Investigation of Architectures for Models of Neural Responses to Electrical Brain Stimulation

Cynthia Steinhardt[1], Pierre Sacré[1], Sara K. Inati[2], Sridevi V. Sarma[1], Kareem A. Zaghloul[3]

***Abstract*— Electrical brain stimulation is used clinically to target pathological regions of the brain for treatment of diseases, such as Parkinson's disease, epilepsy and depression. Conventional treatments involve chronic implants that disrupt activity through a fixed periodic train of pulses or bursts of pulses applied to the affected region. However, stimulating one region of the brain necessarily affects other structurally and/or functionally connected areas. Understanding how connected regions of the brain are affected by stimulation at the implant site could improve treatment efficacy by informing optimal placement and stimulation patterns. In this study, we build predictive input-output models from intracranial recordings obtained from 10 epilepsy patients implanted with electrodes. Specific contacts within each subject were electrically stimulated (inputs), and evoked responses were simultaneously captured from all contacts (outputs). From these data, we constructed and compared four different dynamical models that contain causal linear and nonlinear components. All model architectures successfully predicted evoked responses to stimulation with single pulses and sequences of pulses. Results suggest that a linear time-invariant model in series with a quadratic non-linearity best captures the relationship between stimulation amplitudes and evoked responses.**

I. Introduction

Neural modulation studies using electrical brain stimulation (EBS), stimulation of brain with current pulses from internal electrodes, have increased in the last few decades. EBS has been shown to affect cognitive processes, such as decision making [1], or disrupt pathological neural activity [2], [3]. Other studies used EBS to uncover functional connectivity by stimulating in one region and recording evoked responses from other brain regions [4], [5].

While these studies have indicated causal effects of stimulation, which have inspired treatments, such as deep brain stimulation for Parkinsons disease [6] and depression [7], the effects of EBS on large scale neural circuit activity is less understood. Consequently, clinical implant placement and programming is based on relative improvement of perceived symptoms and safe levels of stimulation. Then, the fixed-amplitude stimulation settings are typically used throughout the course of treatment [6], [8]. For example, when programming deep brain stimulation in a Parkinson's patient, the pulse width, pulse amplitude, and pulse frequency are tuned to reduce tremor and bradykineasia. The goal is to suppress such symptoms *and* avoid undesirable side effects (i.e. impulsivity, paresthesia) that occur when neighboring regions are affected [6]. More precise treatments could be made if one could test all combinations of stimulus parameters, and measure the behavioral and neural responses. Unfortunately, this is a combinatorial search and thus impractical.

In such scenarios, predictive models can play an important role. If one can construct an input–output model, wherein the model computes the neural responses in multiple neighboring regions (output) to stimulation in a target region (input), then one can formally optimize stimulation parameters to achieve specific patterns in the neural network. Clinical treatment of medically refractory epilepsy (MRE) patients has created an opportunity to develop such models. Some MRE patients undergo invasive monitoring, wherein electrode contacts are implanted intracranially and stimulated for clinical evaluation of seizures. Past studies involving MRE subjects used fixed, high amplitude pulses, cortico-cortical evoked potentials (CCEPs), to induce evoked responses in other regions to understand connectivity and evoked responses shape [9].

In a previous study, we considered a two system model in which the response to negative (N) / cathodic-anodic biphasic pulses, and positive (P) / anodic-cathodic biphasic pulses of amplitude drawn from a uniform distribution, were modeled as two separate linear time-invariant systems [10]. This study aims to improve on this previous study by finding a model that can better predict the response to any amplitude of current within a safe range for stimulation. A distribution of amplitudes of current pulses were used to stimulate one site in the brain while recording from all other implanted electrodes in MRE patients. Criteria were made for determining whether the population of neurons close to an electrode was responsive to stimulation at the stimulation site. Then, four model architectures were compared for their ability to predict the response to single pulses and sequences of pulses of varying amplitudes.

