, x_{n-\ell}^{(j)}]^T$. The vectors $\mathbf{y}_n^{(j)}$, $\mathbf{w}_n^{(j)}$, and $\mathbf{z}_{n-1}^{(j)}$ can be further concatenated as columns of the matrices \mathbf{Y}_j , \mathbf{Z}_j , and \mathbf{W}_j to write:

$$\mathbf{Y}_j = \Theta \mathbf{Z}_j + \mathbf{W}_j \tag{10}$$

where $\mathbf{Y}_j = [\mathbf{y}_{n_0+1}^{(j)}, \dots, \mathbf{y}_{N_j}^{(j)}]$, $\mathbf{Z}_j = [\mathbf{z}_{n_0}^{(j)}, \dots, \mathbf{z}_{N_j-1}^{(j)}]$, and $\mathbf{W}_j = [\mathbf{w}_{n_0+1}^{(j)}, \dots, \mathbf{w}_{N_j}^{(j)}]$. This expression takes the form of a linear regression model, and we can obtain an ordinary least square (OLS) estimate of (Θ, \mathbf{Q}) as (Lütkepohl, 2006, chap. 10.3):

$$\hat{\Theta} = \left(\sum_{j=1}^J \mathbf{Y}_j \mathbf{Z}_j^T \right) \left(\sum_{j=1}^J \mathbf{Z}_j \mathbf{Z}_j^T \right)^{-1},$$

$$\hat{\mathbf{Q}} = \frac{1}{N_t} \sum_{j=1}^J (\mathbf{Y}_j - \hat{\Theta} \mathbf{Z}_j) (\mathbf{Y}_j - \hat{\Theta} \mathbf{Z}_j)^T, \quad (11)$$

where $N_t = \sum_{j=1}^J N_j - n_0 J$. If $\mathbf{w}_n^{(j)}$ is Gaussian, then the OLS estimate $(\hat{\Theta}, \hat{\mathbf{Q}})$ is also the maximum-likelihood estimate of (Θ, \mathbf{Q}) (Lütkepohl, 2006).

3.3. MODEL SELECTION WITH CROSS-VALIDATION

In practice the order p could be chosen using numerous different model selection criteria, including Akaike information criterion and the Bayesian information criterion (McQuarrie and Tsai, 1998; Lütkepohl, 2006). Here we use CV to determine p in a data-driven fashion [see Cheung et al. (2012) for another example of using CV to select model parameters with neurophysiological data]. The data $\mathbf{y}_n^{(j)}$ and input $x_n^{(j)}$ are partitioned into training and test sets. The goal is to choose the value p that produces the best prediction of test data when the model $\Theta = [\mathbf{A}, \mathbf{B}]$ is estimated from the training data. We consider two components in assessing model predictive capability. The first is the one-step prediction error, a measure of the model’s ability to track the sample-to-sample and epoch-to-epoch fluctuations in the data. The second is the error between the average evoked response predicted by the model and the measured average response. This measures the quality of the model’s response to the stimulus.

Partition the epochs of available data into training sets R_m and test sets S_m and assume there are $m = 1, 2, \dots, M$ such partitions. Assume the sets S_m are non-overlapping and are of approximately the same size. Let Θ_m be the model estimated from R_m as described in the preceding subsection. The one-step prediction error at time n , $\mathbf{e}_n^{(j)}(\Theta_m)$ is the difference between the recording $\mathbf{y}_n^{(j)}$ and the one-step prediction made by Θ_m using the n_0 samples prior to time n , that is, $\mathbf{z}_{n-1}^{(j)}$:

$$\mathbf{e}_n^{(j)}(\Theta_m) = \mathbf{y}_n^{(j)} - \hat{\mathbf{y}}_n^{(j)}(\Theta_m) \quad (12)$$

where the one-step prediction $\hat{\mathbf{y}}_n^{(j)}(\Theta_m) = \Theta_m \mathbf{z}_{n-1}^{(j)}$. Similarly we define the average response error as

$$\epsilon_n(\Theta_m) = \bar{\mathbf{y}}_n(S_m) - \hat{\bar{\mathbf{y}}}_n(\Theta_m, S_m) \quad (13)$$

where the average evoked response $\bar{\mathbf{y}}_n(S_m) = 1/|S_m| \cdot \sum_{j \in S_m} \mathbf{y}_n^{(j)}$ and the average model response $\hat{\bar{\mathbf{y}}}_n(\Theta_m, S_m)$ over epochs in S_m , $\hat{\bar{\mathbf{y}}}_n(\Theta_m, S_m) = 1/|S_m| \cdot \sum_{j \in S_m} \hat{\mathbf{y}}_n^{(j)}(\Theta_m)$. Here $\hat{\mathbf{y}}_n^{(j)}(\Theta_m)$ is generated using Θ_m as described following Equation (8). We define a CV score as a weighted combination of the one-step prediction and average response errors averaged over all training/test data partitions

$$CV(p) = \frac{1}{M} \sum_{m=1}^M \left[\frac{CV_e(p, m)}{w_e} + \frac{CV_\epsilon(p, m)}{w_\epsilon} \right] \quad (14)$$

where $CV_e(p, m)$ is the mean square one-step prediction error of a p -th order model $\Theta_m(p)$ in predicting data in S_m :

$$CV_e(p, m) = \frac{1}{|S_m|} \sum_{j \in S_m} \frac{1}{N_j - n_0} \sum_{n=n_0+1}^{N_j} \|\mathbf{e}_n^{(j)}(\Theta_m(p))\|_2^2 \quad (15)$$

and $CV_\epsilon(p, m)$ is the mean square value of the average response error on S_m :

$$CV_\epsilon(p, m) = \frac{1}{N} \sum_{n=1}^N \|\epsilon_n(\Theta_m(p))\|_2^2. \quad (16)$$

Here N is the assumed duration of the average response. The weights w_e and w_ϵ vary the emphasis between the one-step prediction error and average response error. In the analysis below, we set w_e and w_ϵ to the medians of $CV_e(p, m)$ and $CV_\epsilon(p, m)$, respectively, for $m = 1, \dots, M$ and all p considered. This approach places approximately equal emphasis on the two errors. The model order p is chosen as the p that minimizes $CV(p)$ over the range of p evaluated.

Several practical issues require attention for computing the average response error. First, use of an average evoked response assumes the stimulus is nominally identical for each epoch. Second, care must be taken in computing the average response of the model Θ to the stimulus $x_n^{(j)}$ over epochs in S_m if the effects of preceding stimuli extend into S_m . In such a case the brain is not “at rest” upon the arrival of the new stimulus in S_m , but is still responding to the preceding stimulus. This situation occurs when the response time of the cortex is longer than the inter-stimulus interval. We mimic this aspect of the measured data when computing the average model response by presenting the entire train of stimuli to the model and averaging over the responses corresponding to epochs in S_m .

