INVITED PAPER

Information Systems Opportunities in Brain–Machine Interface Decoders

This paper reviews systems that convert neural signals from motor regions of the brain into control signals to guide prosthetic devices, with a particular focus on how computational neuroscience knowledge informs the design of feedback control methods.

By Jonathan C. Kao, *Student Member IEEE*, Sergey D. Stavisky, *Student Member IEEE*, David Sussillo, Paul Nuyujukian, *Member IEEE*, and Krishna V. Shenoy, *Senior Member IEEE*

ABSTRACT | Brain–machine interface (BMI) systems convert neural signals from motor regions of the brain into control signals to guide prosthetic devices. The ultimate goal of BMIs is to improve the quality of life for people with paralysis by providing direct neural control of prosthetic arms or computer cursors. While considerable research over the past 15 years has led to compelling BMI demonstrations, there remain several challenges to achieving clinically viable BMI systems. In this review, we focus on the challenge of increasing BMI performance and robustness. We review and highlight key aspects of intracortical BMI decoder design, which is central to the conversion of neural signals into prosthetic control signals, and discuss emerging opportunities to improve intracortical BMI decoders. This is one of the primary research opportunities where information systems engineering can directly impact the future success of BMIs.

KEYWORDS | Brain–computer interface (BCI); brain–machine interface (BMI); control algorithm; decode algorithm; intracortical array; neural network; neural prosthesis

Manuscript received October 17, 2013; revised January 31, 2014; accepted February 6, 2014. Date of publication April 10, 2014; date of current version April 28, 2014. The work of J. C. Kao and S. D. Stavisky was supported by the National Science Foundation (NSF) Graduate Research Fellowships. The work of S. D. Stavisky was also supported by the National Science Foundation under IGERT Grant 0734683. The work of P. Nuyujukian was supported by the Stanford Medical Scientist Training Program, the Stanford Medical Scholars Program, the Howard Hughes Medical Institute Medical Research Fellows Program, and the Paul and Daisy Soros Fellowship. The work of K. V. Shenoy was supported by the Defense Advanced Research Projects Agency "REPAIR Program" under Contract N66001-10-C-2010, the National Institutes of Health (NIH) T-R01NS076460, and an NIH Director's Pioneer Award 8DP1HD075623.

J. C. Kao and **D. Sussillo** are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: jcykao@npl.stanford.edu; sussillo@stanford.edu).
S. D. Stavisky is with the Neurosciences Program, School of Medicine, Stanford University, Stanford, CA 94305 USA (e-mail: sergey.stavisky@stanford.edu).
P. Nuyujukian is with the Department of Bioengineering and the School of Medicine, Stanford University, Stanford, CA 94305 USA (e-mail: paul@npl.stanford.edu).
K. V. Shenoy is with the Department of Electrical Engineering, the Department of Bioengineering, the Department of Neurobiology, the Neurosciences Graduate Program, the Bio-X Program, and the Stanford Neurosciences Institute, Stanford University, Stanford, CA 94305 USA (e-mail: shenoy@stanford.edu).

Digital Object Identifier: 10.1109/JPROC.2014.2307357

I. INTRODUCTION

Millions of people worldwide suffer from motor-related neurological injury or disease, which in some cases is so severe that even the ability to communicate is lost (e.g., [1] and [2]). For people with lost motor function, brain–machine interfaces (BMIs), also known as neural prostheses or brain–computer interfaces (BCIs), have the potential to increase quality of life and enable greater interaction with the world.

Over the last 15 years, significant progress has been made toward realizing clinically viable BMI systems. As illustrated in Fig. 1, BMI systems comprise three major components: 1) sensors recording neural activity, typically from motor cortical regions of the brain; 2) a decoder, which translates the neural recordings into control signals;

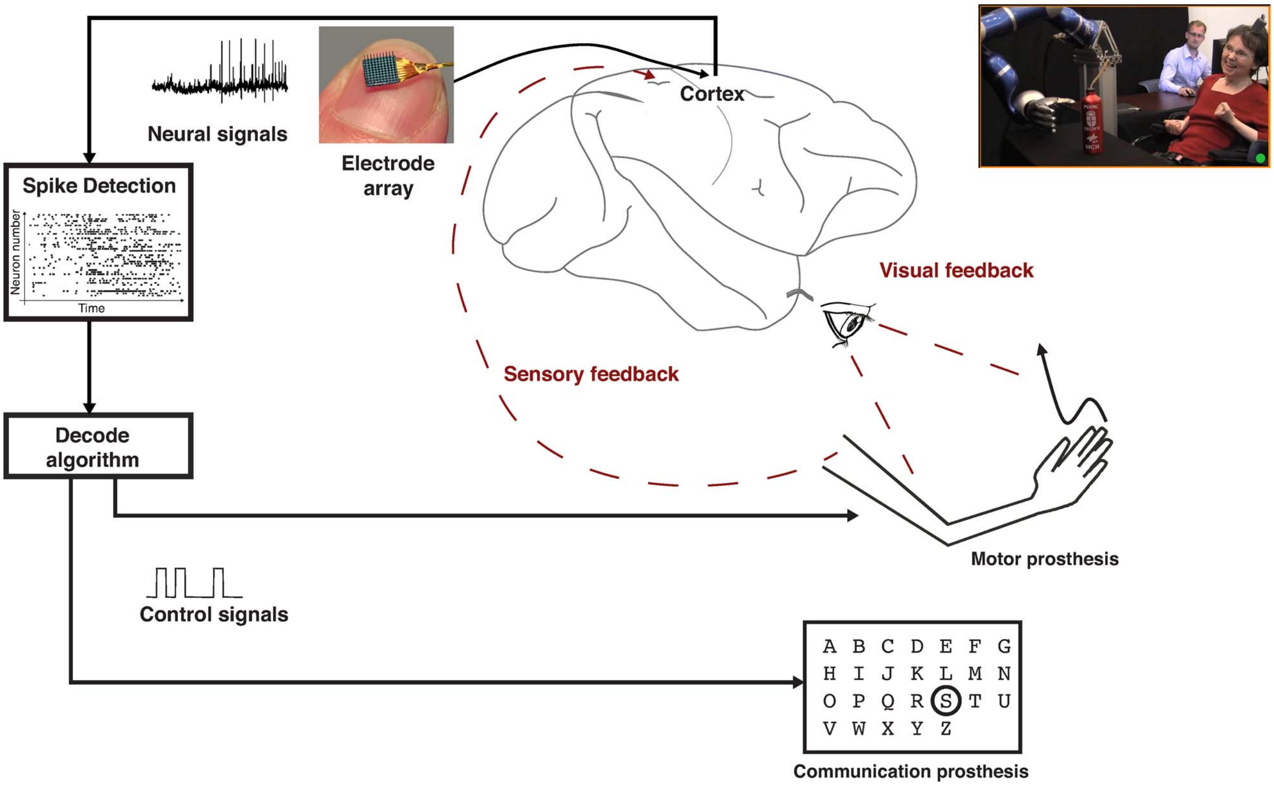


Fig. 1. ***BMI system overview. In an intracortical BMI system, neural signals are recorded from electrode arrays typically implanted in motor cortical regions of the brain. The raw neural signals (also shown in Fig. 2) are then passed through a spike detection algorithm, such as threshold detection, where a spike is detected if the measured electrode voltage crosses a preset threshold value. In the "spike detection" block, a black dot denotes that an action potential was measured. The neural spiking data are then sent to a decode algorithm, which outputs control signals (e.g., a digital signal) that guide a prosthetic device. The movement of the prosthetic device is observed by the subject, which closes a feedback loop. Though in its infancy, it is also possible to "write in" somatosensory and proprioceptive information into the brain, which may increase the controllability of the BMI system (e.g., [15] and [16]). This figure is adapted from [17] while the image of a BrainGate participant controlling a prosthetic arm is from [10].***

and 3) a prosthesis, such as a computer cursor on a screen or a robotic arm, controlled by the decoder.

BMIs have been based on several different neural information sources, including electroencephalographic (EEG) and electrocorticographic (ECoG) technologies. EEG and ECoG technologies measure average activity from large numbers of neurons with electrodes that reside on the scalp or surface of the brain, respectively (e.g., [3]–[8]). In this work, we focus on another major information source: intracortical neural signals. For modern BMI systems, intracortical neural signals are measured from electrodes that reside in the outer few millimeters of motor cortical regions of the brain. These electrodes measure action potentials from individual neurons and local field potentials (LFPs), as shown in Fig. 2. Action potentials, also known as "spikes," are the fundamental currency of information in the brain. Intracortical BMI systems have demonstrated compelling levels of performance in FDA phase-I clinical trials (e.g., [9]–[12]) as well as higher performance than BMIs based on alternative information sources (e.g., [13] and [14]).

Many challenges remain to achieving clinically viable BMI systems, including: 1) increasing BMI performance and robustness; 2) increasing the functional lifetime of implanted sensors; 3) replacing wires with wireless data telemetry and wireless powering; and 4) improving BMI ease of use, so that constant technician supervision is not required. We primarily focus on the first of these challenges: increasing BMI performance and robustness. The performance and robustness of BMI systems depend greatly on the decode algorithm (or "decoder"), which converts spiking activity from motor cortex into the kinematics of a prosthetic device. Because the decoder is integral to BMI performance and clinical viability, decode algorithm design must be optimized to provide subjects with high-quality neural control of a prosthetic device. To this end, the design of modern BMI decoders requires multidisciplinary research efforts which bring together a broad range of neuroscience, including systems and cognitive neuroscience, and multiple facets of information systems engineering, including statistical signal processing, estimation theory, machine learning, control theory, and information theory.

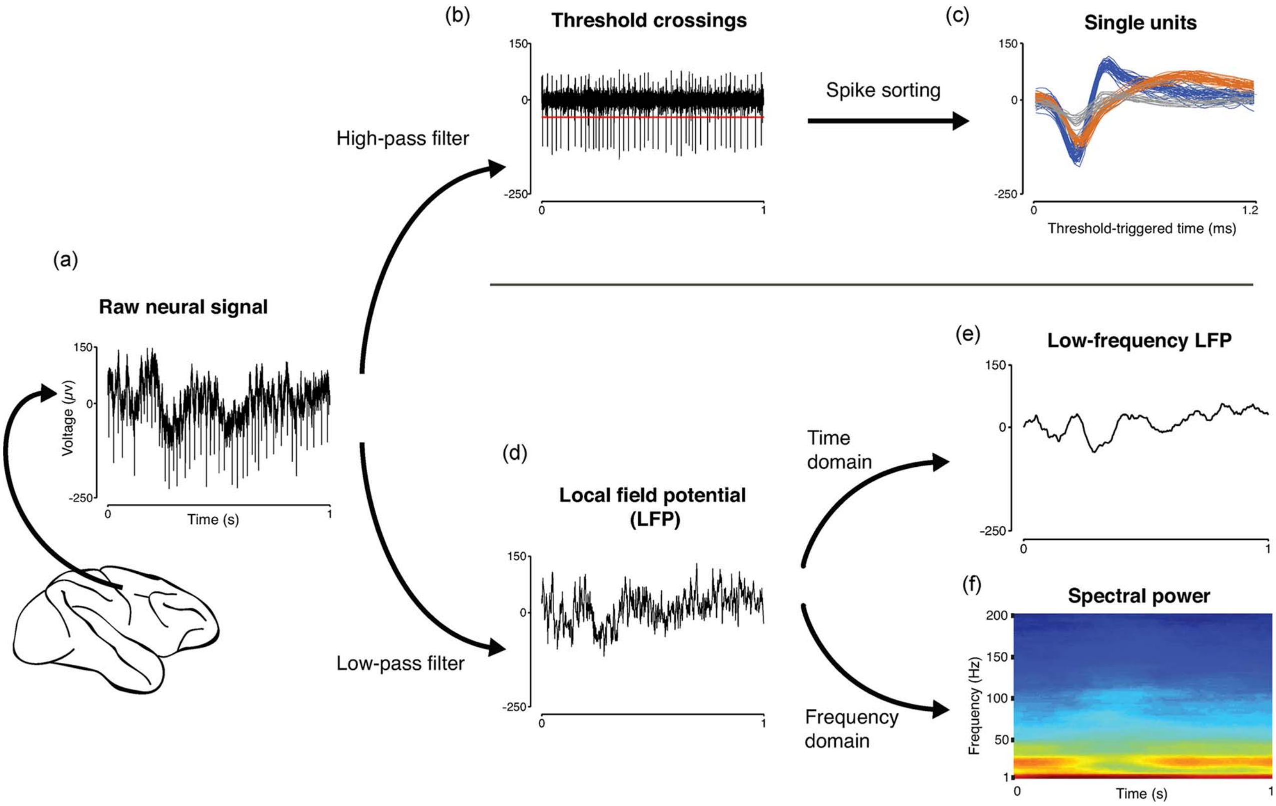


Fig. 2. *Raw neural signals and feature extraction. (a) The raw neural signal voltage is measured from an electrode in motor cortex. From the raw neural signal, two main signals can be extracted: action potentials (spikes) as shown in (b) and (c), and LFPs, as shown in (d)–(f). (b) For spikes, the raw neural signal is high-pass filtered, and a threshold is set (depicted in red) so that any voltage deflection crossing the threshold is counted as a spike. (c) It is occasionally the case that an electrode will measure spikes from different neurons simultaneously. A spike sorting algorithm can be used to separate spikes from different neurons by differentiating their waveforms. For example, waveforms arising from action potentials of two different neurons are shown in blue and orange, while gray represents activity that is not sorted. (d) For LFPs, the raw neural signal is low-pass filtered to remove spiking activity. (e) and (f) Various features of the LFP signal can be used. For example, different time-domain features of the LFP can be extracted, as in (e), or the spectral power in different frequency bands across time can be used, as in (f). While state-of-the-art BMI systems have relied on spiking activity only (e.g., [10], [11], and [14]), recent work has demonstrated BMI control using LFP activity, as further discussed in Section IV-B.*

In this review, we present aspects of BMI system design that may be of particular interest and relevance to information systems engineers. We first review decoder design approaches over approximately the last 15 years. In Section II, we discuss general classes of decode algorithms, which have been guided by both a neuroscientific understanding of motor cortex and statistical signal processing techniques. In Section III, we discuss how decode algorithms can be augmented by feedback control approaches, and present future directions and opportunities in decoder design. In Section IV, we briefly review recent BMI clinical studies and discuss challenges and opportunities that will be important for furthering clinical translation. While this review will present a particular perspective on decode algorithm design, as well as information systems opportunities that will be important for improving decode algorithms, we note that other review articles have also highlighted decoders, sensor interfaces, clinical translation, and other important challenges and opportunities facing BMIs (e.g., [16] and [18]–[20]).