II. Methods

A. Participants

Ten individuals (8 male; 32.5 ± 0.9 yr) with drug resistant epilepsy underwent a surgical procedure in which platinum recording contacts were implanted subdurally on the cortical surface. Placement of the contacts was determined by the clinical team in order to best localize epileptogenic regions for resection. Data were collected at the Clinical Center at the National Institutes of Health (Bethesda, MD). The research

The National Science Foundation Graduate Research Fellowship under Grant No. DG31746891 and the National Institutes of Health under grant 5T32GM119998-02 provided support for C.S., and R21 R21NS103113 provided support for S.V.S.

[1]Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218 (csteinh2@jhu.edu)

[2]Office of the Clinical Director, NINDS, National Institutes of Health, Bethesda, MD 20892

[3]Surgical Neurology Branch, NINDS, National Institutes of Health, Bethesda, MD 20892

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

protocol was approved by the Institutional Review Board, and informed consent was obtained from all participants.

B. Stimulation and Recording Protocol

While participants were monitored in the Epilepsy Monitoring Unit, we electrically stimulated two adjacent subdural contacts with biphasic pulses using a programmable neurostimulator (CereStim, Blackrock Microsystems, LLC., Salt Lake City, UT) and a custom built GUI (Fig 1a). In two individuals, we performed the same experiments at two separate stimulation locations. Thus, we considered there to be 12 independent stimulation datasets.

This experimental setup is similar to previous studies with CCEPs [5]. However, each pulse consists of a square-wave biphasic pulse, where each phase has a duration of 0.3 ms and was separated by a 0.05 ms gap (Fig. 1b). We used a biphasic pulse to avoid charge buildup on the cortical surface (for safety) and polarization of the electrode contacts which could reduce current density [11].

We chose stimulation sites based on two criteria. The clinical team verified that the locations were not directly involved in seizure activity and were relatively central to the entire set of implanted electrodes in a participant. These criteria maximize the chances of observing evoked responses at other contacts. Evoked responses were recorded from all other electrodes.

Electrodes with variance over a standard deviation away from the average variance of all electrodes were excluded from future analysis. A common average correction followed by filtering with a notch filter at 60 Hz (line noise), a high pass filter at 2 Hz to remove effects of electrode drift, and a 200 Hz low pass filter (at the upper bound of neural activity) were used to pre-process the data before further analysis [12]. The stimulation artifact zone was considered to be the 10 ms after stimulation and ignored in future analysis. Electrodes were divided into responsive and non-responsive categories using two criteria. The average evoked response to the training data needed to cross zero at least twice, and the energy within 11 to 311 ms of stimulation was significantly greater than the energy 312 to 612 ms after stimulation by a threshold determined using a knee point algorithm. This required the response to have at least a unimodal shape and to have a stronger response in the range of a typical evoked response than afterwards. This resulted in 52 responsive electrodes.

C. Training and Testing Experiments

To characterize the responses to EBS, we applied sequences of individual biphasic pulses to a pair of neighboring electrodes, while recording from the remaining electrodes (Fig. 1a). In most participants, we delivered pulses once every 800–1000 ms with a random jitter of 10 ms, resulting in average inter-stimulus interval across all participants of 867.0 ± 41.3 ms. For each individual pulse, we used a stimulation amplitude that was randomly drawn from a uniform distribution between approximately 8 and −8 mA