3.4. MODEL QUALITY ASSESSMENT

A key assumption for the consistency of the OLS estimates is that the residuals $\mathbf{w}_n^{(j)}$ be serially uncorrelated, that is, temporally white. Serial correlation in $\mathbf{w}_n^{(j)}$ may be a sign of mis-specifying the model or incorrect selection of order (p, ℓ) (Hong, 1996; Duchesne and Roy, 2004). We use a consistency test developed in Duchesne and Roy (2004) to validate our models. Denote by $\Gamma_{\mathbf{w}}(r) = E\{\mathbf{w}_n^{(j)} (\mathbf{w}_{n-r}^{(j)})^T\}$ the covariance at lag r , the hypotheses of interest are:

$$H_0 : \Gamma_{\mathbf{w}}(r) = \mathbf{0}, \text{ for all } \mathbf{r} \neq \mathbf{0} \quad \text{vs.}$$

$$H_1 : \Gamma_{\mathbf{w}}(r) \neq \mathbf{0}, \text{ for some } \mathbf{r} \neq \mathbf{0}. \quad (17)$$

Let the residual at time n in epoch j be $\hat{\mathbf{w}}_n^{(j)} = \mathbf{y}_n^{(j)} - \hat{\Theta} \mathbf{z}_{n-1}^{(j)}$. Let $q(\cdot)$ be a window function of bounded support L , that is, $q(r) > 0$, for $|r| \leq L$ and $q(r) = 0$ for $|r| > L$. Suppose that the last epoch is of length longer than $(J - 1)L$, that is, $N_j > (J - 1)L$. The test

statistic derived in Duchesne and Roy (2004) for testing H_0 vs. H_1 is

$$T_{N_c} = \frac{N_c \sum_{r=1}^L q^2(r) \text{tr}[\mathbf{C}_{\hat{\mathbf{w}}}(r) \mathbf{C}_{\hat{\mathbf{w}}}^{-1}(0) \mathbf{C}_{\hat{\mathbf{w}}}(r) \mathbf{C}_{\hat{\mathbf{w}}}^{-1}(0)] - d^2 M_{N_c}(q)}{[2d^2 V_{N_c}(q)]^{1/2}} \quad (18)$$

where $N_c = \sum_{j=1}^J N_j - (J - 1)L$ and

$$\mathbf{C}_{\hat{\mathbf{w}}}(r) = \frac{1}{N_c} \left[\sum_{j=1}^{J-1} \sum_{n=r+1}^{N_j} \hat{\mathbf{w}}_n^{(j)} (\hat{\mathbf{w}}_{n+r}^{(j)})^T + \sum_{n=r+1+(J-1)(L-r)}^{N_j} \hat{\mathbf{w}}_n^{(j)} (\hat{\mathbf{w}}_{n+r}^{(j)})^T \right], \quad (19)$$

for $r = 0, 1, \dots, L$, are the estimated residual covariance matrices. The functionals $M_{N_c}(q)$ and $V_{N_c}(q)$ of $q(\cdot)$ and N_c are defined as (Duchesne and Roy, 2004):

$$M_{N_c}(q) = \sum_{i=1}^{L-1} \left(1 - \frac{i}{N_c}\right) q^2(i) \quad (20)$$

$$V_{N_c}(q) = \sum_{i=1}^{L-2} \left(1 - \frac{i}{N_c}\right) \left(1 - \frac{(i+1)}{N_c}\right) q^4(i). \quad (21)$$

We use the Bartlett window defined as $q(j) = 1 - |j|/L, j \leq L$ and $q(j) = 0, j > L$ with a window width $L = \lceil 3N_c^{0.3} \rceil$ as suggested in Duchesne and Roy (2004). For example, in our datasets the longest possible single epoch would have $N_c = 3000$ samples, which leads to the maximum value $L = 34$. Thus the test statistic Equation (18) is based on estimated residual covariance matrices at lags less than or equal to 34. Under the assumption that both $\mathbf{y}_n^{(j)}$ and $\mathbf{x}_n^{(j)}$ are stationary, the test statistic is one-sided and asymptotically standard normally distributed (see Duchesne and Roy, 2004, Theorem 1). It declares that the residuals are serially correlated if $T_N > z_{1-\alpha}$ and are white otherwise, where $z_{1-\alpha}$ is the value of the inverse cumulative distribution function of the standard normal distribution at $1 - \alpha$ and α is the significance level of the test.

4. RESULTS

4.1. MODEL PARAMETERS

We have varying definitions and lengths of epochs throughout our data processing procedures. For detection of outlying epochs we

choose all epochs to be of length $N_j = 100$ samples based on the time between subsequent current stimuli. In model estimation and assessment of residual whiteness, the epochs are defined as the maximum contiguous segments between the time segments removed by the outlier detection process. This minimizes the impact of the initial conditions $\mathbf{z}_{n_0}^{(j)}$ required at the start of each epoch. Hence, N_j varies across epochs and conditions. In CV, the epoch lengths are set to be equal with $N_j = 100$. This, along with choosing the test sets S_m to contain approximately the same number of epochs, makes the test sets span roughly the same amount of time.

As shown in **Table 1**, the number of outlying epochs is generally larger in sleep than in wakefulness, most likely due to the presence of slow waves during sleep. The number of partitions of the available epochs used in the CV procedure for determining model order p and the corresponding model order is shown in **Table 2**. We did not consider model orders higher than $p = 30$. We also evaluated an unconnected model consisting of d univariate ARX models to assess the importance of the coupling or connectivity between channels. The univariate models were estimated by applying the procedure described above to each channel. With the exception of Subject B, stimulus location 1 (L1), the CV procedure picks a higher model order for the unconnected model and in many cases chooses the maximum order considered.