II. A VIEW OF DECODE ALGORITHM DESIGN

The decode algorithm, which translates recorded neural population activity into prosthesis control signals, is essential for high-performance BMI systems. Historically, decoder design has been inspired by neuroscientific views of motor cortex as well as by linear estimation, statistical inference, and neural network theory. BMI decode algorithms are trained in a supervised fashion with simultaneous observations of real arm or prosthesis kinematics (e.g., [21]) and neural population activity. For example, a subject with motor neurological disease or injury may be asked to imagine mimicking the movements of an automated computer cursor while neural activity is recorded. A regression could then be performed to learn a mapping from the subject's recorded neural population

Table 1 A Categorization of BMI Algorithms

	Linear	Nonlinear
No State Space Model	II.A: Optimal linear estimator (e.g., [11], [26]) II.A: Population vector (e.g., [13], [27]) II.B: Wiener filter (e.g., [9], [25])	Artificial neural network (e.g., [28]–[30])
State Space Model	II.C: Kalman filter (KF) (e.g., [31], [32]) III.B: ReFIT-KF (control feedback approach) (e.g., [14])	II.D: Particle filter (e.g., [33]–[36]) II.D: Unscented KF (e.g., [37]) II.D: Laplace-Gaussian filter (e.g., [24], [38]) II.E: Echo state network (e.g., [39])

activity to the kinematics of the automated computer cursor. Then, during real-time BMI control, also called "online" or "closed-loop" control, the computer cursor would be causally controlled by the decoder, which uses the subject's real-time neural population activity to predict the prosthesis kinematics.

Closed-loop BMIs pose an additional challenge that other applications in information systems engineering do not routinely face. Consider training a supervised algorithm that infers a variable x from an observed variable y. To do so, one must learn a mapping $f(\cdot)$ so that $\hat{x} = f(y)$, where $\hat{x}$ is the estimate of x. A common approach is to learn f from observations of (x, y) ("training data") such that a desired error metric $\varepsilon(x, \hat{x})$ is minimized when evaluated on data not in the training set ("testing" or "cross-validation" data). In BMI settings, this approach can lead to suboptimal decoders. One reason for this is because the subject controlling the BMI system continuously observes the movements of the prosthesis and can make online corrections to compensate for inaccurately decoded kinematics (e.g., [22]–[24]). From a systems perspective, the subject *closes the feedback loop*, generating corrective neural responses that are absent in the training data. Thus, it is typically the case that BMIs running in closed-loop operate on data distributions that differ substantially from the data distributions of the training set (e.g., [25]). As a result of this, it is difficult to evaluate the performance of a putative decoder without running closed-loop experiments (e.g., [22] and [24]). While this poses a challenge for decoder design, there are opportunities to augment BMI systems by incorporating concepts from feedback control theory. Using feedback control approaches to increase the performance of BMI systems will be further discussed in Section III.

In this review, we will focus on decode algorithms which have been evaluated in closed-loop experiments using neural spiking activity. Although LFP activity is also measured from intracortical electrodes, as shown in Fig. 2, these signal sources have not been used in closed-loop BMI systems until only recently; we reserve discussion of the LFP signal to Section IV-B. A classification of BMI algorithms is shown in Table 1, which highlights the general categories of algorithms used in BMI systems. Throughout this review, we will use the following conventions: $\mathbf{x}_k \in \mathbb{R}^M$ denotes a column vector containing the M *observed* prosthesis kinematic variables at time k, $\hat{\mathbf{x}}_k$ denotes a vector containing the *decoded* prosthesis kinematic variables at time k, and $\mathbf{y}_k \in \mathbb{R}^N$ denotes a vector containing the recorded neural spiking activity of N neurons at time k. As an example, if the kinematic variables of interest are the position $(\mathbf{p}_k)$ and velocity $(\mathbf{v}_k)$ of a robotic arm at time k, then $\mathbf{x}_k$ would be the vertical concatenation of vectors $\mathbf{p}_k$ and $\mathbf{v}_k$. If y_k^i is the neural spiking activity of the ith neuron at time k, then $\mathbf{y}_k = [y_k^1 \ y_k^2 \ \ldots \ y_k^N]^{\mathsf{T}}$ is the activity of all the recorded neurons at time k, where $[\cdot]$ denotes horizontal concatenation and $\mathbf{y}^{\mathsf{T}}$ denotes the transpose of vector $\mathbf{y}$. For convenience, we also define matrices $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \ldots \ \mathbf{x}_K]$ and $\mathbf{Y} = [\mathbf{y}_1 \ \mathbf{y}_2 \ \ldots \ \mathbf{y}_K]$, with K denoting the number of observed time instances.

As this review focuses on spike-based decoders, the neural spiking activity y_k^i is typically the "binned spike counts" of neuron i. This quantity is computed by counting the number of times neuron i spikes in non-overlapping intervals (or bins) of length Δt. The interval Δt tends to be on the order of tens of milliseconds [22]. By binning time and counting the number of spikes within those bins, one is estimating an underlying neural firing rate. The time length of the data sampled in matrices $\mathbf{X}$ and $\mathbf{Y}$ is $\mathcal{T} = K\Delta t$.

A. Linear Vector Algorithms

An early BMI decoder algorithm proposed by Georgopoulos *et al.* is the population vector (PV) algorithm, which is based on a neurophysiological result: under certain conditions, the cosine of the reach direction can, in part, describe the firing rate of neurons in macaque motor

Table 2 Performance of BMI Algorithms

Study	Algorithm	Fitts throughput (bps)
Kim et al., 2008 [32]	VKF	0.52
Taylor et al., 2002 [27]	PV	0.53
Gilja et al., 2012 [14]	VKF	0.69
Ganguly et al., 2009 [53]	WF	0.95
Sussillo et al., 2012 [39]	ESN	1.45
Gilja et al., 2012 [14]	ReFIT-KF	1.81

This table shows the performance, measured by Fitts throughput (e.g., [46]), of closed-loop BMI studies using the population vector algorithm (PV), Wiener filters (WF), velocity-based Kalman filters (VKF), echo state networks (ESN), and a Kalman filter inspired from a control feedback approach (ReFIT-KF). Although Fitts throughput has units of bits per second (bps), this metric is distinct from Shannon information rates. Furthermore, Fitts throughput is not an achieved bitrate, but a relative measure for comparing BMI closed-loop performance across studies. Indeed, other works demonstrate theoretical and achieved BMI communication rates that are significantly higher than the Fitts throughput (e.g., [54], [55]). Although Fitts throughput has been suggested as a standardized metric for the assessment of neural prosthetics [47], due to several factors, including the variability of experimental conditions, tasks, electrode array quality, and subject skill across studies, the comparisons shown here cannot be exact. Fitts throughput is taken from Supplementary Table 3.1 of the study by Gilja and colleagues [14], while the Fitts throughput of the echo state network was calculated using the study by Sussillo and colleagues [39].

cortex, as discussed further below [40]. The PV algorithm has been used in several BMI systems (e.g., [13], [27], and [41]). In the PV algorithm, $\mathbf{x}_k$ is typically the velocity of the prosthetic device, so that the goal of the decoder is to estimate the direction and speed of movement. The PV algorithm is based on a *representational* view of motor cortex, in which the neural activity of individual neurons represents kinematic variables (e.g., [40] and [42]). According to this view, the activity of the ith neuron at any time can be described as a function of the kinematic variables: $y_k^i = f_i(\mathbf{x}_k)$. In this manner, a "tuning curve" can be built, where the average firing rate of a neuron is computed for different reach directions. These firing rates are subsequently interpolated (across reach directions) with one period of a cosine wave [43], so that $y_k^i \propto \cos(\theta_k + \phi)$, where θ_k is the angle of reaching and ϕ is a learned parameter. The direction in the kinematic space for which the neuron i is modeled to fire most strongly is called the preferred direction of neuron i, denoted by a unit vector $\mathbf{d}^i \in \mathbb{R}^M$. The contribution of each vector $\mathbf{d}^i$ to the decoded kinematics $\hat{\mathbf{x}}_k$ is linearly proportional to the firing rate of the neuron. Hence,

$$\hat{\mathbf{x}}_k = \frac{c}{N}\sum_{i=1}^{N} \frac{y_k^i - b^i}{\alpha^i}\mathbf{d}^i \tag{1}$$

where b^i is an offset term (typically the mean firing rate of neuron i) such that, when $y_k^i < b^i$, the neuron contributes movement in the direction $-\mathbf{d}^i$. The variable α^i is a weighting term, typically chosen to be the modulation depth of the neuron, or a variance normalizing term. The term c is a constant to make the sum proportional to speed. If we define $z_k^i = (y_k^i - b^i)/\alpha^i$ to be the normalized firing rates, then the PV algorithm can be written as $\hat{\mathbf{x}}_k = (c/N)\sum_{i=1}^N z_k^i \mathbf{d}^i$. For convenience, we also define vector $\mathbf{z}_k = [z_k^1\ z_k^2\ \ldots\ z_k^N]^\mathsf{T}$ and matrix $\mathbf{Z} = [\mathbf{z}_1\ \mathbf{z}_2\ \ldots\ \mathbf{z}_K]$. We also note that in some studies, the neural data are smoothed by low-pass filtering (e.g., [13]).

An algorithm which generalizes the PV algorithm, known as the optimal linear estimator (OLE), similarly takes a representational view of motor cortex, but does not find the preferred directions $\mathbf{d}^i$ by computing the tuning curve of each neuron. Instead, the vectors are found by considering the correlations of the neurons and their cross correlation to kinematics: $\mathbf{d}_{\mathrm{OLE}}^i = \sum_{j=1}^N \mathbf{R}_{ij}^{-1}\mathbf{V}_j^\mathsf{T}$ for $\mathbf{R}_{ij} = \mathbb{E}[z_k^i z_k^j]$, the (i,j) entry of the correlation matrix of the normalized firing rates, and $\mathbf{V}_j = \mathbb{E}[z_k^j \mathbf{x}_k^\mathsf{T}]$, the jth row of the cross-correlation matrix of the firing rates and kinematics [26]. The expectations are evaluated over time. If the distribution of the vectors $\mathbf{d}^i$ is uniform, then it can be shown that $\mathbf{d}_{\mathrm{OLE}}^i/\|\mathbf{d}_{\mathrm{OLE}}^i\| = \mathbf{d}^i$ so that the OLE and PV algorithms are equivalent. $\mathbf{R}$ and $\mathbf{V}$ are typically estimated by their time-averaged estimates $\mathbf{R} = \mathbf{Z}\mathbf{Z}^\mathsf{T}$ and $\mathbf{V} = \mathbf{Z}\mathbf{X}^\mathsf{T}$. Thus, OLE corresponds to a least squares solution. By defining $\mathbf{L}_{\mathrm{OLE}} = \mathbf{R}^{-1}\mathbf{V}$, it is apparent that $\mathbf{L}_{\mathrm{OLE}}$ minimizes the squared error of $\mathbf{X} - (\mathbf{L}_{\mathrm{OLE}})^\mathsf{T}\mathbf{Z}$. Here, the ith row of $\mathbf{L}_{\mathrm{OLE}}$ corresponds to the preferred direction $\mathbf{d}_{\mathrm{OLE}}^i$. In closed-loop control, the kinematics can be decoded by calculating $\hat{\mathbf{x}}_k = (\mathbf{L}_{\mathrm{OLE}})^\mathsf{T}\mathbf{z}_k$. Typically, a constant bias term $\mathbf{b}_{\mathrm{OLE}}$ is also included so that $\hat{\mathbf{x}}_k = (\mathbf{L}_{\mathrm{OLE}})^\mathsf{T}\mathbf{z}_k + \mathbf{b}_{\mathrm{OLE}}$. Recently, a clinical demonstration [11] used a variant of OLE called indirect OLE [44], in which a linear tuning model $\mathbf{Z} = \mathbf{B}^\mathsf{T}\mathbf{X}$ was learned and subsequently used to infer $\mathbf{L}_{\mathrm{OLE}}$ by setting $\mathbf{L}_{\mathrm{OLE}} = (\mathbf{B}^\dagger)$ (e.g., [23], [24], and [44]) where $\mathbf{B}^\dagger$ denotes the pseudoinverse of $\mathbf{B}$. It is worth noting that the performance of PV and OLE decoders is comparable in closed-loop systems (e.g., [23] and [24]).