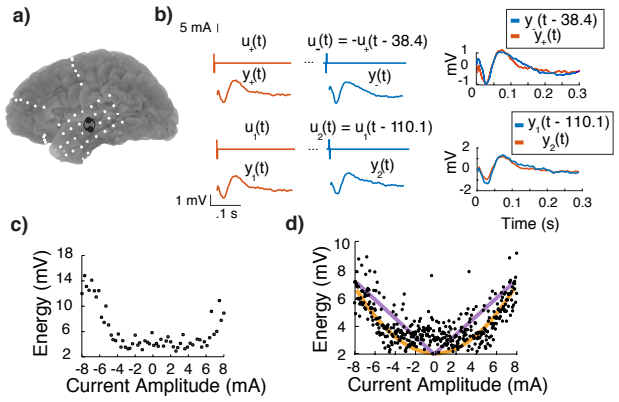

Fig. 1. **(a)** Stimulation site in one patient. **(b)** Sequence of biphasic pulses and resulting evoked responses on another electrode. Evoked responses to pulses of the same amplitude overtime and overlaid. Evoked responses to pulses of opposite polarity overtime and overlaid. The energy of response to a five minute training session, **(c)** showing asymmetric scaling of response to P and N pulses and **(d)** overlaid with a linear and quadratic fit.

(-7.93 ± 0.07 mA to 7.84 ± 0.16 mA). In the training session, we stimulated for five minutes, using approximately 300 individual stimulation pulses (Fig. 1a).

In the test session, we presented a novel sequence of individual pulses for two minutes (approximately 120 individual stimulation pulses). We used this novel sequence of pulses and their responses to test how well each model could predict the evoked response to novel stimulation.

Five participants received stimulation with multiple pulses per trial instead of one pulse per trial, where two to seven pulses could be given within 20 to 50 ms of one another per trial. These sessions were used to determine the ability of each model to predict the response to a signal comprised of a single pulse and multiple pulses.

D. System Models

The system was modeled with variations on a time invariant (TI) model due to the similarity of outputs to the same stimulation amplitude. Two shifted pulses of the same amplitude $u_1(t) = u_2(t - \tau)$, where $\tau = 110.1$ s, resulted in nearly identical evoked responses $y_1(t)$ and $y_2(t)$. This indicates a TI system (Fig. 1b). Two pulses $u_+(t)$ and $u_-(t)$, where $u_-(t) = -u_+(t)$ were considered to be time shifted versions of each other modulo a sign (Fig. 1b). Evoked responses to both inputs were also nearly identical, indicating a rectifying nonlinearity with amplitude. Nonlinearity selection was also data driven. Some electrodes showed asymmetric scaling based on pulse polarity (Fig. 1c). Additionally, the energy of evoked responses showed near linear and quadratic increase with amplitude (Fig. 1d). Models possessing various combinations of these features were probed to find the best explanatory model. All models share an LTI block and transform stimulation input $x(t)$ into evoked response $y(t)$.

1) Modeling the stimulation inputs: The biphasic stimulation input was idealized as a train of impulses with amplitudes drawn from a uniform distribution between −8 and 8 mA.

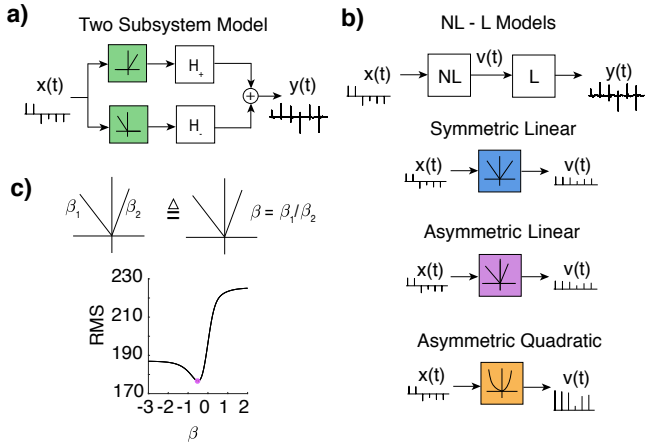

Fig. 2. **(a)** The architecture for the 2S model. **(b)** The architectures for NL-L models including how $x(t)$ transforms to $v(t)$ for the SL, AL, and AQ models. **(c)** Finding the beta ratio by grid search optimization.