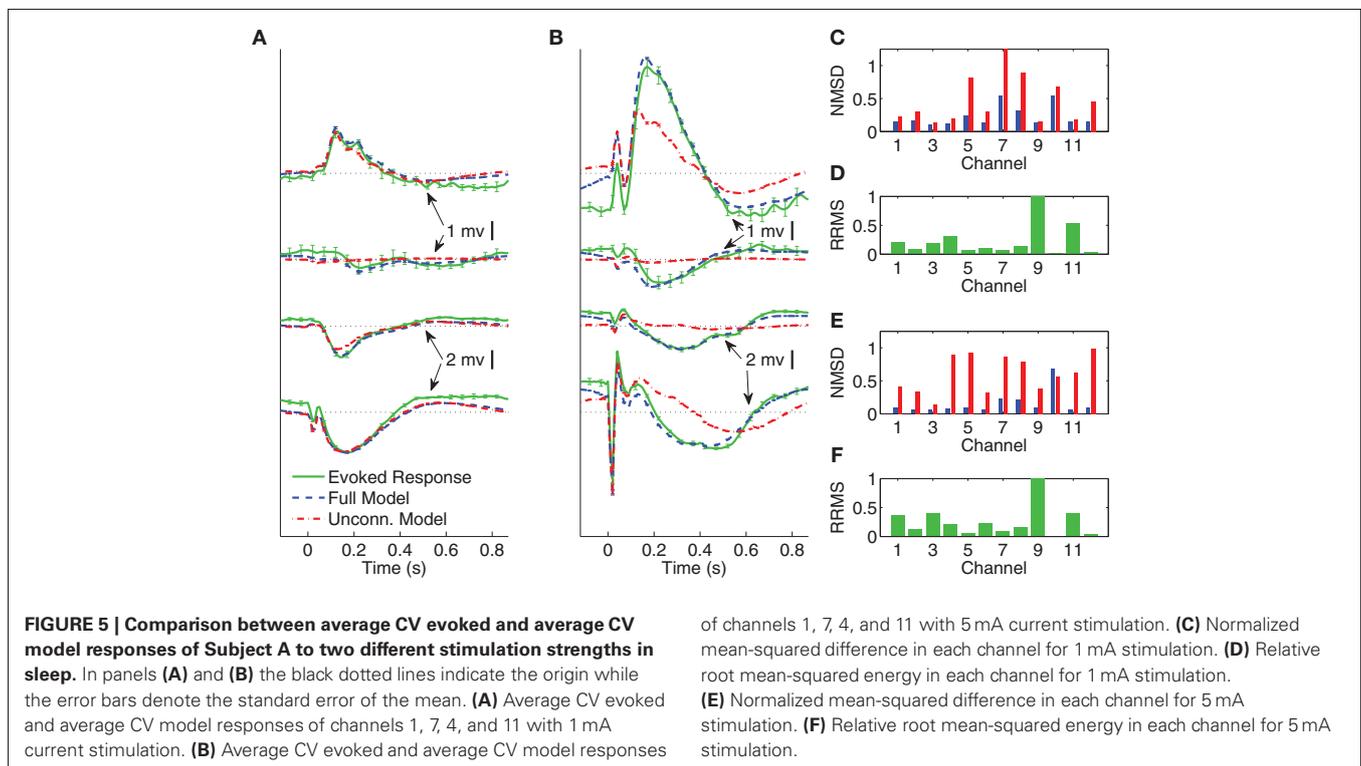
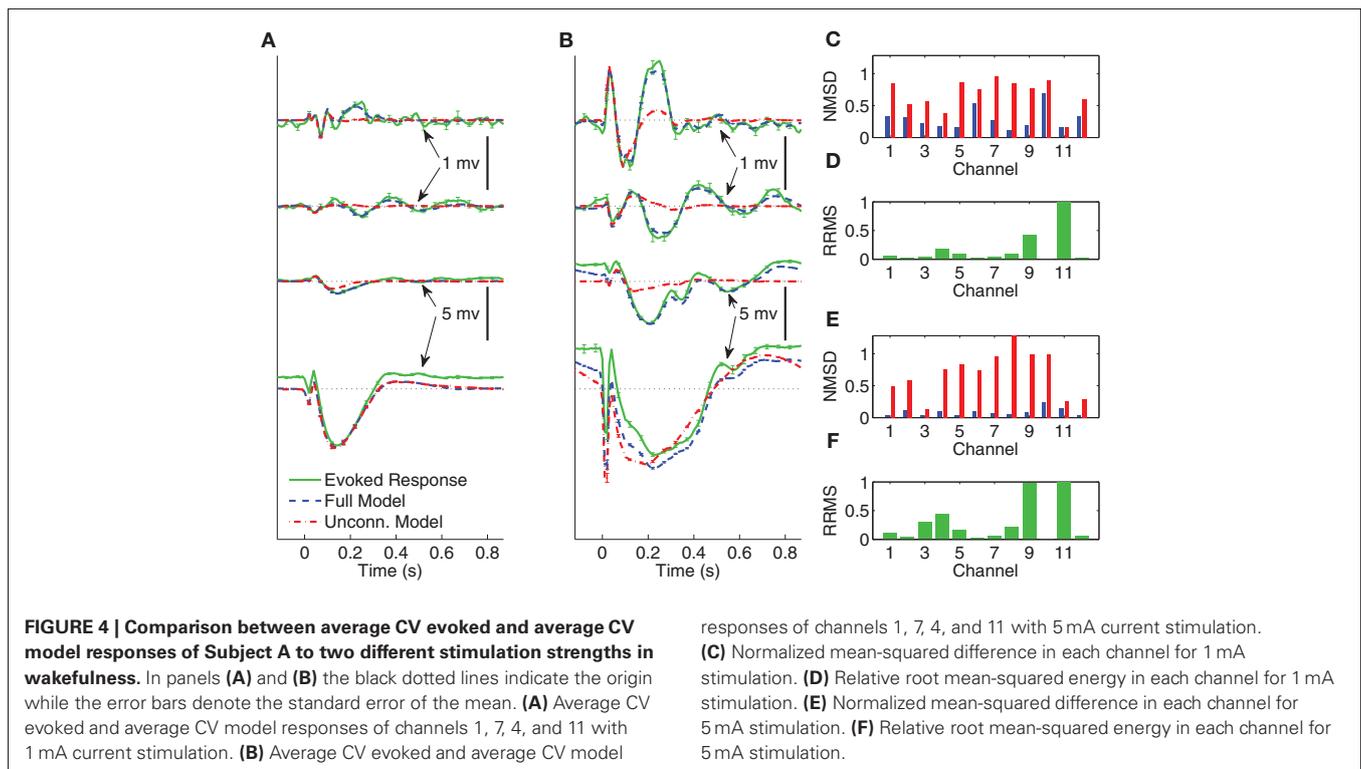
The whiteness test described in section 3.4 was applied to the residuals from all models using a significance level $\alpha = 0.1$. Note that since exceeding the threshold implies the residuals are not white, use of a relatively large value for α leads to a more stringent test, that is, makes it easier to declare the residuals are not white. The MVARX models passed the whiteness test for every data set, while the unconnected models failed the test for every data set.

4.2. EVOKED RESPONSE MODEL PERFORMANCE

In **Figures 4–6** we compare the average evoked response and average model response for a subset of subjects and conditions. The average responses are generated following the CV approach described in section 3.3. **Figures 4A,B** show the average CV evoked responses $\bar{\mathbf{y}}_n(S) = M^{-1} \sum_{m=1}^M \bar{\mathbf{y}}_n(S_m)$ and average CV model responses $\hat{\mathbf{y}}_n(\Theta, S) = M^{-1} \sum_{m=1}^M \hat{\mathbf{y}}_n(\Theta_m, S_m)$ in channels 1, 4, 7, and 11 of Subject A in wakefulness for 1 and 5 mA stimulation, respectively. Here 0s on the time axis corresponds to the stimulus onset. The averaging is first done within the testing block for each CV partition, then a second phase of averaging is done over the average responses of the test blocks for all CV

Table 2 | Model order parameters for wakefulness and sleep data sets.

Dataset	Wakefulness			Sleep		
	CV Part.	MVARX p	ARX p	CV Part.	MVARX p	ARX p
Subject A, 1 mA	7	20	30	8	20	30
Subject A, 5 mA	7	26	30	11	26	26
Subject B, L1	10	30	28	8	30	24
Subject B, L2	10	18	22	7	22	30
Subject C	10	16	30	7	12	30



partitions. The average CV model response of the MVARX model (blue dashed line) follows the dynamics of the average CV evoked response (green solid line) in each channel, for both stimulus amplitudes and a range of channel response levels. In contrast,

the average CV model response of the unconnected model (red dashed dot line) only tracks the average CV evoked response in channels with the largest amplitudes, even though the univariate model is fit independently to each channel. In the figures,