Recent studies have noted that linear vector methods have demonstrated poorer quality control than Bayesian algorithms in closed-loop BMI systems (e.g., [24] and [45]). Although it is difficult to compare the closed-loop performance of decoders across experimental studies, Fitts throughput [46] has been suggested as a potential metric to perform this comparison (e.g., [14] and [47]). We note that due to several factors, including the variability of tasks, array quality, and subject skill across studies, these comparisons cannot be exact. Fitts throughput is further discussed in Supplementary Materials section 3.1 of the study by Gilja *et al.* [14]. Using Fitts throughput, we note that the performance of linear vector algorithms tends to be lower than others reported in the literature, as shown in Table 2. One reason for this performance difference may be due to additional modeling assumptions in other algorithms that are not present in linear vector techniques, such as smoothness in the decoded kinematic variables or the incorporation of noise models. Another potential reason for this performance difference may result from the static "preferred direction" assumption, where each neuron only encodes velocity in a single direction. Indeed, recent studies report that motor cortical neuron responses cannot be explained by static preferred directions alone. For example, studies demonstrated that the preferred direction of a neuron can change significantly based on the speed of a reach [48], or even over the course of a reaching movement [49]. Furthermore, learning the tuning curve of a neuron requires an approach where the neural data are modeled to be a function of the kinematic variables (e.g., [40]) or intended kinematic variables (e.g., [50]) so that $y_k^i = f_i(\mathbf{x}_k)$. However, the temporal responses of the neural activity may potentially be far more complex than the kinematics used to describe them, which would pose a limitation for this model (e.g., [49]). Some recent studies put forward a model with opposite causality, where kinematic variables are modeled to be functions of the neural population activity in motor cortex (e.g., [42], [51], and [52]) so that $\mathbf{x}_k = g(\mathbf{y}_k, \mathbf{y}_{k-1}, \ldots)$. Depending on these assumptions, decoder implementations will be somewhat different.

Interestingly, the linear vector methods, while inspired from a neurophysiological approach, have a natural and standard interpretation from an engineering viewpoint. The OLE method, which generalizes the PV algorithm, can be viewed as the least squares regression between a sequence of kinematic data and a corresponding sequence of neural data. While least squares has been standard in estimation literature as far back as Gauss and Legendre, linear estimation has developed significantly since that time [56]. We next review more recent BMI decoders stemming from advances in linear estimation theory.

B. Wiener Filters

The Wiener filter was a seminal contribution in estimation theory, helping to bring a statistical point of view into communication and control theory [57]. Both Wiener [58] and Kolmogorov [59] independently developed filtering theory in which a noisy sequence of observations $\mathbf{y}_1, \ldots, \mathbf{y}_k$ is used to calculate a linear estimate of a signal $\mathbf{x}_k$, given by $\hat{\mathbf{x}}_k = \sum_{j=1}^k \mathbf{L}_j^\top \mathbf{y}_j$, where $\mathbf{L}_j \in \mathbb{R}^{N \times M}$. In the Wiener–Kolmogorov filtering theory, the goal is to learn parameters $\mathbf{L}_1, \ldots, \mathbf{L}_k$ such that the squared error in predicting $\mathbf{x}_k$ is minimized. A major distinction of the Wiener–Kolmogorov approach, in contrast to linear vector techniques, is the incorporation of neural history $(\mathbf{y}_{k-1}, \mathbf{y}_{k-2}, \ldots)$ into the regression problem. In this section, we will describe the implementation of the Wiener filter by referring to the binned spike counts $\mathbf{y}_k$, but the Wiener filter could also be implemented using normalized spike counts $\mathbf{z}_k$.

In BMI systems, the Wiener filter (e.g., [9], [25], and [32]) is typically implemented in the following fashion: for a history of length $p\Delta t$, the decoded kinematics are

$$\hat{\mathbf{x}}_k = \sum_{j=0}^{p-1} \mathbf{L}_j^\top \mathbf{y}_{k-j}. \tag{2}$$

(As in the OLE case, a constant bias term can also be included.) By defining $\mathbf{L}_W = [\mathbf{L}_0^\top \mathbf{L}_1^\top \ldots \mathbf{L}_{p-1}^\top]^\top$, the vertical concatenation of the matrices $\mathbf{L}_0, \mathbf{L}_1, \ldots, \mathbf{L}_{p-1}$, the Wiener filter solution can be obtained by solving

$$\mathbf{L}_W = \begin{bmatrix} \mathbf{R}_{yy}(0) & \mathbf{R}_{yy}(1) & \ldots & \mathbf{R}_{yy}(p-1) \\ \mathbf{R}_{yy}(1) & \mathbf{R}_{yy}(0) & \ldots & \mathbf{R}_{yy}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{yy}(p-1) & \mathbf{R}_{yy}(p-2) & \ldots & \mathbf{R}_{yy}(0) \end{bmatrix}^{-1} \times \begin{bmatrix} \mathbf{R}_{yx}(0) \\ \mathbf{R}_{yx}(1) \\ \vdots \\ \mathbf{R}_{yx}(p-1) \end{bmatrix} \tag{3}$$

where $\mathbf{R}_{yy}(j) = \mathbb{E}(\mathbf{y}_k \mathbf{y}_{k+j}^\top)$ and $\mathbf{R}_{yx}(j) = \mathbb{E}(\mathbf{y}_k \mathbf{x}_{k+j}^\top)$ for all $j = 0, \ldots, p-1$. The index j refers to autocorrelations ($\mathbf{R}_{yy}(j)$) or cross-correlations ($\mathbf{R}_{yx}(j)$) at a lag of time j. We note that when $p = 1$ (i.e., no neural history), this approach reduces to the OLE method, since $\mathbf{L}_{\mathrm{OLE}} = \mathbf{R}_{zz}^{-1}(0)\mathbf{R}_{zx}(0)$, where $\mathbf{R}_{zz}(0)$ and $\mathbf{R}_{zx}(0)$ are the normalized firing rate analogs of $\mathbf{R}_{yy}(0)$ and $\mathbf{R}_{yx}(0)$. The autocorrelations and cross-correlations are typically estimated by their time-averaged estimates. We let $\mathbf{X}_{[i:j]}$ denote $[\mathbf{x}_i \ \mathbf{x}_{i+1} \ \ldots \ \mathbf{x}_j]$ for $i < j$, and define the following matrix:

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \mathbf{y}_p & \mathbf{y}_{p+1} & \cdots & \mathbf{y}_K \\ \mathbf{y}_{p-1} & \mathbf{y}_p & \cdots & \mathbf{y}_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_1 & \mathbf{y}_2 & \ldots & \mathbf{y}_{K-p+1} \end{bmatrix}. \tag{4}$$

Then, the time-averaged estimate of $\mathbf{L}_W$ can be calculated as $\mathbf{L}_W = (\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^\top)^{-1}(\tilde{\mathbf{Y}}\mathbf{X}_{[p:K]}^\top)$. For correlation-ergodic signals, as K approaches infinity, this formulation converges to the solution in (3) [60]. Several studies have used this approach for BMI decoding (e.g., [9], [25], and [28]).

An important parameter to choose in fitting the Wiener filter is p. If the time required to make a reach with the prosthesis is approximately τ, then choosing p such that

$p\Delta t > \tau$ is illogical, since the regression would be considering autocorrelations at lags longer than the timescale of the reach. Rather, p should be chosen so as to match the timescales at which the neural data are informative of the kinematics while not contributing significant lag to the system. Assigning significant weight to neural data relatively far into the past will likely cause the decoder to have significant lag in responding to the subject's changing intention. Hence, one approach to choose p is to evaluate Wiener filters for varying p in closed-loop BMI control. In this manner, the potential to overfit coefficients $\mathbf{L}_j$ for large j (corresponding to neural data further in the past) is minimal. Another approach to avoid overfitting the large number of coefficients is to regularize the regression using a technique such as ridge regression (e.g., [37]).

An interpretation of the Wiener–Kolmogorov filtering approach is that it provides optimal smoothness over a history of neural data of length $p\Delta t$ that is least squares optimal. While, at first, the Wiener filter may seem to be an extension of OLE to neural data with multiple time lags, there is a distinct difference between Wiener filter techniques and linear vector methods: with Wiener filters, the preferred directions associated with a neuron can be different at distinct time lags. This is apparent when considering that the ith rows (i.e., the "preferred directions") of $\mathbf{L}_j$ and $\mathbf{L}_k$ can be different for $j \neq k$. Therefore, the Wiener filter is not built in a framework that assumes static preferred directions. A consequence of this filtering framework is that because movements in different directions will have different temporal neural responses, the direction that a neuron drives the prosthesis can differ for different targets. This is a departure from linear vector methods, where a neuron can only move the prosthesis along a single direction. To our knowledge, no closed-loop comparisons within the same study have been performed between decoders using the Wiener–Kolmogorov approach and the linear vector approach. We note that across closed-loop studies in the literature, decoders using the Wiener–Kolmogorov approach have achieved higher Fitts throughput than linear vector methods, as shown in Table 2. The Wiener–Kolmogorov approach was used in the first BMI clinical studies with intracortical electrode arrays, demonstrating that a human could control a computer cursor and perform rudimentary actions with a multijointed robotic arm [9].

C. Kalman Filters

In 1960, Kalman introduced the state–space framework to filtering, which was a crucial and enabling insight facilitating finite-time and nonstationary analyses [61], [62]. The Kalman filter is a recursive algorithm that estimates the current state of a dynamical system given an observation of the output of the dynamical system and the previous state estimate. In general, state-of-the-art BMI systems using Kalman filtering model the prosthesis kinematics as the state of a linear dynamical system with certain dynamical update laws. In this dynamical modeling, it is typically assumed that the kinematics obey physical laws and are smooth over time (e.g., [10], [14], [32], and [63]).

In 2003, Wu *et al.* proposed a Kalman filter technique to estimate the kinematics of a prosthetic device given observations of the neural population activity $\mathbf{y}_k$ [31]. The dynamical model proposes that the kinematics of the prosthesis $\mathbf{x}_k$ are the state of a linear time-invariant dynamical system, while the neural activity $\mathbf{y}_k$ is the output of the dynamical system. The state and output process are both modeled to have Gaussian noise. Therefore, the system can be written as

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{w}_k \tag{5}$$
$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{q}_k \tag{6}$$

with $\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{W})$ and $\mathbf{q}_k \sim \mathcal{N}(0, \mathbf{Q})$. Because sequences $\{\mathbf{x}_k\}_{k=1,\ldots,K}$ and $\{\mathbf{y}_k\}_{k=1,\ldots,K}$ are observed in the training set while $\mathbf{w}_k$ and $\mathbf{q}_k$ are zero mean terms, $\mathbf{A}$ and $\mathbf{C}$ can be learned via least squares regression: $\mathbf{A} = \mathbf{X}_{[2:K]}\mathbf{X}_{[1:K-1]}^{\mathsf{T}}(\mathbf{X}_{[1:K-1]}\mathbf{X}_{[1:K-1]}^{\mathsf{T}})^{-1}$ and $\mathbf{C} = \mathbf{Y}\mathbf{X}^{\mathsf{T}}(\mathbf{X}\mathbf{X}^{\mathsf{T}})^{-1}$. After learning $\mathbf{A}$ and $\mathbf{C}$, $\mathbf{W}$ is calculated as the sample covariance of the residuals $\mathbf{X}_{[2:K]} - \mathbf{A}\mathbf{X}_{[1:K-1]}$ while $\mathbf{Q}$ is analogously the sample covariance of the residuals $\mathbf{Y} - \mathbf{C}\mathbf{X}$. Given $\mathbf{A}, \mathbf{W}, \mathbf{C}, \mathbf{Q}$ as well as an initial state condition, $\mathbf{x}_0$ (often set to be zero), the Kalman filter recursively estimates the current state $\hat{\mathbf{x}}_k$, given the current neural observation $\mathbf{y}_k$, and the previous state estimate $\hat{\mathbf{x}}_{k-1}$ [64]. Several laboratory and clinical demonstrations have used these kinematic-state Kalman filters in closed-loop BMI systems (e.g., [10], [14], and [32]).