TABLE I
NONLINEARITIES IN NL-L CASCADE MODEL $\mathcal{H}_{\text{NL-L}}$.

		symmetric $\beta = 1$	asymmetric $\beta \neq 1$
linear	$F(x) = \lvert x \rvert$	$\mathcal{H}_{\text{SL}}$	$\mathcal{H}_{\text{AL}}$
quadratic	$F(x) = \lvert x \rvert^2$	–	$\mathcal{H}_{\text{AQ}}$

2) Modeling the evoked responses: We consider two classes of models (Fig. 2).

The first model class is the parallel interconnection of two rectified linear time-invariant subsystems, a two subsystem (2S) model (Fig. 2a):

$$y(t) = h_+(t) * f_+(x(t)) + h_-(t) * f_-(x(t))$$

with $f_+(x) = \max(0, x)$ and $f_-(x) = \max(0, -x)$. This model structure is motivated by an initial assumption that P and N pulses evoked entirely different responses from the neural population.

The second model class is the cascade interconnection of a static nonlinear block and a linear time-invariant system

$$y(t) = \mathcal{H}_{\text{NL-L}}\{x(t)\} = h(t) * f_{\text{NL}}(x(t))$$

with

$$f_{\text{NL}}(x) = \begin{cases} F(x), & \text{if } x \geq 0, \\ \beta\, F(x), & \text{otherwise.} \end{cases}$$

Depending on the value that β takes and the nonlinearity $F(x)$, we have three models (see Table I): the symmetric linear (SL), the asymmetric linear (AL), and the asymmetric quadratic (AQ).

We assumed that each LTI block is causal and has a finite-duration impulse response, that is, $h(t) = 0$ for all $t < 0$ and $t > T$.

E. System Identification

Our goal is to estimate a model based on observing the response $y(t)$ of our system to an input $x(t)$. A widely used method, called minimum mean-square estimator (MMSE), is to choose the model $\mathcal{H}$ such that

$$\mathcal{H}_{\text{MMSE}} = \underset{\mathcal{H}}{\operatorname{argmin}}\, \mathbf{E}(\mathcal{H}\{x(t)\} - y(t))^2 \tag{1}$$

that is, to choose as our estimate a value that minimizes the mean-square error (MSE). See [13], [14] for details. When it is not fixed, β was found by gridding over values between 10^{-3} and 10^2 and by choosing the value of β that minimized the mean-square error.

F. Model Comparison

To compare the fit of the models on the data, predictions were made with each model by transforming the test input $x_{\text{test}}(t)$ by the optimal function found during the training session to produce a prediction of the evoked response $\hat{y}_{\text{test}}(t)$. We then computed the mean-square error of the predicted response as follow

$$\text{MSE} = \frac{1}{N} \sum_{t=0}^{N-1} (\hat{y}_{\text{test}}(t) - y_{\text{test}}(t))^2. \tag{2}$$

The distribution of MSE across responsive electrodes was compared across models through their mean and SEM. A lower and upper bound on error was measured. The lower bound MS_{noise} was the mean of the squares of the trace in the last 300 ms after stimulation, when no evoked response should occur; this is the contribution of noise to error. The upper bound $\text{MS}_{\text{signal}}$ was the mean of the squares of the trace in the first 300 ms after stimulation, which is proportional to the energy of the signal. Normalized Percent Signal Explained (NPSE) for some model was calculated as:

$$\text{NPSE}_{\text{model}} = 1 - \frac{\text{MSE}_{\text{model}} - \text{MS}_{\text{noise}}}{\text{MS}_{\text{signal}} - \text{MS}_{\text{noise}}} \tag{3}$$

which assess prediction similarity, discounting effects of noise.

Additionally, models were compared with coefficient of determination (R^2) and nAIC for robustness of model fit and overparameterization.