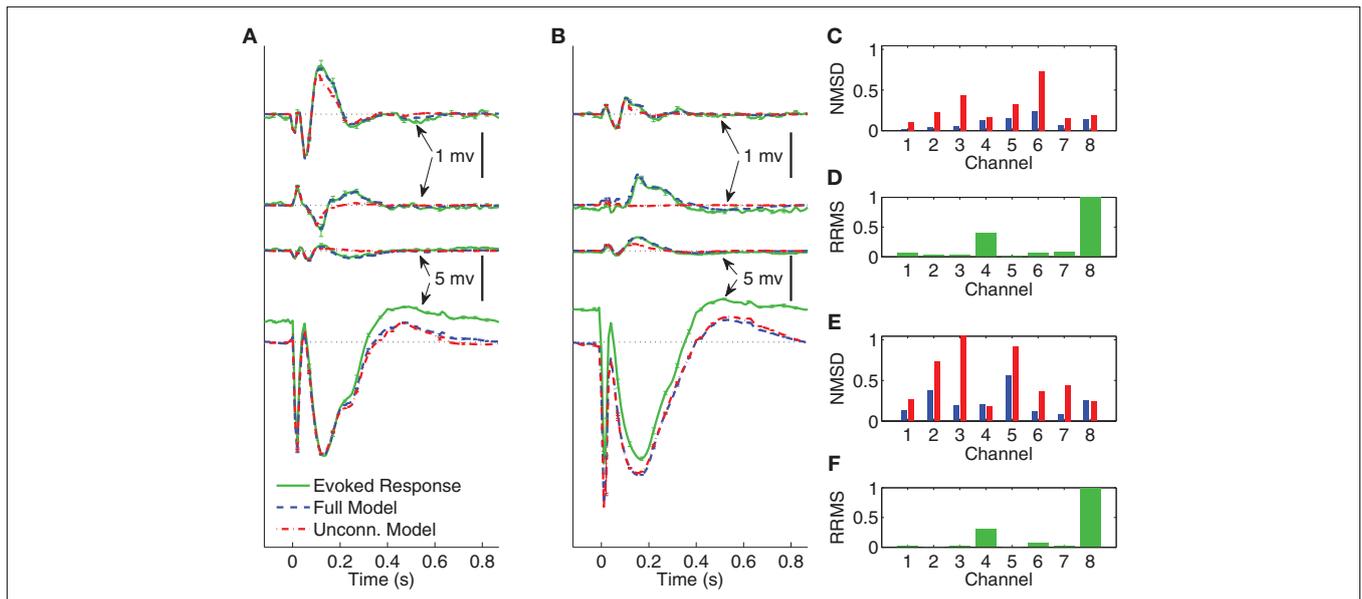


FIGURE 6 | Comparison between average CV evoked responses and average CV model responses of Subject B with two different stimulating locations in wakefulness. In panels (A) and (B) the black dotted lines indicate the origin while the error bars denote the standard error of the mean. (A) Average CV evoked and average CV model responses of channels 1, 3, 6, and 8 when the stimulating channel is L1. (B) Average CV evoked and average

CV model responses of channels 1, 3, 6, and 8 when the stimulating channel is L2. (C) Normalized mean-squared difference in each channel when the stimulating channel is L1. (D) Relative root mean-squared energy in each channel when the stimulating channel is L2. (E) Normalized mean-squared difference in each channel when the stimulating channel is L1. (F) Relative root mean-squared energy in each channel with the stimulating channel is L2.

error bars indicating one standard error are displayed every five samples. **Figures 4C–F** summarize the model performance on a channel-by-channel basis. Let $\bar{y}_{i,n}(S)$ and $\hat{y}_{i,n}(\Theta, S)$ be the average CV evoked response and average CV model response at time n in the i -th channel. **Figures 4C,E** depict the normalized mean-squared difference (NMSD) between the average CV evoked and average CV model response for 1 and 5 mA stimulation, respectively, where the NMSD in channel i is defined as

$$\text{NMSD}(i) = \frac{\sum_{n=1}^N (\bar{y}_{i,n}(S) - \hat{y}_{i,n}(\Theta, S))^2}{\sum_{n=1}^N \bar{y}_{i,n}^2(S)}. \quad (22)$$

Figures 4D,F depict the relative root mean-squared (RRMS) energy for 1 and 5 mA stimulations, respectively, for each channel. The RRMS for channel i is defined as the ratio of the root mean-squared energy in channel i to that of the channel with the largest root mean-squared energy. More precisely,

$$\text{RRMS}(i) = \frac{\sqrt{\sum_{n=1}^N \bar{y}_{i,n}^2(S)}}{\max_{i'=1, \dots, d} \sqrt{\sum_{n=1}^N \bar{y}_{i',n}^2(S)}}. \quad (23)$$

The unconnected model only gives comparable NMSD to that of full model in channel 11, which has the largest energy. The difference between the MVARX model and the unconnected model in terms of per-channel NMSD is less significant for the 1 mA stimulation, than for the 5 mA stimulation.

Figures 5A,B depict the average CV evoked and average CV model responses for Subject A during NREM sleep with current

stimulation of 1 and 5 mA, respectively. The four traces, from top to bottom, show the responses in channels 1, 7, 4, and 11, respectively. Panels (C) and (E) depict the NMSD, while (D) and (F) depict RRMS for 1 and 5 mA stimulation, respectively, as a function of channel.

The average CV evoked responses and the average CV model responses in wakefulness for Subject B, with two different stimulating sites L1 and L2, and both with current stimulus of 5 mA, are shown in panels (A) and (B) of **Figure 6**. The four traces, from top to bottom, depict the responses in channels 1, 3, 6, and 8, respectively. The difference between the two stimulating sites lies mainly in channels with smaller energy, i.e., channels 1, 3, and 6. Panels (C) and (E) depict NMSD in each channel when the stimulating channel is L1 and L2, respectively. Panels (D) and (F) show the RRMS in each channel.

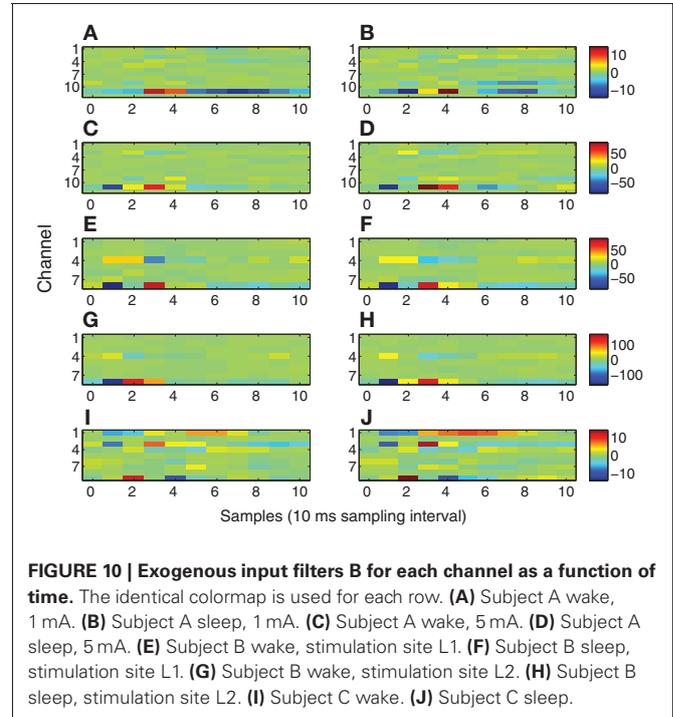
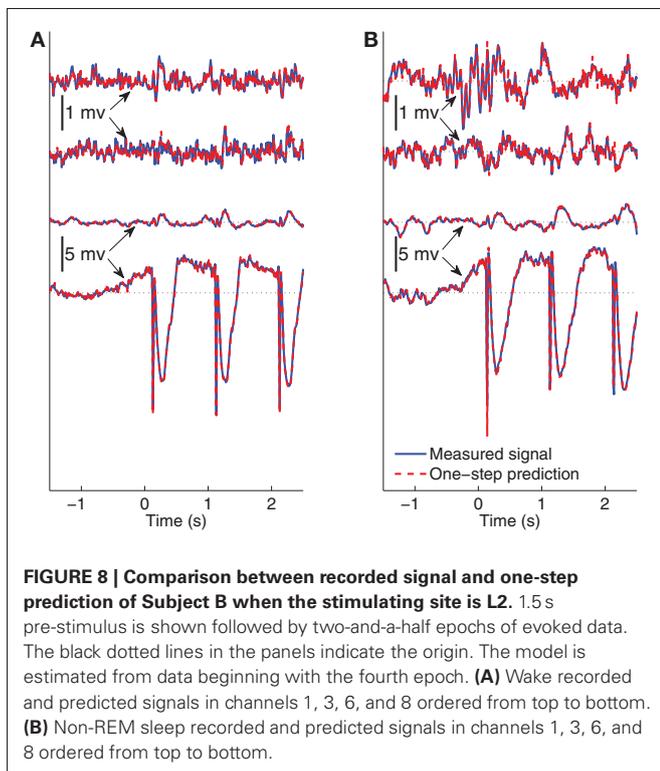
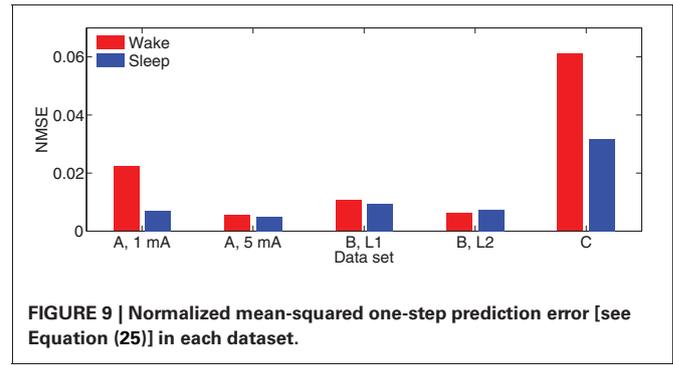
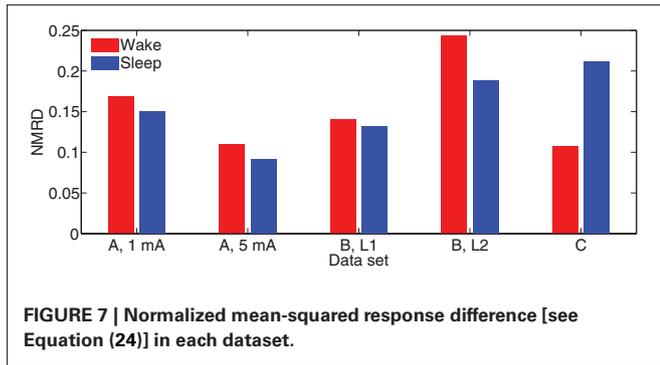
Define the normalized mean-squared response difference (NMRD) over all channels as the ratio of the NMRD to the mean-squared average CV evoked response. That is,