A benefit in modeling a dynamical update law for the kinematic variables is the ability to enforce that the prosthesis movements obey physical kinematic laws. For example, if $\mathbf{x}_k = [\mathbf{p}_k^{\mathsf{T}}\ \mathbf{v}_k^{\mathsf{T}}]^{\mathsf{T}}$, then the $\mathbf{A}$ matrix can be additionally designed such that the position obeys $\mathbf{p}_{k+1} = \mathbf{p}_k + \mathbf{v}_k\Delta t$. Further, the $\mathbf{A}$ matrix provides a measure of smoothing or low-pass filtering over the kinematic variables. This is important for ensuring that the kinematics are not discontinuous or jarring to the subject controlling the prosthesis. The Kalman filter also casts BMI systems into a Bayesian framework, where it is now possible to model noise processes, effectively weighting neurons based on modeled noise properties. However, one potential limitation of Kalman filtering is that the state and observation noise processes are typically modeled to be Gaussian, which is an oversimplified assumption. We also note that the output model of the linear dynamical system (6) is inherently a *representational* approach, where the kinematics are generative of the neural data, as shown in Fig. 3(c).

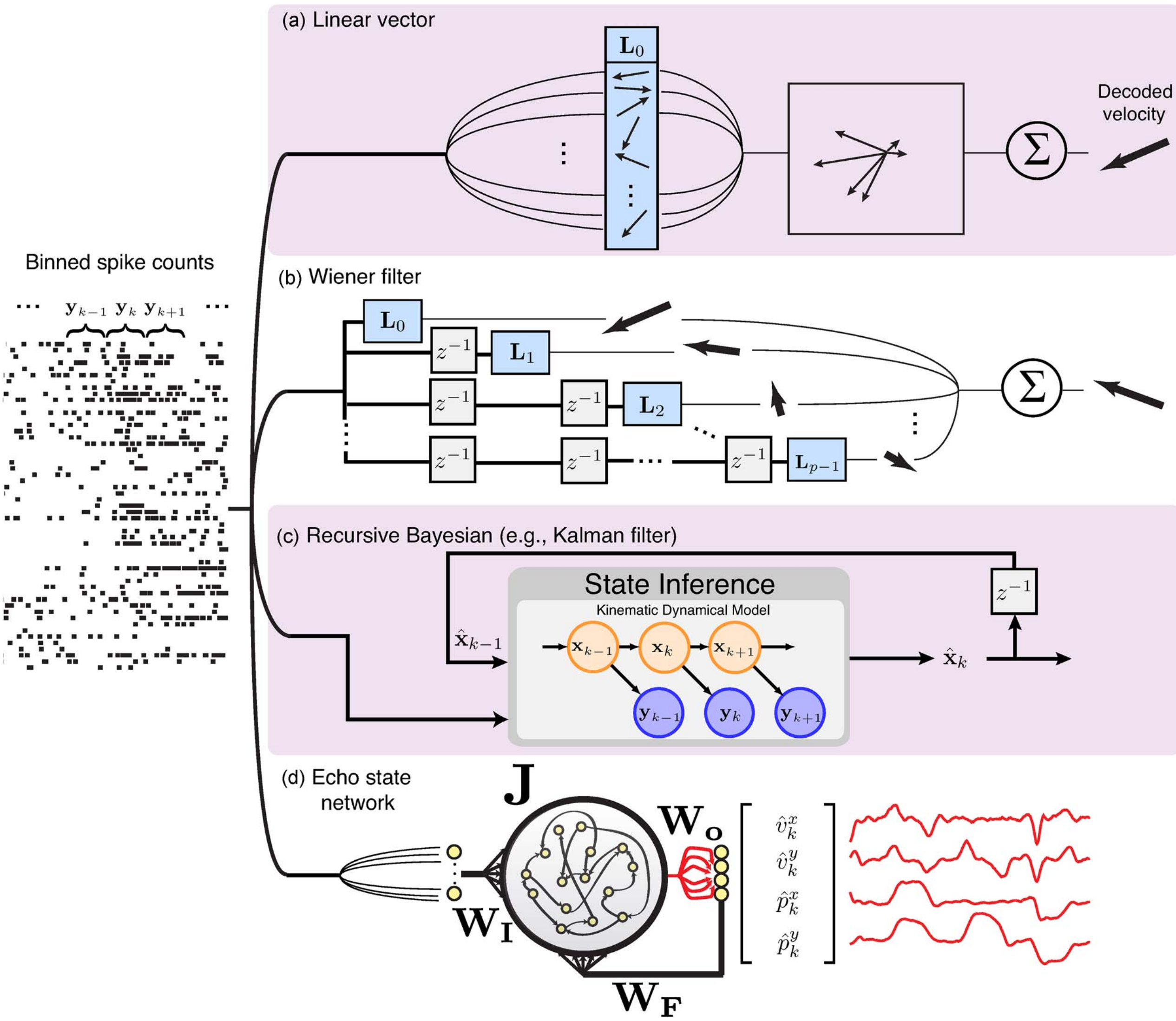


Fig. 3. ***Schematic of BMI algorithms. The spikes on N channels (left) are binned, and subsequently used by a decode algorithm. (a) The population vector/optimal linear estimator algorithm. The binned spike counts for each of N neurons are modeled to encode directions of movement in the workspace. The amount the neuron fires indicates the speed of movement along a certain direction. These vectors are then summed to give a final decoded velocity. (b) The Wiener filter algorithm. The neural activity (including neural history up to $p-1$ bins in the past) is weighted and then linearly summed to give decoded kinematics. The decoded kinematics need not be velocity, but could also be, for example, position. (c) Recursive Bayesian algorithms. The neural spiking activity is the output of a dynamical system model where the prosthesis kinematics are the underlying state. A recursive Bayesian algorithm (e.g., the Kalman filter) is used to infer the kinematic state $\hat{\mathbf{x}}_k$, given the current neural observation $\mathbf{y}_k$ and the previously predicted kinematic state $\hat{\mathbf{x}}_{k-1}$. (d) Echo state network algorithm. The neural spiking activity drives a recurrent, randomly connected, neural network. The kinematics $\hat{\mathbf{v}}_k = [\hat{v}_k^x\ \hat{v}_k^y]^\mathsf{T}$ and $\hat{\mathbf{p}}_k = [\hat{p}_k^x\ \hat{p}_k^y]^\mathsf{T}$ are decoded by linear readout and fed back into the network through the coefficients $\mathbf{W}_\mathbf{F}$.***

Kalman filters with time-invariant parameters, such as those used in BMI applications, converge to a steady-state form, typically in a matter of seconds [63]

$$\hat{\mathbf{x}}_k = \mathbf{M}_1\hat{\mathbf{x}}_{k-1} + \mathbf{M}_2\mathbf{y}_k \tag{7}$$

and, therefore, the Kalman filter can be interpreted analogously to the Wiener–Kolmogorov filtering approach. To make the correspondence, we note that the Kalman filter can be approximated in the form of (2) with certain structure: $\mathbf{L}_j^\mathsf{T} = \mathbf{M}_1^j\mathbf{M}_2$ for $j = 0, \ldots, p-1$. While such structure may be beneficial to the decoder, providing a form of regularization, it also imposes constraints that have important consequences. For example, in velocity Kalman filters (where $\mathbf{x}_k = \mathbf{v}_k$, e.g., [32]), matrix $\mathbf{M}_1$ is of the form $\mathbf{M}_1 \approx \alpha\mathbf{I}$ with $\alpha < 1$, since BMI training paradigms tend to sample kinematic velocities uniformly in all directions. Therefore, the velocity Kalman filter is analogous to a linear vector method with smoothing of past neural data, since each neuron will contribute velocities in approximately the same direction over all time. Kim *et al.* [32] demonstrated that a VKF performs superiorly to a Wiener filter with $p\Delta t = 1$ s. However, other studies have commented or shown that a Wiener filter outperforms Kalman filter techniques in offline simulations (e.g., [25]). Additionally, the Fitts throughput of BMI systems using Wiener filtering techniques tends to be higher than those of BMI systems using velocity Kalman filtering, as shown in Table 2.

D. Nonlinear Bayesian Algorithms

While linear vector, Wiener filter, and Kalman filter techniques have resulted in respectable performance, their modeling power and computational capacity are limited by their linearity. Neural computation is nonlinear, suggesting that BMI performance may be improved by using nonlinear decoding techniques. A benefit of using a Kalman filter approach is the incorporation of noise models and dynamical modeling, providing a Bayesian framework for BMI systems. However, while allowing the modeling of noise parameters, the linear dynamical system assumptions underlying Kalman filters may be oversimplified. For example, the output process of the linear dynamical system cannot model neural activity as a nonlinear function of the kinematics. To address this limitation, Li *et al.* implemented an unscented Kalman filter with a quadratic dynamical output process and demonstrated higher closed-loop performance than a Kalman filter [37]. Other studies have proposed particle filtering and point process based approaches (e.g., [33]–[36] and [65]) as well as Laplace–Gaussian filtering (e.g., [24] and [38]). Studies have reported that decoders using nonlinear Bayesian approaches achieved higher closed-loop performance than a population vector decoder (e.g., [24] and [45]).

E. Nonlinear Recurrent Neural Networks

One particular nonlinear modeling tool, the recurrent neural network (RNN), has seen much development over the last decade. In particular, the echo state network (ESN) [66] has seen wide spread application and has been investigated in both offline demonstrations (e.g., [67]) and closed-loop BMI systems (e.g., [39]). An ESN is an RNN with learning limited to the output weights. Specifically, the continuous-time ESN is defined by

$$\tau \dot{\mathbf{s}}_k = -\mathbf{s}_k + \mathbf{J}\mathbf{r}_k + \mathbf{W}_I\mathbf{y}_k + \mathbf{W}_F\hat{\mathbf{x}}_k \qquad (8)$$

where $\mathbf{s}_k$ is the hidden state of the recurrent network. The hidden units interact through matrix $\mathbf{J}$. The continuous variable $\mathbf{r}_k$ is the "instantaneous firing rate" and is defined as $r_k^i = \tanh(s_k^i)$. Interesting dynamics arise in the network due to this nonlinear coupling. The inputs $\mathbf{y}_k$ enter the system through weights $\mathbf{W}_I$ while a linear readout of the kinematics $\hat{\mathbf{x}}_k = \mathbf{W}_O\mathbf{r}_k$ is fed back to the hidden units through feedback weights $\mathbf{W}_F$.

Typically, training an RNN uses an algorithm called "backpropagation through time." Due to limitations in this algorithm [68], alternative network architectures and training methods have been developed to sidestep backpropagation through time. One such architecture is the ESN. The defining features of the ESN are a randomized $\mathbf{J}$ matrix and limited supervised training of only the output weights $\mathbf{W}_O$. Because the output is fed back to the hidden state, modifying the output weights additionally modifies the network dynamics, essentially driving a nonlinear spatio–temporal kernel with signal $\hat{\mathbf{x}}_k$ (along with input $\mathbf{y}_k$). Because learning is focused exclusively on $\mathbf{W}_O$, ESN training methods can be as simple as linear regression. Thus, the ESN architecture allows for powerful nonlinear modeling while sidestepping the full learning problem in RNNs.

We applied the ESN, as shown in Fig. 3(d), to the closed-loop BMI reaching task [39]. For the input, we used spiking activity (threshold crossings; see Section IV-B) from motor cortex, while for the training signal, we used the velocity and position of the reaching arm. We trained $\mathbf{W}_O$ with the FORCE learning rule [69]. We found that in a closed-loop BMI, the ESN performed over twice as well as a velocity Kalman filter across two test subjects [39]. Moreover, as shown in Table 2, the RNN is able to achieve higher Fitts throughput than linear vector techniques, Wiener filters, and velocity Kalman filters. Furthermore, we observed that the prosthesis kinematics decoded by the ESN were more like the hand kinematics than the prosthesis kinematics decoded by the velocity Kalman filter [39]. This study indicates that nonlinear RNNs merit further investigation as a viable BMI decode methodology.

III. EMERGING OPPORTUNITIES IN DECODER DESIGN

A design philosophy in machine learning and supervised classification is to design algorithms that capitalize on aspects or features specific to the system and data being analyzed. While the population vector algorithm was inspired from neurophysiological results, other techniques such as recursive Bayesian filtering are applied to BMI systems in a standard fashion without addressing unique aspects of BMI systems or motor cortical neural data. Hence, there is the potential to further increase BMI performance by augmenting decode algorithms with techniques that account for unique features in BMI systems.

We specifically focus on two opportunities where taking into account aspects of BMI systems and motor cortical neural data may further increase the performance of BMI systems. The first opportunity is to address the closed-loop nature of BMI systems: the subject controlling the BMI continuously observes how the prosthesis is moving. As a result, the user of the BMI may adopt novel strategies to control the prosthesis in addition to making online adjustments and corrections based on visual observations of the prosthesis movements. Hence, by augmenting decode algorithms with ideas from feedback control theory, it may be possible to increase the performance of BMIs, potentially irrespective of the specific type algorithm being used. The second opportunity is to incorporate recent neurophysiological evidence regarding the dynamical behavior of neural population ensembles. Designing decode algorithms that incorporate neuroscientific findings regarding motor cortex function has the potential to further increase BMI performance.