III. Results and Discussion

Twelve sites in MRE patients were stimulated with a sequence of P and N pulses of varying amplitudes, while evoked responses were recorded from the remaining electrodes. A training session of pulses was used to derive impulse response functions and optimize nonlinearities. These functions were then used to predict evoked responses to single pulses and sequences of pulses of stimulation.

The consistency of evoked responses over time and scaling of energy with stimulation amplitude inspired the assumption of a time-invariant model with a linear component. Prediction accuracy was considered the amount of the signal above the noise threshold that was predicted by the model (3) [10].

When predictions were made for evoked responses to single pulses, the SL model, which assumed identical responses to negative and positive stimulation, performed significantly worse than all models, but it captured over 60 % of the signal

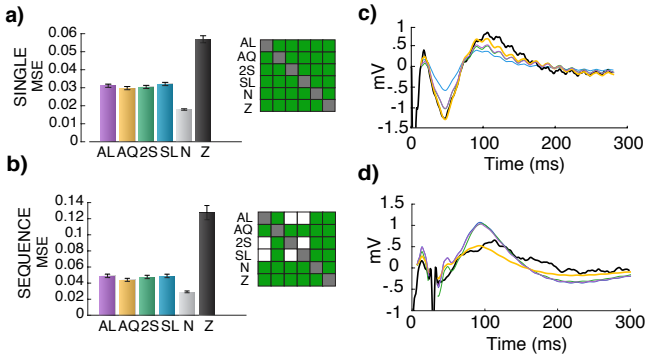

Fig. 3. Comparison of MSE of all models to error bounds for **(a)** single pulse test sessions across responsive electrodes and **(b)** multiple pulse test sessions with significance (right), where green is significant ($p < 0.05$) and white is not. A sample prediction of response to **(c)** a single pulse and **(d)** multiple pulses for all models.

TABLE II
MODEL COMPARISON METRICS

	Single Pulses			Multiple Pulses		
	R^2	nAIC	NPSE [%]	R^2	nAIC	NPSE [%]
Two Subsystem	0.288	8.45	65.3	0.359	8.44	81.2
Asymmetric Linear	0.286	6.45	63.7	0.356	6.45	79.8
Asymmetric Quadratic	0.302	6.44	65.6	0.387	6.42	84.8
Symmetric Linear	0.287	6.45	63.9	0.358	6.45	79.9

(Fig. 2b). This indicates a significant portion of the responses to N and P pulses are identical. The AL and AQ models add a degree of freedom via asymmetric scaling. The AL model has PSE slightly higher than the 2S model. However, the AQ model significantly outperforms the 2S model.

The PSE for predictions of the response to a sequence of pulses increases, indicating the the TI assumption holds. Performance differences between the SL, AL, and 2S model decrease to within 1 %. Meanwhile, the AQ model outperforms the other models by about 5% (Table II). The outperformance of the 2S model indicates overfitting. nAIC values also show the 2S model is relatively overparameterized. Asymmetric scaling does not have a large effect on PSE, given insignificant differences between SL and AL models. Instead, quadratic scaling has a significant effect that implies a quadratic scaling more closely reflects the local field response to EBS of increasing amplitude.

The success of the AQ model likely captures the difference between a biphasic P and N pulse. A biphasic P (N) pulse is an anodic (cathodic) pulse, followed by a cathodic (anodic) pulse. Cathodic and anodic pulses of the same current level have an axonal activation function with trimodal shape, which occurs at the same locations and has a central peak with polarity opposite the two side lobes. Cathodic pulses are shown to create an activation function that activates centrally and suppresses distally (which prevents an action potential) [15]. Thus, a P pulse leads to activation then suppression of the most local axons and activation of most distal axons, while N pulses activate both populations. This explains how P and N pulses lead to different responses, as in our model.