$$\text{NMRD} = \frac{\sum_{n=1}^N \|\bar{y}_n(S) - \hat{y}_n(\Theta, S)\|_2^2}{\sum_{n=1}^N \|\bar{y}_n(S)\|_2^2}. \quad (24)$$

Figure 7 depicts NMRD of the MVARX models for all five data sets considered. Generally the MVARX models captures the dynamics in average evoked response reasonably well with NMRD no larger than 0.25.

4.3. ONE-STEP PREDICTION MODEL PERFORMANCE

The ability of the model to predict the present recorded value of the data given past recordings reflects a different attribute than



the modeling of the average evoked response. One-step prediction performance indicates the model's ability to follow spontaneous fluctuations in the data. **Figure 8** compares the recording $y_n^{(j)}$ and one-step prediction $\hat{y}_n^{(j)}(\Theta)$ of the signals recorded from Subject B for 1.5 s of pre-stimulus data followed by two and a half epochs of evoked data, when the stimulating site is L2. The models used to perform prediction in **Figure 8** are trained from data excluding the data plotted. Panels (A) and (B) shows the signals in wakefulness and sleep, respectively. Similar results are obtained for the other epochs, subjects, and conditions. The traces show the signals in channels 1, 3, 6, and 8, respectively. These results indicate that the MVARX model performs accurate one-step prediction in wakefulness and sleep and for both pre-stimulus and evoked data segments.

Define the normalized mean-squared one-step (NMSE) prediction error as the ratio of the mean-squared prediction error

over the samples to the mean-squared energy. That is,

$$\text{NMSE} = \frac{\frac{1}{J(N-n_0)} \sum_{j=1}^J \sum_{n=n_0+1}^N \|y_n^{(j)} - \hat{y}_n^{(j)}(\Theta)\|_2^2}{\frac{1}{JN} \sum_{j=1}^J \sum_{n=1}^N \|y_n^{(j)}\|_2^2}. \quad (25)$$

As a reference, the NMSE of the model $\Theta = \mathbf{0}$ is approximately 1. The bar diagrams in **Figure 9** show the NMSE of the MVARX models for all five datasets considered. Overall, our models give NMSE less than 0.06 for one-step prediction of the recordings and less than 0.02 in seven of the ten data sets studied.

4.4. B MATRICES

Figure 10 depicts the exogenous input filters B matrices estimated for all 10 datasets as color plots. The i -th row of each matrix represents the FIR filter coefficients representing the path from the stimulus site to the i -th channel. Hence, rows with greater extremes of color have the strongest paths from the stimulus site.

5. APPLICATION TO CONSCIOUSNESS ASSESSMENT

Numerous network characteristics can be obtained from an MVARX model. For example, graphs with partially directed coherence or conditional Granger causality as edges can be obtained by computing partially directed coherence or conditional Granger causality from the MVARX parameters. In this section we demonstrate the application of the model to assessment of consciousness by measuring the integrated information of the estimated MVARX model. The integrated information theory (Tononi, 2004, 2008, 2010) starts from two self-evident axioms about consciousness: every experience is one out of many and generates information because it differs in its own way from the large repertoire of alternative experiences; and every experience is one, that is, integrated, because it cannot be decomposed into independent parts. The theory formalizes these notions by postulating that a physical system generates information by reducing uncertainty about which previous states could have caused its present state, and that this information is integrated to the extent that it cannot be partitioned into the information generated by parts of the system taken independently. The theory predicts that integrated information in wakefulness is higher than that in sleep. Integrated information can be measured rigorously in models such as the MVARX model presented here. The integration of information is captured by \mathbf{A} and \mathbf{Q} in the MVARX model— \mathbf{B} only indicates how stimulation enters the network. In this section we contrast integrated information in wakefulness and sleep using a variation on the procedure introduced in Barrett and Seth (2011) for obtaining a bipartition approximation to integrated information in MVAR systems. Our variation is based on use of “effective information” (Kullback–Leibler divergence) (Balduzzi and Tononi, 2008) in place of the difference in mutual information and ensures that integrated information is always positive (Cover and Thomas, 2006).

Suppose \mathbf{y}_n describes a stable MVAR(p) process:

$$\mathbf{y}_n = \sum_{i=1}^p \mathbf{A}_i \mathbf{y}_{n-i} + \mathbf{w}_n, \quad (26)$$

where \mathbf{w}_n are i.i.d. zero-mean Gaussian noise vectors with covariance \mathbf{Q} . Then the MVAR(p) process is wide sense stationary and $\mathbf{y}_n \sim \mathcal{N}(0, \Sigma(\mathbf{y}))$ with $\Sigma(\mathbf{y}) = E\{\mathbf{y}_n \mathbf{y}_n^T\}$. Given that the state at time n , $\mathbf{y}_n = \mathbf{y}$, the conditional distribution of the state τ samples prior to sample n , $\mathbf{y}_{n-\tau}$, follows

$$\mathbf{y}_{n-\tau} | (\mathbf{y}_n = \mathbf{y}) \sim \mathcal{N}(\Gamma_\tau(\mathbf{y}) \Sigma(\mathbf{y})^{-1} \mathbf{y}, \Sigma(\mathbf{y}_{n-\tau} | \mathbf{y}_n)) \quad (27)$$

where $\Gamma_\tau(\mathbf{y}) = E\{\mathbf{y}_{n-\tau} \mathbf{y}_n^T\}$ and

$$\Sigma(\mathbf{y}_{n-\tau} | \mathbf{y}_n) = \Sigma(\mathbf{y}) - \Gamma_\tau(\mathbf{y}) \Sigma(\mathbf{y})^{-1} \Gamma_\tau(\mathbf{y})^T. \quad (28)$$

Given \mathbf{A} and \mathbf{Q} , the matrices $\Sigma(\mathbf{y})$ and $\Gamma_\tau(\mathbf{y})$ for $\tau = 1, \dots, \rho$, with $\rho \geq p - 1$, are computed as described in Barrett and Seth (2011).

Let the set of the channels be $S = \{1, 2, \dots, d\}$. A bipartition $\mathcal{B} = \{M^1, M^2\}$, divides the channels into two mutually non-overlapping and non-empty sub-networks, $S = M^1 \cup M^2$.