A. Decoder Retraining and Intention Estimation

As early as 1969, it was demonstrated that a monkey could modulate the activity of a particular neuron using operant conditioning (where the activity of the neuron is shown to the monkey, and a specific change in firing is rewarded) [70]. More recently, such neural adaptation has been demonstrated in BMI systems, where the properties of some neurons may change while the subject controls a closed-loop BMI (e.g., [27], [53], and [71]). For example, a recent study demonstrated that neurons which do not contribute to the decoder have a relative decrease in modulation compared to neurons that contribute to the decoder [72]. In these cases, the feedback component of the BMI, whereby the subject of the BMI observes how the prosthesis responds to user intention, leads to neural adaptation. This adaptation may reflect, for example, the adoption of a cognitive strategy to move the prosthesis. When coupled with the fact that the subject makes real-time adjustments to guide the prosthesis in a desired fashion, it is clear that the distribution of neural data during closed-loop BMI control is different than the distribution of the neural data used in training sets.

Because the distributions of neural data in the training set and closed-loop control are different, evaluating decoders on withheld "offline" data is not a reliable indicator of closed-loop performance (e.g., [23] and [24]). In addition to this, decoder parameter optimization through offline cross validation may result in parameters that are suboptimal for closed-loop control [22]. One method to better match the distributions between the training data and subsequent closed-loop BMI control data is to retrain the decoder with data from a closed-loop BMI control session. This approach, called decoder retraining (e.g., [73]–[75]), is a multistage process. In the first stage, a decoder is learned from training data with natural reaching or imagined movements. In the second stage, the learned decoder is used in closed-loop control, and these data subsequently serve as a training set for the learning of a new decoder (e.g., [10], [11], [13], [14], [27], and [73]–[75]). The second stage can be repeated, so that the decoder parameters continue to be updated based off of the most recently collected closed-loop BMI data. Decoder retraining in part accounts for potential neural adaptation that may result from being in the closed-loop control context (e.g., [53] and [72]–[75]) and was shown to decrease changes in neural preferred directions between the training set and closed-loop BMI control [73].

However, merely retraining the decoder with data from a closed-loop BMI session does not necessarily result in superior decoder performance. Indeed, we found that decoder retraining alone decreased decoder performance, in part because the closed-loop training set contains several instances where the prosthesis movements controlled by the decoder are discordant with the subject's intentions [73]. For example, when controlling an imperfect decoder, the subject may intend to move the prosthesis to the right, but the decoder instead moves the prosthesis to the left. This weakens the training set correlations between the neural data and the kinematics. Hence, an important technique to augment decoder retraining is *intention estimation*, which can help to improve the kinematic-neural data correlations [14], [73]. With intention estimation, the training set kinematics are modified to reflect the intent of the subject [14], [76]. As an example, our recent decoder, the "ReFIT–KF" algorithm, utilizes an intention estimation modification where it is assumed that the subject is always intending to move a computer cursor to the prompted target (goal) [14]. In this fashion, any observed training set velocities, which may even move the cursor away from the goal, are rotated to point toward the goal. We note that this modification is only performed on the training set, and no goal information is made available to the algorithm when used in closed-loop control. By combining decoder retraining and intention estimation, the performance of a BMI can be significantly increased [14]. Most of the performance improvement is a result of intention estimation rather than decoder retraining [73]. Recently, it was demonstrated that repeatedly updating decoder parameters with decoder retraining and intention estimation causes the performance of the decoder to increase until a steady-state convergence of decoder parameters [74], [75].

Another variant of this approach has the subject perform closed-loop BMI control with assistance, where the prosthesis movements are partially controlled by the decoder as well as by an automated controller that guides the prosthesis to the target (e.g., [11] and [13]). The amount of assistance provided is decreased as the subject learns to use the BMI well. Subsequently, these data constitute a new training set for decoder calibration [13].

B. A Feedback Control Intervention

Another innovation of the ReFIT–KF algorithm is to incorporate feedback control assumptions to model the visual feedback component of the BMI system. The ReFIT–KF algorithm makes the assumption that the user of the BMI observes and internalizes the decoded position

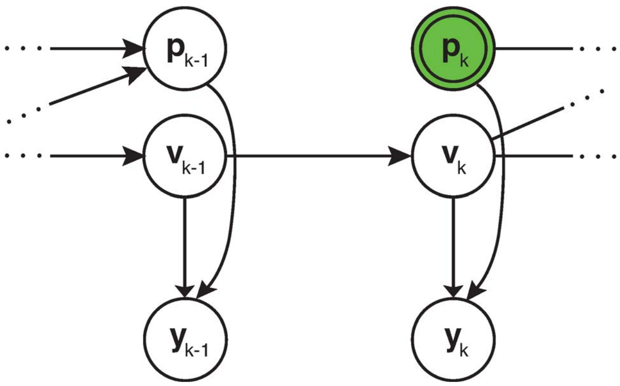


Fig. 4. ***Feedback control causal intervention. This is the linear dynamical model used in Kalman filter decoders, where x_k encompasses both position p_k and velocity v_k of the prosthesis. The position at time k is given by $p_k = p_{k-1} + v_{k-1}\Delta t$. Therefore, uncertainty in both p_{k-1} and v_{k-1} should propagate to uncertainty in p_k. However, because the subject observes the position of the prosthesis at time k (as indicated by the green shading), an assumption is made that the subject has no uncertainty in position p_k. Therefore, the uncertainty from p_{k-1} and v_{k-1} is not propagated to p_k. Figure from [14].***

of the cursor with complete certainty. Therefore, any uncertainty in the decoded position, which would otherwise arise from propagated uncertainty in the decoded velocity, is set to zero. This is demonstrated in the graphical model of Fig. 4. The position at time k is observed by the user, as set by the decoder (which is called a *causal intervention* [77], highlighted in green) and incoming arrows to $\mathbf{p}_k$ are removed, indicating that no uncertainty is propagated to $\mathbf{p}_k$. The ReFIT–KF algorithm also estimates the contribution of position to the neural activity by finding the matrix $\mathbf{C}_p$ that minimizes the squared error $\mathbf{y}_k - \mathbf{C}_\mathrm{p}\mathbf{p}_k$. Given the assumption that there is no uncertainty in the decoded position $\hat{\mathbf{p}}_k$, the ReFIT–KF algorithm subtracts the position contribution to the neural signal by calculating $\tilde{\mathbf{y}}_k = \mathbf{y}_k - \mathbf{C}_\mathrm{p}\hat{\mathbf{p}}_k$. Subsequently, the position subtracted neural data $\tilde{\mathbf{y}}_k$ are used as a neural observation of the Kalman filter. Combined with decoder retraining and intention estimation, the ReFIT–KF algorithm increased the performance of state-of-the-art BMI systems by approximately twofold [14].

Feedback control approaches can increase the performance of BMI systems while being somewhat agnostic to the type of algorithm being used. For example, intention estimation modifications and decoder retraining can be applied to most decoders. Therefore, developing techniques that account for the feedback aspect of BMI systems may further increase the performance of BMI systems.

C. Future BMIs: A Neural Dynamical Perspective

Whereas the linear vector techniques described in Section II-A were informed by a neuroscientific perspective, much of recent BMI algorithm development has relied on linear estimation and neural network theory. We therefore ask: What is the place of neuroscience in decoder design?

Over the past decade, a line of scientific evidence has proposed a *dynamical* perspective of motor cortex (e.g., [51], [52], and [78]–[80] and reviewed in [42]). In this perspective, motor cortex is described as a dynamical machine that generates movements. A key component of this theory is that the neural population activity at time k is informative of the neural population activity at time $k+1$. This is captured by introducing a "neural state." The neural state, which can be inferred from observations of motor cortical activity, summarizes the neural population activity, and is governed by a dynamical model that describes the neural state at time $k+1$ as a function of the neural state at time k. Several studies have investigated the characteristics of these dynamics (e.g., [52], [79], and [81]), while other studies have demonstrated that the trajectories of the neural state are informative of behavioral correlates (e.g., [78] and [82]). At the crux of the dynamical perspective is a departure from modeling single neuron tuning to modeling population-level neural interactions and dynamics.

Current techniques in the BMI literature do not incorporate dynamical models of the neural population activity. For example, the Kalman filter incorporates a dynamical model, but it is only a model of the physical kinematic laws of the prosthesis (resulting in temporal smoothing of the kinematics) which are learned without neural data (e.g., [10], [14], [31], and [32]). While these models are able to capture how neural activity is *externally driven* by kinematic activity, they do not capture how the neural activity has its own *internal drive*, with rules that govern how the neural population modulates itself over time. If one can learn an adequate dynamical model of the neural population activity, modeling this temporal structure has the potential to increase BMI performance. One reason to expect improvement is because a prediction of future neural population activity (obtained through a dynamical model) can be used to augment noisy observations of the neural activity. Our recent study demonstrates that modeling even a simple linear time-invariant approximation of the neural dynamics can significantly increase the performance of a BMI system [83]. Therefore, incorporating ideas from recent studies of neural dynamics may be important for enabling next-generation, high-performance BMI systems.

IV. TOWARD CLINICAL TRANSLATION

The ultimate goal of BMI systems is to improve the quality of life for people with paralysis. To this end, many of the design choices in BMI systems are guided by a motivation to increase the clinical viability of BMI systems. While increasing decoder performance is essential to clinical translation, other important challenges remain that will be essential for bringing BMI systems to clinical viability. In this section, we describe a brief history of BMI clinical translation, and discuss three additional opportunities in BMI systems that may be of interest to information systems engineers.

A. A Brief History of Clinical Translation

Until recently, progress in BMI technology has largely come from advances in statistical signal processing and motor neuroscience based on preclinical nonhuman primate studies [21]. However, the field's driving motivation has always been to move toward creating clinically viable prostheses to restore movement to people with paralysis. The first intracortical BMI tested in a person consisted of just two chronically implanted electrodes, and gave the subject basic control of a computer cursor following extensive training [84], [85]. In 2004, the first participant in the BrainGate FDA phase-I clinical trial was implanted with a 96-electrode Utah array, similar to the one used in many previous monkey studies. This study provided critical evidence that movement intention-related signals persist in motor cortex even many years after paralysis-causing injury. Today there are multiple ongoing clinical trials of investigatory closed-loop intracortical BMI systems, with compelling demonstrations of individuals with tetraplegia using these devices to more accurately control a computer cursor [9], [32], [86]–[88] and use a robotic arm to manipulate objects in their environment [10], [11]. While there remain a number of challenges to be solved on the way to clinical translation, here we will focus on recent progress made toward increasing the longevity of BMI system use as well as low power implementations of BMI systems.

B. Maintaining BMI Performance in the Face of Signal Loss

A particularly pressing challenge is to develop neural prostheses that will sustain high performance for many years after device implantation, with the ultimate goal being lifetime functionality [16], [89]. The number of discriminable neurons recorded by an implanted array degrades over time (e.g., [86] and [90]–[93]) due to several factors. These factors include biological failures such as gliosis and meningitis, material failures such as insulation leakage, and electrode mechanical failure (e.g., [92] and [93]). The risks and costs of sensor reimplantation in a human patient are quite real; thus, their usable lifespan must be maximized.

Throughout this review, we have referred to the neural observations of BMIs being the spikes of individual neurons measured from electrodes. However, the electrical voltages recorded on these electrodes tend to attenuate over time, making the detection of spiking activity from individual neurons more difficult. Recent studies have demonstrated that one way of maintaining performance in the face of degrading electrode recording quality is to measure activity derived from multiunit spiking threshold crossings, rather than single neuron spiking activity. These threshold crossing events are measured by counting the number of times the voltage on an electrode falls below a predetermined value (e.g., some multiple of the root mean square voltage on the channel) regardless of whether the activity is from a single neuron. For example, one could measure action potentials from two neurons on a single electrode, as shown in Fig. 2(b) and (c), but the threshold crossing observation would not differentiate between spikes coming from one neuron or the other. A concern in using threshold crossings is the loss of information incurred from not separating out distinct neural sources; one could easily imagine the deleterious effect of combining activity from two neurons with opposite tuning. However, previous studies have demonstrated that these effects do not significantly decrease BMI performance. Indeed, the performance of BMIs using threshold crossings is comparable to those using single unit activity [45], [91]. Moreover, a potential loss in performance is outweighed by the ability of threshold crossings to mitigate decoder performance drop-off in the presence of decreasing signal-to-noise ratio. It was reported that as long as the neural signal is above the noise floor, BMI performance based on threshold crossing activity is largely independent of action potential voltage [91]. As a manifestation of this result, several studies have demonstrated that BMIs which decode threshold crossing activity can perform at a consistently high level even when the arrays are multiple years post implantation [14], [86]. This approach has successfully translated to human clinical studies (e.g., [10] and [12]). An example plot of performance of the ReFIT–KF algorithm, retrained daily using threshold crossings, is shown in Fig. 5. For multiple years, across two macaques, the decoder performance remained approximately constant. By using threshold crossings, the neural signals measured from these arrays continue to result in comparable BMI performance to this day, surpassing the six-year mark in Monkey L.