Current is thought to fall off with Euclidean distance from the stimulation site, a linear scaling. The quadratic nonlinearity could related to connectivity of neurons in this activated population to multiple neurons which drives exponential increase in activation that appears locally quadratic. Studies with two-photon stimulation have shown quadratic dependence on the probability for channel opening based on the intensity of stimulation, so this nonlinearity could also relate to higher current being more likely to drive the internal voltage of the axon higher and open channels [16].

The relative success of our AQ model indicates a static non-linearity functional neural responses to EBS across the brain. This more realistic model of neural responses across the brain to EBS offers a more accurate method for exploring how to induce desired effects when modulating brain activity when designed stimulation-based treatments for neurological disorders.

REFERENCES

[1] M. R. Cohen and W. T. Newsome, "What electrical microstimulation has revealed about the neural basis of cognition," *Current opinion in neurobiology*, vol. 14, no. 2, pp. 169–177, 2004.

[2] M. T. Salam, J. L. P. Velazquez, and R. Genov, "Seizure suppression efficacy of closed-loop versus open-loop deep brain stimulation in a rodent model of epilepsy," *IEEE Trans Neural Syst Rehabil Eng*, vol. 24, pp. 710–719, 2016.

[3] L. W. Lim *et al.*, "Electrical stimulation alleviates depressive-like behaviors of rats: investigation of brain targets and potential mechanisms," *Translational psychiatry*, vol. 5, no. 3, p. e535, 2015.

[4] J. Jacobs *et al.*, "Direct electrical stimulation of the human entorhinal region and hippocampus impairs memory," *Neuron*, vol. 92, no. 5, pp. 983–990, 2016.

[5] R. Matsumoto *et al.*, "Functional connectivity in the human language system: a cortico-cortical evoked potential study," *Brain*, vol. 127, no. 10, pp. 2316–2330, 2004.

[6] A. L. Benabid *et al.*, "Deep brain stimulation of the subthalamic nucleus for the treatment of parkinson's disease," *The Lancet Neurology*, vol. 8, no. 1, pp. 67–81, 2009.

[7] H. S. Mayberg *et al.*, "Deep brain stimulation for treatment-resistant depression," *Neuron*, vol. 45, no. 5, pp. 651–660, 2005.

[8] M. Rodriguez-Oroz *et al.*, "Bilateral deep brain stimulation in parkinson's disease: a multicentre study with 4 years follow-up," *Brain*, vol. 128, no. 10, pp. 2240–2249, 2005.

[9] T. Kunieda, Y. Yamao, T. Kikuchi, and R. Matsumoto, "New approach for exploring cerebral functional connectivity: review of cortico-cortical evoked potential," *Neurologia medico-chirurgica*, vol. 55, no. 5, pp. 374–382, 2015.

[10] C. Steinhardt *et al.*, "Using time-invariant models to construct whole brain neural responses to stimulation," *under review*.

[11] S. Brummer and M. Turner, "Electrochemical considerations for safe electrical stimulation of the nervous system with platinum electrodes," *IEEE Transactions on Biomedical Engineering*, no. 1, pp. 59–63, 1977.

[12] S. Gonzalez *et al.*, "Very high frequency oscillations as a predictor of movement intentions," *Neuroimage*, vol. 32, no. 1, pp. 170–179, 2006.

[13] L. Ljung, *System Identification: Theory for the User*. P T R Prentice Hall, 1987.

[14] P. Z. Marmarelis and V. Z. Marmarelis, *Analysis of physiological systems: the white-noise approach*, ser. Computers in biology and medicine. New York: Plenum Press, 1978.

[15] F. Rattay, "Ways to approximate current-distance relations for electrically stimulated fibers." *Journal of theoretical biology*, vol. 125, no. 3, pp. 339–349, 1987.

[16] B. K. Andrasfalvy, B. V. Zemelman, J. Tang, and A. Vaziri, "Two-photon single-cell optogenetic control of neuronal activity by sculpted light," *Proceedings of the National Academy of Sciences*, vol. 107, no. 26, pp. 11 981–11 986, 2010.