Denote two sub-systems \mathbf{m}_n^1 and \mathbf{m}_n^2 within which are the measurements in the channels corresponding to the elements in M^1 and M^2 at time n , respectively. Given $\Sigma(\mathbf{y})$ and $\Gamma_\tau(\mathbf{y})$, we have $\Sigma(\mathbf{m}^i) = [\Sigma(\mathbf{y})]_{M^i, M^i}$ and $\Gamma_\tau(\mathbf{m}^i) = [\Gamma_\tau(\mathbf{y})]_{M^i, M^i}$, for $i = 1, 2$. Hence, given the present state, the conditional distribution of the sub-system i at τ samples into the past is given by $\mathbf{m}_{n-\tau}^i | (\mathbf{m}_n^i = \mathbf{m}^i) \sim \mathcal{N}(\Gamma_\tau(\mathbf{m}^i) \Sigma(\mathbf{m}^i)^{-1} \mathbf{m}^i, \Sigma(\mathbf{m}_{n-\tau}^i | \mathbf{m}_n^i))$, for $i = 1, 2$, where $\Sigma(\mathbf{m}_{n-\tau}^i | \mathbf{m}_n^i) = \Sigma(\mathbf{m}^i) - \Gamma_\tau(\mathbf{m}^i) \Sigma(\mathbf{m}^i)^{-1} \Gamma_\tau(\mathbf{m}^i)^T$.

Define the effective information for the system \mathbf{y} over a lag of τ samples under partition \mathcal{B} as [see Barrett and Seth (2011), (0.32)]

$$\varphi(\mathbf{y}; \tau, \mathcal{B}) = \frac{1}{2} \left[-\log_2(\det(\Sigma(\mathbf{y}_{n-\tau} | \mathbf{y}_n))) + \sum_{i=1}^2 \log_2(\det(\Sigma(\mathbf{m}_{n-\tau}^i | \mathbf{m}_n^i))) \right] \text{ bits.} \quad (29)$$

The effective information is the Kullback–Leibler divergence between a system consisting of two mutually independent sub-systems \mathbf{m}_n^1 and \mathbf{m}_n^2 and the system \mathbf{y}_n . The integrated information measured at a time difference of τ is defined as

$$\phi(\mathbf{y}; \tau) = \varphi(\mathbf{y}; \tau, \mathcal{B}^{\text{MIB}}) \quad (30)$$

where the minimum information bipartition (MIB) is defined as

$$\mathcal{B}^{\text{MIB}} = \arg \min_{\mathcal{B}} \left(\frac{\varphi(\mathbf{y}; \tau, \mathcal{B})}{K_2(\mathcal{B})} \right) \quad (31)$$

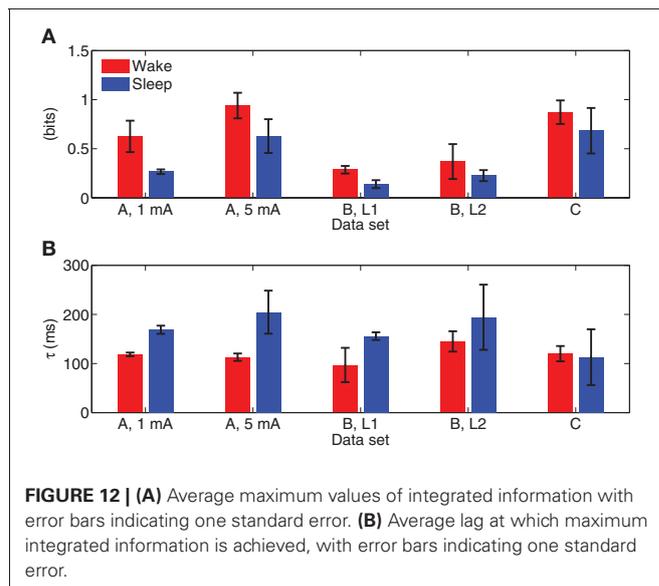
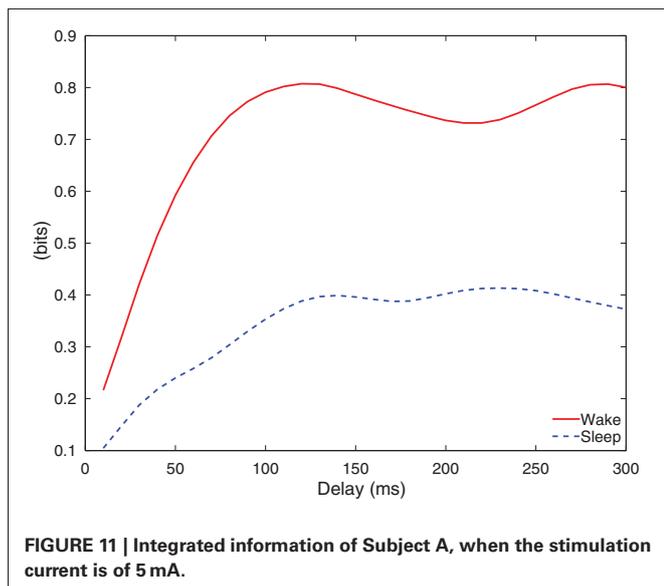
with

$$K_2(\mathcal{B}) = \min(H(\mathbf{m}_n^1), H(\mathbf{m}_n^2)) \quad (32)$$

and the differential entropy of \mathbf{m}_n^i , $H(\mathbf{m}_n^i)$ is given by

$$H(\mathbf{m}_n^i) = \frac{1}{2} \log_2 \left((2\pi e)^{|M^i|} \det(\Sigma(\mathbf{m}^i)) \right). \quad (33)$$

Figure 11 depicts the integrated information of Subject A for stimulus of 5 mA, as the time difference τ varies from 10 to 300 ms. The integrated information in wakefulness is higher than that in sleep. In both wakefulness and sleep, the integrated information increases until the time difference is approximately 100 ms and then remains approximately constant. We further used the CV procedures described in section 3.3 to study the difference between integrated information in wakefulness and sleep. Specifically, we estimated a model from the training set of each CV partition and compute integrated information for each CV partition. This provides M different estimates of integrated information for each data set, where M is the number of CV partitions. We compare the maximum values of the estimates of integrated information for each CV partition in wakefulness and sleep using the Wilcoxon rank sum test, which tests the null (H_0) hypothesis that the measured maximum integrated information values in wakefulness and sleep for all CV partitions are samples from continuous distributions with equal medians, against H_1 that they are not. The p -values of the rank sum test for each conditions are shown in **Table 3**. With the exception of Subject C, all of the cases



have p -values below 0.05, and Subject C is only slightly above 0.05. **Figure 12** depicts the average maximum value of integrated information and average time delay τ at which the maximum value is achieved, where the averaging is done across CV results, and error bars indicates one standard error.

6. DISCUSSION

The results demonstrate the effectiveness of the MVARX model for intracerebral electrical stimulation data. Excellent agreement between measured and modeled evoked responses is found across channels, two stimulus amplitudes, vigilance states, stimulus sites, and subjects (**Figures 4–7**). One-step prediction is used to show that the MVARX model also accurately captures the spontaneous fluctuations in the measured signals (**Figures 8** and **9**). We contrast the MVARX models with a series of univariate ARX models, one for each channel, to illustrate the importance of accounting for the interaction between cortical signals (**Figures 4–6**). In some channels for some subjects/conditions the univariate ARX model describes the evoked response as well as the MVARX model. However, in general modeling interactions between cortical signals is necessary to capture the measured response. For example, in **Figure 4B** the univariate model fails to model the responses in channels 1, 4, and 7 beyond 200 ms after the stimulation.