A different and complimentary method to maintain performance in the face of degrading electrode recording quality is to make use of additional types of neural signals that may be available from the same sensor. As shown in Fig. 2, spiking activity is extracted by high-pass filtering the raw voltage signal coming off of the electrode, but there is also information in the low-frequency component of the signal (up to several hundred hertz). This signal is called the local field potential and is the superposition of extracellular electrical currents resulting from action potentials, postsynaptic potentials, and other membrane currents of cells in the vicinity of the recording electrode (e.g., [94]–[97]). Several reports have conjectured that LFP may be more stable than spikes over time (e.g., [98] and [99]) and have shown that LFP can be measured even in the absence of spikes (e.g., [100]). A number of studies

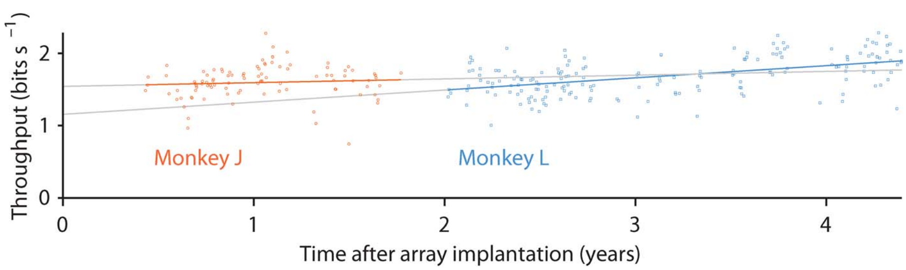


Fig. 5. *Robustness of threshold crossings across years. The performance, measured as Fitts throughput (bits per second), is shown for the ReFIT–KF algorithm using threshold crossings across years. Each dot corresponds to the performance as measured on an experimental day, and the different colors correspond to different subjects. Each day, the ReFIT–KF was trained using the recorded neural threshold crossings measured on that day. There is no observed decline in performance of the ReFIT–KF algorithm across approximately four years, as indicated by the regression lines. Figure from [14].*

have examined the LFP recorded during natural reaches and shown that this signal contains considerable information about the movement (e.g., [98] and [101]–[109]).

Only very recently have these offline studies been followed up by closed-loop demonstrations with macaques controlling a BMI using LFP signals [75], [110]. Because continuously controlling a BMI cursor using LFP signals is only in its infancy, there may be room to improve LFP decoding performance. For example, a fundamental difference between LFP and spiking signals is that LFP is an analog signal and is thus subject to a variety of analysis techniques. While point-process spiking activity is typically preprocessed to form an estimate of the firing rate through binning or smoothing, there are a myriad of choices of possible time-domain and frequency-domain LFP features that can be derived from the raw recorded LFP voltage, as shown in Fig. 2. Hence, algorithmic investigations (as discussed in Section II) may have to be revisited to delineate how various aspects of the LFP, such as different time- and frequency-domain features, can result in effective high-performance decoders. While one study has demonstrated that decoding from threshold crossings leads to better performance than decoding from various LFP features [110], it remains to be seen whether incorporating the LFP signal into a spike-based BMI can improve performance or robustness over a system driven solely by spiking activity. Importantly, developing decoders that beneficially combine these signals would be a major step toward increasing the clinical viability of BMI systems.

C. Decoder Longevity Without Retraining

Another important challenge facing BMI systems is that recorded neural signals can be nonstationary from one day to the next, so that a decoder trained on a previous day may not be effective on a subsequent day. While current clinical studies have been instrumental in demonstrating the capabilities of BMI systems, they have always required constant expertise and supervision by trained technicians. In particular, the technicians, among other tasks, must daily collect a training set, which is used to subsequently train a decode algorithm. However, to facilitate widespread BMI use, it would be useful for BMI systems to be autonomous, capable of running for days without recalibration or technician supervision.

To this end, some recent work has been devoted to building robust decoders that are capable of being used for multiple weeks without the need for retraining sessions or recalibration by a technician. Encouragingly, it has been shown that LFP and threshold crossing activity provide a level of robustness for decoding not previously afforded by single units [55], [110]. In these studies, a static decoder was used for extended periods of time (up to a year) without retraining. These studies demonstrate a stabilizing of the relationship between neural activity and cursor movement during online BMI control [110] and suggest that BMIs may be capable of robust performance over long timescales. However, more extensive experimentation is warranted over longer periods of time and with more subjects.

D. Low-Power Neuromorphic Implementations

Clinical and laboratory studies currently require multiple computers and recording systems to function, which result in bulky and significant hardware infrastructure. However, for clinical use, BMI systems should be portable, low power, and ideally implantable without significant burden on the subject. The two approaches to this problem are differentiated based on where the neural decode occurs: remotely or locally. In remote decode systems, the neural data is measured, amplified, and optionally thresholded before being transmitted wirelessly to a receiver which performs the rest of the processing and decoding [111]–[113]. An alternative approach is to perform the decode locally, with transmission of only the low data rate kinematic information. Conventional digital hardware systems, including ASICS, may still require too much power, and could lead to excessive heating of the brain and surrounding tissue. The limit for power dissipation set by the American Association of Medical Instrumentation is 10 mW within a 6×6 mm^2 area [114], [115]. This power constraint may be met with a *neuromorphic* approach [116], which uses analog hardware modeling neural architectures to perform computations. The advantage of this approach is that the decode could be performed locally on a neuromorphic chip with no need for broadband neural data transmission, cutting down the wireless data rate by approximately four orders of magnitude to 3 kb/s [17]. In this approach, standard algorithms, such as the Kalman filter, are translated into spiking neural network implementations. For example, a recent study shows that a 2000-artificial neuron spiking neural network can adequately mimic a Kalman filter decoder in closed-loop BMI control [17], [117]. In neuromorphic chip implementations, an artificial neuron spiking at 100 Hz dissipates approximately 50 nW of power, which could lead to significantly lower power implementations that may potentially be fully implantable. Fully implantable chips (e.g., [118] and [119]) may be important for reducing infection risks and mechanical forces that may cause electronic and array failure. Combined with the neuromorphic approach, BMI systems may be safely implanted while drawing very little power, making them more accessible for clinical use.

V. CONCLUSION

In the past 15 years, great strides have been made to bring intracortical BMI systems from concept, to laboratory implementation, and finally to clinical studies. In particular, information systems engineering has played a significant role in the design of decode algorithms, which are at the core of BMI systems. These designs have resulted in compelling laboratory and clinical studies, and continue to march us toward the goal of bringing BMIs to clinical viability. As reviewed here, there are numerous information systems engineering challenges and opportunities that will be important to achieving this goal. ■

Acknowledgment

The authors would like to thank M. Mazariegos, J. Aguayo, M. Wechsler, C. Sherman, E. Morgan, and L. Yates for expert surgical assistance and veterinary care; B. Oskotsky for IT support; and B. Davis, E. Casteneda, and S. Eisensee for administrative assistance.

REFERENCES

[1] K. D. Anderson, "Targeting recovery: Priorities of the spinal cord-injured population," *J. Neurotrama*, vol. 21, pp. 1371–1383, 2004.

[2] Christopher & Dana Reeve Foundation, "One degree of separation: Paralysis and spinal cord injury in the United States," 2009.

[3] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, and D. W. Moran, "A brain-computer interface using electrocorticographic signals in humans," *J. Neural Eng.*, vol. 1, no. 2, pp. 63–71, Jun. 2004.

[4] E. C. Leuthardt, G. Schalk, J. Roland, A. Rouse, and D. W. Moran, "Evolution of brain-computer interfaces: Going beyond classic motor physiology," *Neurosurgical Focus*, vol. 27, no. 1, Jul. 2009, DOI: 10.3171/2009.4.FOCUS0979.

[5] D. W. Moran, "Evolution of brain-computer interface: Action potentials, local field potentials and electrocorticograms," *Current Opinion Neurobiol.*, vol. 20, no. 6, pp. 741–745, Dec. 2010.

[6] G. Schalk, J. Kubánek, K. J. Miller, N. R. Anderson, E. C. Leuthardt, J. G. Ojemann, D. Limbrick, D. W. Moran, L. A. Gerhardt, and J. R. Wolpaw, "Decoding two-dimensional movement trajectories using electrocorticographic signals in humans," *J. Neural Eng.*, vol. 4, no. 3, pp. 264–275, Sep. 2007.

[7] G. Schalk, K. J. Miller, N. R. Anderson, J. A. Wilson, M. D. Smyth, J. G. Ojemann, D. W. Moran, J. R. Wolpaw, and E. C. Leuthardt, "Two-dimensional movement control using electrocorticographic signals in humans," *J. Neural Eng.*, vol. 5, no. 1, pp. 75–84, Mar. 2008.

[8] Z. Wang, A. Gunduz, P. Brunner, A. L. Ritaccio, Q. Ji, and G. Schalk, "Decoding onset and direction of movements using electrocorticographic (ECoG) signals in humans," *Front. Neuroeng.*, vol. 5, p. 15, Jan. 2012, DOI: 10.3389/fneng.2012.00015.

[9] L. R. Hochberg, M. D. Serruya, G. M. Friehs, J. A. Mukand, M. Saleh, A. H. Caplan, A. Branner, D. Chen, R. D. Penn, and J. P. Donoghue, "Neuronal ensemble control of prosthetic devices by a human with tetraplegia," *Nature*, vol. 442, no. 7099, pp. 164–171, Jul. 2006.

[10] L. R. Hochberg, D. Bacher, B. Jarosiewicz, N. Y. Masse, J. D. Simeral, J. Vogel, S. Haddadin, J. Liu, S. S. Cash, P. van der Smagt, and J. P. Donoghue, "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm," *Nature*, vol. 485, no. 7398, pp. 372–375, May 2012.

[11] J. L. Collinger, B. Wodlinger, J. E. Downey, W. Wang, E. C. Tyler-Kabara, D. J. Weber, A. J. C. McMorland, M. Velliste, M. L. Boninger, and A. B. Schwartz, "High-performance neuroprosthetic control by an individual with tetraplegia," *Lancet*, vol. 381, no. 9866, pp. 557–564, Feb. 2013.

[12] B. Jarosiewicz, N. Y. Masse, D. Bacher, S. S. Cash, E. Eskandar, G. Friehs, J. P. Donoghue, and L. R. Hochberg, "Advantages of closed-loop calibration in intracortical brain-computer interfaces for people with tetraplegia," *J. Neural Eng.*, vol. 10, no. 4, Aug. 2013, 046012.

[13] M. Velliste, S. Perel, M. C. Spalding, A. S. Whitford, and A. B. Schwartz, "Cortical control of a prosthetic arm for self-feeding," *Nature*, vol. 453, no. 7198, pp. 1098–1101, Jun. 2008.

[14] V. Gilja, P. Nuyujukian, C. A. Chestek, J. P. Cunningham, B. M. Yu, J. M. Fan, M. M. Churchland, M. T. Kaufman, J. C. Kao, S. I. Ryu, and K. V. Shenoy, "A high-performance neural prosthesis enabled by control algorithm design," *Nature Neurosci.*, vol. 15, no. 12, pp. 7–10, Nov. 2012.

[15] J. E. O'Doherty, M. A. Lebedev, P. J. Ifft, K. Z. Zhuang, S. Shokur, H. Bleuler, and M. A. L. Nicolelis, "Active tactile exploration using a brain-machine-brain interface," *Nature*, vol. 479, no. 7372, pp. 228–231, Nov. 2011.

[16] V. Gilja, C. A. Chestek, I. Diester, J. M. Henderson, and K. V. Shenoy, "Challenges and opportunities for next-generation intracortically based neural prostheses," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 7, pp. 1891–1899, Jul. 2011.

[17] J. Dethier, P. Nuyujukian, S. I. Ryu, K. V. Shenoy, and K. Boahen, "Design and validation of a real-time spiking-neural-network decoder for brain-machine interfaces," *J. Neural Eng.*, vol. 10, no. 3, Apr. 2013, 036008.

[18] R. E. Kass, V. Ventura, and E. N. Brown, "Statistical issues in the analysis of neuronal data," *J. Neurophysiol.*, vol. 94, pp. 8–25, 2005.

[19] A. M. Green and J. F. Kalaska, "Learning to move machines with the mind," *Trends Neurosci.*, vol. 34, no. 2, pp. 61–75, 2011.

[20] M. L. Homer, A. V. Nurmikko, J. P. Donoghue, and L. R. Hochberg, "Sensors and decoding for intracortical brain computer interfaces," *Annu. Rev. Biomed. Eng.*, vol. 15, pp. 383–405, Jan. 2013.

[21] P. Nuyujukian, J. M. Fan, V. Gilja, P. S. Kalanithi, C. A. Chestek, and K. V. Shenoy, "Monkey models for brain-machine interfaces: The need for maintaining diversity," in *Proc. 33rd Annu.*

Conf. IEEE Eng. Med. Biol. Soc., Jan. 2011, vol. 2011, pp. 1301–1305.

[22] J. P. Cunningham, P. Nuyujukian, V. Gilja, C. A. Chestek, S. I. Ryu, and K. V. Shenoy, "A closed-loop human simulator for investigating the role of feedback control in brain-machine interfaces," *J. Neurophysiol.*, vol. 105, pp. 1932–1949, 2011.

[23] S. M. Chase, A. B. Schwartz, and R. E. Kass, "Bias, optimal linear estimation, and the differences between open-loop simulation and closed-loop performance of spiking-based brain-computer interface algorithms," *Neural Netw.*, vol. 22, no. 9, pp. 1203–1213, 2009.

[24] S. Koyama, S. M. Chase, A. S. Whitford, M. Velliste, A. B. Schwartz, and R. E. Kass, "Comparison of brain-computer interface decoding algorithms in open-loop and closed-loop control," *J. Comput. Neurosci.*, vol. 29, no. 1–2, pp. 73–87, Aug. 2010.

[25] J. M. Carmena, M. A. Lebedev, R. E. Crist, J. E. O'Doherty, D. M. Santucci, D. F. Dimitrov, P. G. Patil, C. S. Henriquez, and M. A. L. Nicolelis, "Learning to control a brain-machine interface for reaching and grasping by primates," *PLoS Biol.*, vol. 1, no. 2, Nov. 2003, E42.

[26] E. Salinas and L. F. Abbott, "Vector reconstruction from firing rates," *J. Comput. Neurosci.*, vol. 1, no. 1–2, pp. 89–107, Jun. 1994.

[27] D. M. Taylor, S. I. H. Tillery, and A. B. Schwartz, "Direct cortical control of 3D neuroprosthetic devices," *Science*, vol. 296, no. 5574, pp. 1829–1832, Jun. 2002.

[28] N. G. Hatsopoulos, J. Joshi, and J. G. O'Leary, "Decoding continuous and discrete motor behaviors using motor and premotor cortical ensembles," *J. Neurophysiol.*, vol. 92, no. 2, pp. 1165–1174, Aug. 2004.

[29] J. C. Sanchez, D. Erdogmus, Y. N. Rao, S.-P. Kim, M. A. L. Nicolelis, J. Wessberg, and J. C. Principe, "Interpreting neural activity through linear and nonlinear models for brain machine interfaces," in *Proc. 25th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2003, no. 2, pp. 2160–2163.

[30] J. C. Sanchez, J. C. Principe, J. M. Carmena, M. A. Lebedev, and M. A. L. Nicolelis, "Simultaneous prediction of four kinematic variables for a brain-machine interface using a single recurrent neural network," in *Proc. 26th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2004, pp. 5321–5324.

[31] W. Wu, M. J. Black, Y. Gao, E. Beinenstock, M. D. Serruya, A. Shaikhouni, and J. P. Donoghue, "Neural decoding of cursor motion using a Kalman filter," *Advances in Neural Information and Processing Systems 15*. Cambridge, MA, USA: MIT Press, 2003, pp. 133–140.

[32] S.-P. Kim, J. D. Simeral, L. R. Hochberg, J. P. Donoghue, and M. J. Black, "Neural control of computer cursor velocity by decoding motor cortical spiking activity in humans with tetraplegia," *J. Neural Eng.*, vol. 5, no. 4, pp. 455–476, Dec. 2008.

[33] A. E. Brockwell, A. L. Rojas, and R. E. Kass, "Recursive Bayesian decoding of motor cortical signals by particle filtering," *J. Neurophysiol.*, vol. 91, no. 4, pp. 1899–1907, Apr. 2004.

[34] Y. Gao, M. J. Black, E. Bienenstock, S. Shoham, and J. P. Donoghue, "Probabilistic inference of hand motion from neural activity in motor cortex," *Advances in Neural Information and Processing Systems 14*. Cambridge, MA, USA: MIT Press, 2002, pp. 213–220.

[35] Y. Gao, M. J. Black, E. Bienenstock, W. Wu, and J. P. Donoghue, "A quantitative comparison of linear and non-linear models of motor cortical activity for the encoding and decoding of arm motions," in *Proc. 1st Int. IEEE EMBS Conf. Neural Eng.*, 2003, pp. 189–192.

[36] S. Shoham, L. M. Paninski, M. R. Fellows, N. G. Hatsopoulos, J. P. Donoghue, and R. A. Normann, "Statistical encoding model for a primary motor cortical brain-machine interface," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 7, pp. 1312–1322, Jul. 2005.

[37] Z. Li, J. E. O'Doherty, T. L. Hanson, M. A. Lebedev, C. S. Henriquez, and M. A. L. Nicolelis, "Unscented Kalman filter for brain-machine interfaces," *PLoS ONE*, vol. 4, no. 7, Jan. 2009, e6243.

[38] S. Koyama, L. C. Pérez-Bolde, C. R. Shalizi, and R. E. Kass, "Approximate methods for state-space models," *J. Amer. Stat. Assoc.*, vol. 105, no. 489, pp. 170–180, Mar. 2010.

[39] D. Sussillo, P. Nuyujukian, J. M. Fan, J. C. Kao, S. D. Stavisky, S. I. Ryu, and K. V. Shenoy, "A recurrent neural network for closed-loop intracortical brain-machine interface decoders," *J. Neural Eng.*, vol. 9, no. 2, Apr. 2012, 026027.

[40] A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner, "Neuronal population coding of movement direction," *Science*, vol. 233, no. 4771, pp. 1416–1419, Sep. 1986.

[41] R. Wahnoun, J. He, and S. I. Helms Tillery, "Selection and parameterization of cortical neurons for neuroprosthetic control," *J. Neural Eng.*, vol. 3, no. 2, pp. 162–171, Jun. 2006.

[42] K. V. Shenoy, M. Sahani, and M. M. Churchland, "Cortical control of arm movements: A dynamical systems perspective," *Annu. Rev. Neurosci.*, vol. 36, pp. 337–359, Jul. 2013.

[43] A. P. Georgopoulos, J. F. Kalaska, R. Caminiti, and J. T. Massey, "On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex," *J. Neurosci.*, vol. 2, no. 11, pp. 1527–1537, Nov. 1982.

[44] W. Wang, S. S. Chan, D. A. Heldman, and D. W. Moran, "Motor cortical representation of position and velocity during reaching," *J. Neurophysiol.*, vol. 97, no. 6, pp. 4258–4270, Jun. 2007.

[45] G. W. Fraser, S. M. Chase, A. S. Whitford, and A. B. Schwartz, "Control of a brain-computer interface without spike sorting," *J. Neural Eng.*, vol. 6, no. 5, Oct. 2009, DOI: 10.1088/1741-2560/6/5/055004.

[46] S. K. Card, W. K. English, and B. J. Burr, "Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT," *Ergonomics*, vol. 21, no. 8, pp. 601–613, 1978.

[47] J. P. Donoghue, J. D. Simeral, S.-P. Kim, G. M. Friehs, L. R. Hochberg, and M. J. Black, "Toward standardized assessment of pointing devices for brain-computer interfaces," presented at the Proc. Soc. Neurosci., San Diego, CA, USA, 2007.

[48] M. M. Churchland, G. Santhanam, and K. V. Shenoy, "Preparatory activity in premotor and motor cortex reflects the speed of the upcoming reach," *J. Neurophysiol.*, vol. 96, pp. 3130–3146, 2006.

[49] M. M. Churchland and K. V. Shenoy, "Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex," *J. Neurophysiol.*, vol. 97, pp. 4235–4257, 2007.

[50] W. Truccolo, G. M. Friehs, J. P. Donoghue, and L. R. Hochberg, "Primary motor cortex tuning to intended movement kinematics in humans with tetraplegia," *J. Neurosci.*, vol. 28, no. 5, pp. 1163–1178, Jan. 2008.

[51] K. V. Shenoy, M. T. Kaufman, M. Sahani, and M. M. Churchland, "A dynamical systems view of motor preparation: Implications for neural prosthetic system design," *Progr. Brain Res.*, vol. 192, pp. 33–58, Jan. 2011.

[52] M. M. Churchland, J. P. Cunningham, M. T. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy, "Neural population dynamics during reaching," *Nature*, vol. 487, no. 7405, pp. 51–56, Jul. 2012.

[53] K. Ganguly, L. Secundo, G. Ranade, A. L. Orsborn, E. F. Chang, D. F. Dimitrov, J. D. Wallis, N. M. Barbaro, R. T. Knight, and J. M. Carmena, "Cortical representation of ipsilateral arm movements in monkey and man," *J. Neurosci.*, vol. 29, no. 41, pp. 12 948–12 956, 2009.

[54] G. Santhanam, S. I. Ryu, B. M. Yu, A. Afshar, and K. V. Shenoy, "A high-performance brain-computer interface," *Nature*, vol. 442, pp. 195–198, Jul. 2006.

[55] P. Nuyujukian, J. C. Kao, J. M. Fan, S. D. Stavisky, S. I. Ryu, and K. V. Shenoy, "A high-performance, robust brain-machine interface without retraining," presented at the Front. Neurosci., Comput. Systems Neurosci., Salt Lake City, UT, USA, pp. 190–191, 2012.

[56] H. W. Sorenson, "Least-squares estimation: From Gauss to Kalman," *IEEE Spectrum*, vol. 7, no. 7, pp. 63–68, Jul. 1970.

[57] T. Kailath, "A view of three decades of linear filtering theory," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 146–181, Mar. 1974.

[58] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. New York, NY, USA: Technology Press/Wiley, 1949.

[59] A. N. Kolmogorov, "Sur l'interpolation et extrapolation des suites stationnaires," *Comptes Rendus de l'Academie des Sciences*, vol. 208, pp. 2043–2045, 1939.

[60] S. V. Vaseghi, "Least square error Wiener-Kolmogorov filters," in *Advanced Digital Signal Processing and Noise Reduction*, 4th ed. New York, NY, USA: Wiley, 2008, pp. 173–191, no. 1941.

[61] R. E. Kalman, "New methods of Wiener filtering theory," in *Proc. 1st Symp. Eng. Appl. Random Funct. Theory Probab.*, 1963, pp. 279–388.

[62] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2000.

[63] W. Q. Malik, W. Truccolo, E. N. Brown, and L. R. Hochberg, "Efficient decoding with steady-state Kalman filter in neural interface systems," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 19, no. 1, pp. 25–34, Feb. 2011.

[64] R. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME/J. Basic Eng.*, vol. 82, no. Series D, pp. 35–45, 1960.

[65] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, "A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects," *J. Neurophysiol.*, vol. 93, no. 2, pp. 1074–1089, Feb. 2005.

[66] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication,"

Science, vol. 304, no. 5667, pp. 78–80, Apr. 2004.

[67] Y. N. Rao, S.-P. Kim, J. C. Sanchez, D. Erdogmus, J. C. Principe, J. M. Carmena, M. A. Lebedev, and M. A. L. Nicolelis, "Learning mappings in brain machine interfaces with echo state networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2005, pp. 233–236.

[68] H. Jaeger, "A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the 'echo state network' approach," German Nat. Res. Ctr. Inf. Technol., GMD Rep. 159, 2002.

[69] D. Sussillo and L. F. Abbott, "Generating coherent patterns of activity from chaotic neural networks," *Neuron*, vol. 63, no. 4, pp. 544–557, 2009.

[70] E. E. Fetz, "Operant conditioning of cortical unit activity," *Science*, vol. 163, pp. 955–958, Feb. 1969.

[71] B. Jarosiewicz, S. M. Chase, G. W. Fraser, M. Velliste, R. E. Kass, and A. B. Schwartz, "Functional network reorganization during learning in a brain-computer interface paradigm," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 49, pp. 19486–19491, Dec. 2008.

[72] K. Ganguly, D. F. Dimitrov, J. D. Wallis, and J. M. Carmena, "Reversible large-scale modification of cortical networks during neuroprosthetic control," *Nature Neurosci.*, vol. 14, no. 5, pp. 662–667, May 2011.

[73] J. M. Fan, P. Nuyujukian, J. C. Kao, C. A. Chestek, S. I. Ryu, and K. V. Shenoy, "Intention estimation in brain machine interfaces," *J. Neuroeng.*, vol. 11, no. 1, 2014, 016004.

[74] A. L. Orsborn, S. Dangi, H. G. Moorman, and J. M. Carmena, "Closed-loop decoder adaptation on intermediate time-scales facilitates rapid BMI performance improvements independent of decoder initialization conditions," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 4, pp. 468–477, Jul. 2012.

[75] K. So, S. Dangi, A. L. Orsborn, M. C. Gastpar, and J. M. Carmena, "Subject-specific modulation of local field potential spectral power during brain-machine interface control in primates," *J. Neural Eng.*, vol. 11, no. 2, Feb. 2014, 026002.

[76] L. Shpigelman, H. Lalazar, and E. Vaadia, "Kernel-ARMA for hand tracking and brain-machine interfacing during 3D motor control," *Advances in Neural Information Processing Systems 21*. Cambridge, MA, USA: MIT Press, 2009, pp. 1489–1496.

[77] D. Koller and N. Friedman, *Probabilistic Graphical Models*. Cambridge, MA, USA: MIT Press, 2009.

[78] A. Afshar, G. Santhanam, B. M. Yu, S. I. Ryu, M. Sahani, and K. V. Shenoy, "Single-trial neural correlates of arm movement preparation," *Neuron*, vol. 71, no. 3, pp. 555–564, Aug. 2011.