The MVARX model explicitly represents both evoked and spontaneous (or background) brain activity using a deterministic input term to capture the effect of stimuli and a random input term to generate spontaneous activity. Stimuli generally give rise to a non-zero mean component in the response that varies with time, i.e., is non-stationary. Conventional approaches to MVAR

modeling of cortical event-related potentials (e.g., Ding et al., 2000), subtract the ensemble mean of the data before processing to avoid the negative effects of the non-stationary mean on the MVAR model. However, subtraction of the ensemble mean significantly reduces the SNR of the data and is not necessary if the exogenous input is properly accounted for in the modeling procedure.

The effect of the stimulus on each recording channel is addressed by applying a separate filter in each channel to the stimulus signal. The filter coefficients are estimated jointly with the autoregressive model parameters from the measured evoked data. This approach accounts for the generally unknown and different characteristics of the transmission paths from the stimulation to each measurement site. The length of the filters [ℓ samples in Equation (5)] should be limited based on physiological expectations for the stimulus paradigm. Indeed, the autoregressive coefficients \mathbf{A}_i and filters \mathbf{b}_i are estimated simultaneously and the evoked response ($\mathbf{y}_{n,e}^{(j)}$ in Equation (8)) is often much larger than the spontaneous component [$\mathbf{y}_{n,s}^{(j)}$ in Equation (7)]. If ℓ is set equal to the duration of one epoch of $\mathbf{y}_{n,e}^{(j)}$, then it is possible to perfectly model $\mathbf{y}_{n,e}^{(j)}$ using only the \mathbf{b}_i while setting the $\mathbf{A}_i = 0$. We have shown that the MVARX models are capable of characterizing $\mathbf{y}_{n,s}$ by one-step prediction of data not used to estimate the model (see **Figure 8**). Moreover, the model describes the dynamics in $\mathbf{y}_{n,e}$, as was shown in **Figures 4–6**.

In order to define a practical value for ℓ we refer to previous electrophysiological studies on intracerebral evoked potentials (Matsumoto et al., 2004, 2007, 2012). In these studies Matsumoto and colleagues thoroughly discussed the possible

Table 3 | p -values of the Wilcoxon rank sum test of whether integrated information in wakefulness and sleep are different.

	Subject A, 1 mA	Subject A, 5 mA	Subject B, L1	Subject B, L2	Subject C
p -value	3.18e-4	0.0012	2.06e-4	0.0068	0.0553

generator mechanisms of intracerebral potentials evoked by direct electrical stimulation. In all of these studies it has been shown that the duration of the “purely evoked” response expires within 100 ms. Based on these results and our 100 Hz sampling frequency we set $\ell = 10$. The 100 ms value is also consistent with our data. Indeed, the first 100 ms post-stimulus of the evoked waveforms exhibit quite different character than later portions. Typically the initial 100 ms of the measured response contain relatively sharp, high frequency waveforms, while later portions of the response have a smoother, lower frequency behavior. This suggests two regimes in the modeling process. The exogenous input filters account for the sharp initial response, as evident by the filter impulse responses shown in **Figure 10**. Channels having relatively large impulse response tend to rapidly transition from negative to positive maxima over one or two samples, consistent with the sharp features in the early portions of the evoked response. These sharp inputs to the channels are smoothed by the autoregressive component of the model to obtain the later portions of the response. The filter responses depicted in **Figure 10** decay to relatively small values by the 10-th lag (100 ms) and generally contain most of their energy in the first through sixth lags, that is between 10 and 60 ms. This further supports the choice of $\ell = 10$.

The energy transmission characteristics shown in **Figure 10** are consistent with physiological expectations for modeling stimulation of fibers of passage. There is general consistency between wakefulness and sleep in all subjects (**Figure 10**, left column vs. right column) even though the evoked responses differ markedly (**Figure 4** vs. **Figure 5**); channels with strong and weak responses are the same in wakefulness and sleep, and the shape of the responses in each channel are generally very similar. The subtle differences between wakefulness and sleep may be due to changes in neural excitability. Comparing 1 and 5 mA stimulation in Subject A (**Figures 10A,B** and **C,D**) reveals that channel 11 has the strongest response in both stimulation levels and the strength of the response increases roughly by a factor of 5, consistent with the factor of 5 change in the stimulation level. This is because we used the trigger signal to represent the exogenous input without adjusting its amplitude. However, the shape of the response in channel 11 differs slightly, with the 5 mA case having reduced latency by approximately 10 ms and a higher frequency response reflected by the sharper, shorter duration of the filter. This suggests that the higher stimulus level is associated with a faster response. The two stimulation sites L1 and L2 in Subject B (**Figures 10E,F** and **G,H**) both involve channels 8 and 4 as the strongest response, suggesting similar fibers of passage are excited at the two sites. However, the overall gain differs by a factor of 2 and the shape of the response in channel 8 and 4 differ, especially in wakefulness. Subject C (**Figures 10I,J**) exhibits multiple channels with strong linkage to the stimulus site.

Our MVARX approach assumes the dynamic interactions between evoked and spontaneous cortical signals follow the same model, that is, both evoked and spontaneous activity are described by one set of A_i . The excellent one-step prediction performance in the pre-stimulus interval of **Figure 8** combined with the high quality fitting of the evoked responses suggests this is a

reasonable assumption, at least for these particular data sets. This approach also assumes that the measured signal is the sum of the evoked and spontaneous activity.

The windowed median filtering procedure successfully eliminated the volume conduction artifact while limiting changes to the measured signal to within ± 20 ms of the stimulation. The outlier detection strategy only eliminates epochs that have significant deviation from the average evoked response. Both of these strategies significantly improve model fidelity to the measured data. Seven times as many outlier epochs were identified in sleep than in wakefulness, likely due to the presence of occasional slow waves during an epoch. However, in seven of the ten data sets we analyzed 28 or more of the 30 available epochs, which indicates our artifact detection procedure is not overly aggressive. Subject A had the most outlier epochs and in the worst case (5 mA, sleep) our procedure eliminated 8 of the possible 30 epochs. The CV strategy for choosing MVAR model order is effective, as demonstrated by the fidelity of the model evoked responses (**Figures 4–7**) and the ability of the models to accurately perform one-step prediction on pre-stimulus data (**Figure 8**). Outlier rejection helps the data meet the stationarity assumption of the MVARX model. While it is unlikely that the data are truly stationary, the accuracy with which the model describes the data and the whiteness of the residuals suggests that the stationarity assumption is reasonable.

As a proof of concept application, we used the MVARX model to assess changes in the level of information integration between wakefulness and deep sleep in human subjects. Using a simple, bipartition approximation we found that, as predicted by theoretical considerations (Tononi, 2004; Seth et al., 2008), integrated information is higher in wakefulness than sleep for each subject/condition, supporting the notion that integrated information reflects the capacity for consciousness. We note that the integrated information results presented here only apply to the recordings analyzed. Analysis of the dependence of integrated information on recording coverage is beyond the scope of this paper. Our findings indicate that the human cerebral cortex is better suited at information integration—being both functionally specialized and functionally integrated—when awake and conscious. In contrast, when consciousness fades in deep sleep, the parameters of the system change in such a way that information integration is diminished, in line with theoretical predictions (Tononi, 2004) and consistent with qualitative evidence obtained from experiments employing transcranial magnetic stimulation and high density EEG (Massimini et al., 2005). We also found that the lag at which the maximum level of information integration is attained is consistently longer in sleep than wakefulness. Maximum information integration in wakefulness occurred at lags of 30–110 ms, while those in sleep were from 70 to 140 ms longer, consistent with the increased low frequency activity of sleep.

ACKNOWLEDGMENTS

This research was supported in part by the National Institute of Biomedical Imaging and Bioengineering under grant R21EB009749.

REFERENCES

- Astolfi, L., Cincotti, F., Mattia, D., De Vico Fallani, F., Tocci, A., Colosimo, A., et al. (2008). Tracking the time-varying cortical connectivity patterns by adaptive multivariate estimators. *IEEE Trans Biomed. Eng.* 55, 902–913.
- Babiloni, F., Cincotti, F., Babiloni, C., Carducci, F., Mattia, D., Astolfi, L., et al. (2005). Estimation of the cortical functional connectivity with the multimodal integration of high-resolution EEG and fMRI data by directed transfer function. *Neuroimage* 24, 118–131.
- Baccalá, L. A., and Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. *Biol. Cybern.* 84, 463–474.
- Balduzzi, D., and Tononi, G. (2008). Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput. Biol.* 4:e1000091. doi: 10.1371/journal.pcbi.1000091
- Barnett, L., and Seth, A. K. (2011). Behaviour of Granger causality under filtering: theoretical invariance and practical application. *J. Neurosci. Methods* 201, 404–419.
- Barrett, A. B., and Seth, A. K. (2011). Practical measures of integrated information for time-series data. *PLoS Comput. Biol.* 7:e1001052. doi: 10.1371/journal.pcbi.1001052
- Bernasconi, C., and König, P. (1999). On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings. *Biol. Cybern.* 81, 199–210.
- Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction*. Wiley Series in Probability and Statistics, 2nd Edn. New York, NY: Wiley-Interscience.
- Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., and Bressler, S. L. (2004). Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9849–9854.
- Cheung, B. L. P., Nowak, R. D., Lee, H. C., van Drongelen, W., and Van Veen, B. D. (2012). Cross validation for selection of cortical interaction models from scalp EEG or MEG. *IEEE Trans. Biomed. Eng.* 59, 504–514.
- Cossu, M., Cardinale, F., Castana, L., Citterio, A., Francione, S., Tassi, L., et al. (2005). Stereoelectroencephalography in the presurgical evaluation of focal epilepsy: a retrospective analysis of 215 procedures. *Neurosurgery* 57, 706–718.
- Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, 2nd Edn. New York, NY: Wiley-Interscience.
- Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., and Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn. Sci. (Regul. Ed.)* 10, 204–211.
- Ding, M., Bressler, S. L., Yang, W., and Liang, H. (2000). Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment. *Biol. Cybern.* 83, 35–45.
- Ding, M., Chen, Y., and Bressler, S. L. (2006). *Granger Causality: Basic Theory and Application to Neuroscience*. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA.
- Duchesne, P., and Roy, R. (2004). On consistent testing for serial correlation of unknown form in vector time series models. *J. Multivar. Anal.* 89, 148–180.
- Geweke, J. F. (1984). Measures of conditional linear dependence and feedback between time series. *J. Am. Stat. Assoc.* 79, 907–915.
- Hong, Y. (1996). Consistent testing for serial correlation of unknown form. *Econometrica* 64, 837–864.
- Kamiński, M. J., and Blinowska, K. J. (1991). A new method of the description of the information flow in the brain structures. *Biol. Cybern.* 65, 203–210.
- Korzeniewska, A., Crainiceanu, C. M., Kuś, R., Franaszczuk, P. J., and Crone, N. E. (2008). Dynamics of event-related causality in brain electrical activity. *Hum. Brain Mapp.* 29, 1170–1192.
- Laureys, S. (2005). The neural correlate of (un)awareness: lessons from the vegetative state. *Trends Cogn. Sci. (Regul. Ed.)* 9, 556–559.
- Lütkepohl, H. (2006). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.
- Malekpour, S., Li, Z., Cheung, B., Castillo, E., Papanicolaou, L., Kramer, A., et al. (2012). Interhemispheric effective and functional cortical connectivity signatures of spina bifida are consistent with callosal anomaly. *Brain Connect.* 2, 142–154.
- Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., and Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science* 309, 2228–2232.
- Matsumoto, R., Nair, D. R., Ikeda, A., Fumuro, T., LaPresto, E., Mikuni, N., et al. (2012). Parieto-frontal network in humans studied by cortico-cortical evoked potential. *Hum. Brain Mapp.* 33, 2856–2872.
- Matsumoto, R., Nair, D. R., LaPresto, E., Bingaman, W., Shibasaki, H., and Lüders, H. O. (2007). Functional connectivity in human cortical motor system: a cortico-cortical evoked potential study. *Brain* 130, 181–197.
- Matsumoto, R., Nair, D. R., LaPresto, E., Najm, I., Bingaman, W., Shibasaki, H., et al. (2004). Functional connectivity in the human language system: a cortico-cortical evoked potential study. *Brain* 127, 2316–2330.
- McQuarrie, A. D. R., and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. River Edge, NJ: World Scientific Pub Co Inc.
- Möller, E., Schack, B., Arnold, M., and Witte, H. (2001). Instantaneous multivariate EEG coherence analysis by means of adaptive high-dimensional autoregressive models. *J. Neurosci. Methods* 105, 143–158.
- Munari, C., Hoffmann, D., Francione, S., Kahane, P., Tassi, L., Lo Russo, G., et al. (1994). Stereo-electroencephalography methodology: advantages and limits. *Acta Neurol. Scand. Suppl.* 152, 56–67.
- Nobili, L., De Gennaro, L., Proserpio, P., Moroni, F., Sarasso, S., Pigorini, A., et al. (2012). Local aspects of sleep: observations from intracerebral recordings in humans. *Prog. Brain Res.* 199, 219–232.
- Nobili, L., Ferrara, M., Moroni, F., De Gennaro, L., Russo, G. L., Campus, C., et al. (2011). Dissociated wake-like and sleep-like electro-cortical activity during sleep. *Neuroimage* 58, 612–619.
- Penny, K. I. (1996). Appropriate critical values when testing for a single multivariate outlier by using the mahalanobis distance. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* 45, 73–81.
- Ranck, J. B. (1975). Which elements are excited in electrical stimulation of mammalian central nervous system: a review. *Brain Res.* 98, 417–440.
- Rechtschaffen, A., and Kales, A. (eds.). (1968). *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. NIH Publication No. 204. Washington, DC: US Government Printing Office, National Institute of Health Publication.
- Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., and Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends Cogn. Sci. (Regul. Ed.)* 12, 314–321.
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci.* 5:42. doi: 10.1186/1471-2202-5-42
- Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* 215, 216–242.
- Tononi, G. (2010). Information integration: its relevance to brain function and consciousness. *Arch. Ital. Biol.* 148, 299–322.
- Valentín, A., Alarcón, G., Honavar, M., Garcia Seoane, J. J., Selway, R. P., Polkey, C. E., et al. (2005). Single pulse electrical stimulation for identification of structural abnormalities and prediction of seizure outcome after epilepsy surgery: a prospective study. *Lancet Neurol.* 4, 718–726.
- Valentín, A., Anderson, M., Alarcón, G., Garcia Seoane, J. J., Selway, R., Binnie, C. D., et al. (2002). Responses to single pulse electrical stimulation identify epileptogenesis in the human brain *in vivo*. *Brain* 125, 1709–1718.
- Winterhalder, M., Schelter, B., Hesse, W., Schwab, K., Leistriz, L., Klan, D., et al. (2005). Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. *Signal Process.* 85, 2137–2160.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 24 August 2012; accepted: 07 November 2012; published online: 30 November 2012.

Citation: Chang J-Y, Pigorini A, Massimini M, Tononi G, Nobili L and Van Veen BD (2012) Multivariate autoregressive models with exogenous inputs for intracerebral responses to direct electrical stimulation of the human brain. *Front. Hum. Neurosci.* 6:317. doi: 10.3389/fnhum.2012.00317

Copyright © 2012 Chang, Pigorini, Massimini, Tononi, Nobili and Van Veen. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.