[79] B. Petreska, B. M. Yu, J. P. Cunningham, G. Santhanam, S. I. Ryu, K. V. Shenoy, and M. Sahani, "Dynamical segmentation of single trials from population neural data," *Advances in Neural Information Processing Systems 24*. Cambridge, MA, USA: MIT Press, 2011, pp. 756–764.

[80] L. Buesing, J. H. Macke, and M. Sahani, "Learning stable, regularised latent models of neural population dynamics," *Network, Comput. Neural Syst.*, vol. 23, no. 1–2, pp. 24–47, Jan. 2012.

[81] K. C. Ames, S. I. Ryu, and K. V. Shenoy, "Neural dynamics of reaching following incorrect or absent motor preparation," *Neuron*, vol. 81, no. 2, pp. 438–451, Jan. 2014.

[82] D. Sussillo, M. M. Churchland, M. T. Kaufman, and K. V. Shenoy, "A recurrent neural network that produces EMG from rhythmic dynamics," presented at the Front. Neurosci., Comput. Systems Neurosci., Salt Lake City, UT, USA, 2013.

[83] J. C. Kao, P. Nuyujukian, J. P. Cunningham, M. M. Churchland, S. I. Ryu, and K. V. Shenoy, "Increasing brain-machine interface performance by modeling neural population dynamics," presented at the Proc. Soc. Neurosci., San Diego, CA, USA, 2013.

[84] P. R. Kennedy and R. A. E. Bakay, "Restoration of neural output from a paralyzed patient by a direct brain connection," *Neuroreport*, vol. 9, no. 8, pp. 1707–1711, Jun. 1998.

[85] P. R. Kennedy, R. A. E. Bakay, M. M. Moore, K. Adams, and J. Goldwaithe, "Direct control of a computer from the human central nervous system," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 2, pp. 198–202, Jun. 2000.

[86] J. D. Simeral, S.-P. Kim, M. J. Black, J. P. Donoghue, and L. R. Hochberg, "Neural control of cursor trajectory and click by a human with tetraplegia 1000 days after implant of an intracortical microelectrode array," *J. Neural Eng.*, vol. 8, no. 2, Apr. 2011, 025027.

[87] V. Gilja, C. Pandarinath, C. H. Blabe, L. R. Hochberg, K. V. Shenoy, and J. M. Henderson, "Design and application of a high performance intracortical brain computer interface for a person with amyotrophic lateral sclerosis," presented at the Proc. Soc. Neurosci., San Diego, CA, USA, 2013.

[88] J. D. Simeral, D. Bacher, A. A. Sarma, N. J. Schmansky, S. D. Stavisky, B. Jarosiewicz, T. Milekovic, D. M. Rosler, V. Gilja, C. Pandarinath, A. S. Cornwell, J. M. Henderson, K. V. Shenoy, R. F. Kirsch, J. P. Donoghue, and L. R. Hochberg, "Evolution of the BrainGate real-time brain-computer interface (BCI) platform for individuals with tetraplegia or limb loss," presented at the Proc. Soc. Neurosci., San Diego, CA, USA, 2013.

[89] J. W. Judy, "Opportunities to improve long-term reliability," *IEEE Pulse*, vol. 3, no. 2, pp. 57–60, Apr. 2012.

[90] J. Krüger, F. Caruana, R. D. Volta, and G. Rizzolatti, "Seven years of recording from monkey cortex with a chronically implanted multiple microelectrode," *Front. Neuroeng.*, vol. 3, May 2010, DOI: 10.3389/fneng.2010.00006.

[91] C. A. Chestek, V. Gilja, P. Nuyujukian, J. D. Foster, J. M. Fan, M. T. Kaufman, M. M. Churchland, Z. Rivera-Alvidrez, J. P. Cunningham, S. I. Ryu, and K. V. Shenoy, "Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex," *J. Neural Eng.*, vol. 8, no. 4, Aug. 2011, 045005.

[92] J. C. Barrese, N. Rao, K. Paroo, C. Triebwasser, C. Vargas-Irwin, L. Franquemont, and J. P. Donoghue, "Failure mode analysis of silicon-based intracortical microelectrode arrays in non-human primates," *J. Neural Eng.*, vol. 10, no. 6, Dec. 2013, 066014.

[93] A. Prasad, Q.-S. Xue, V. Sankar, T. Nishida, G. Shaw, W. J. Streit, and J. C. Sanchez, "Comprehensive characterization and failure modes of tungsten microwire arrays in chronic neural implants," *J. Neural Eng.*, vol. 9, no. 5, Oct. 2012, 056015.

[94] U. Mitzdorf, "Current source-density method and application in cat cerebral cortex: Investigation of evoked potentials and EEG phenomena," *Physiol. Rev.*, vol. 65, no. 1, pp. 37–100, Jan. 1985.

[95] S. Katzner, I. Nauhaus, A. Benucci, V. Bonin, D. L. Ringach, and M. Carandini, "Local origin of field potentials in visual cortex," *Neuron*, vol. 61, no. 1, pp. 35–41, Jan. 2009.

[96] G. Buzsáki, C. A. Anastassiou, and C. Koch, "The origin of extracellular fields and currents-EEG, ECoG, LFP and spikes," *Nature Rev. Neurosci.*, vol. 13, no. 6, pp. 407–420, Jun. 2012.

[97] G. T. Einevoll, C. Kayser, N. K. Logothetis, and S. Panzeri, "Modelling and analysis of local field potentials for studying the function of cortical circuits," *Nature Rev. Neurosci.*, vol. 14, no. 11, pp. 770–785, Nov. 2013.

[98] R. D. Flint, E. W. Lindberg, L. R. Jordan, L. E. Miller, and M. W. Slutzky, "Accurate decoding of reaching movements from field potentials in the absence of spikes," *J. Neural Eng.*, vol. 9, no. 4, Jun. 2012, 046006.

[99] E. Stark and M. Abeles, "Predicting movement from multiunit activity," *J. Neurosci.*, vol. 27, no. 31, pp. 8387–8394, Aug. 2007.

[100] D. A. Heldman, W. Wang, S. S. Chan, and D. W. Moran, "Local field potential spectral tuning in motor cortex during reaching," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 180–183, Jun. 2006.

[101] C. Mehring, J. Rickert, E. Vaadia, S. Cardosa de Oliveira, A. Aertsen, and S. Rotter, "Inference of hand movements from local field potentials in monkey motor cortex," *Nature Neurosci.*, vol. 6, no. 50, pp. 1253–1254, Dec. 2003.

[102] H. Scherberger, M. R. Jarvis, and R. A. Andersen, "Cortical local field potential encodes movement intentions in the posterior parietal cortex," *Neuron*, vol. 46, no. 2, pp. 347–354, Apr. 2005.

[103] J. Rickert, S. C. D. Oliveira, E. Vaadia, A. Aertsen, S. Rotter, and C. Mehring, "Encoding of movement direction in different frequency ranges of motor cortical local field potentials," *J. Neurosci.*, vol. 25, no. 39, pp. 8815–8824, Sep. 2005.

[104] J. Zhuang, W. Truccolo, C. E. Vargas-Irwin, and J. P. Donoghue, "Decoding 3-D reach and grasp kinematics from high-frequency local field potentials in primate primary motor cortex," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1774–1784, Jul. 2010.

[105] A. K. Bansal, W. Truccolo, C. E. Vargas-Irwin, and J. P. Donoghue, "Decoding 3D reach and grasp from hybrid signals in motor and premotor cortices: Spikes, multiunit activity, and local field potentials," *J. Neurophysiol.*, vol. 107, no. 5, pp. 1337–1355, Dec. 2011.

[106] A. K. Bansal, C. E. Vargas-Irwin, W. Truccolo, and J. P. Donoghue, "Relationships among low-frequency local field potentials, spiking activity, and three-dimensional reach and grasp kinematics in primary motor and ventral premotor cortices," *J. Neurophysiol.*, vol. 105, no. 4, pp. 1603–1619, Apr. 2011.

[107] D. A. Markowitz, Y. T. Wong, C. M. Gray, and B. Pesaran, "Optimizing the decoding of movement goals from local field potentials in macaque cortex," *J. Neurosci.*, vol. 31, no. 50, pp. 18412–18422, Dec. 2011.

[108] R. D. Flint, C. Ethier, E. R. Oby, L. E. Miller, and M. W. Slutzky, "Local field potentials allow accurate decoding of muscle activity," *J. Neurophysiol.*, vol. 108, no. 1, pp. 18–24, Apr. 2012.

[109] E. J. Hwang and R. A. Andersen, "The utility of multichannel local field potentials for brain-machine interfaces," *J. Neural Eng.*, vol. 10, no. 4, Aug. 2013, 046005.

[110] R. D. Flint, Z. A. Wright, M. R. Scheid, and M. W. Slutzky, "Long term, stable brain machine interface performance using local field potentials and multiunit spikes," *J. Neural Eng.*, vol. 10, no. 5, Aug. 2013, 056005.

[111] C. A. Chestek, V. Gilja, P. Nuyujukian, R. J. Kier, F. Solzbacher, S. I. Ryu, R. R. Harrison, and K. V. Shenoy, "HermesC: Low-power wireless neural recording system for freely moving primates," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 17, no. 4, pp. 330–338, Aug. 2009.

[112] H. Miranda, V. Gilja, C. A. Chestek, K. V. Shenoy, and T. H. Meng, "HermesD: A high-rate long-range wireless transmission system for simultaneous multichannel neural recording applications," *IEEE Trans. Biomed. Circuits Syst.*, vol. 4, no. 3, pp. 181–191, Jun. 2010.

[113] D. A. Borton, M. Yin, J. Aceros, and A. V. Nurmikko, "An implantable wireless neural interface for recording cortical circuit dynamics in moving primates," *J. Neural Eng.*, vol. 10, no. 2, Apr. 2013, 026010.

[114] S. Kim, P. Tathireddy, R. A. Normann, and F. Solzbacher, "Thermal impact of an active 3-D microelectrode array implanted in the brain," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 4, pp. 493–501, Dec. 2007.

[115] P. D. Wolf, "Thermal considerations for the design of an implanted cortical brain-machine interface (BMI)," in *Indwelling Neural Implants: Strategies for Contending With the In Vivo Environment*, W. Reichert, Ed. Boca Raton, FL, USA: CRC Press, 2008.

[116] R. Silver, K. Boahen, S. Grillner, N. Kopell, and K. L. Olsen, "Neurotech for neuroscience: Unifying concepts, organizing principles, and emerging tools," *J. Neurosci.*, vol. 27, no. 44, pp. 11807–11819, Oct. 2007.

[117] J. Dethier, P. Nuyujukian, C. Eliasmith, T. Stewart, S. A. Elassaad, K. V. Shenoy, and K. Boahen, "A brain-machine interface operating with a real-time spiking neural network control algorithm," *Advances in Neural Information Processing Systems 24*. Cambridge, MA, USA: MIT Press, 2011, pp. 2213–2221.

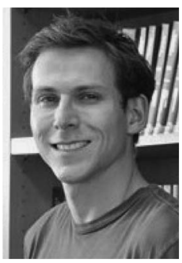
[118] A. V. Nurmikko, J. P. Donoghue, L. R. Hochberg, W. R. Patterson, Y.-K. Song, C. W. Bull, D. A. Borton, F. Laiwalla, S. Park, Y. Ming, and J. Aceros, "Listening to brain microcircuits for interfacing with external world—Progress in wireless implantable microelectronic neuroengineering devices," *Proc. IEEE*, vol. 98, no. 3, pp. 375–388, Mar. 2010.

[119] S. Kim, R. Bhandari, M. Klein, S. Negi, L. Rieth, P. Tathireddy, M. Toepper, H. Oppermann, and F. Solzbacher, "Integrated wireless neural interface based on the Utah electrode array," *Biomed. Microdevices*, vol. 11, no. 2, pp. 453–466, Apr. 2009.

ABOUT THE AUTHORS

Jonathan C. Kao (Student Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 2010, where he is currently working toward the Ph.D. degree in electrical engineering.

His research interests include algorithms for neural prosthetic control, neural dynamical systems modeling, and the development of clinically viable neural prostheses.



Sergey D. Stavisky (Student Member, IEEE) received the Sc.B degree in neuroscience from Brown University, Providence, RI, USA, in 2008. He is currently working toward the Ph.D. degree in neuroscience at Stanford University, Stanford, CA, USA.

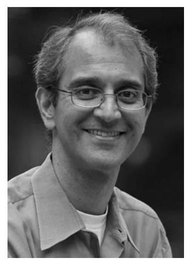
His research interests include developing neural prosthetics and studying how the brain's sensorimotor system uses these new effectors.

David Sussillo received the B.S. degree in computer science from Carnegie Mellon University, Pittsburgh, PA, USA, in 1999 and the M.S. degree in electrical engineering and the Ph.D. degree in neuroscience from the Columbia University, New York, NY, USA, in 2003 and 2009, respectively.

He is currently an Electrical Engineering Postdoctoral Fellow in the Laboratory of Krishna Shenoy at Stanford University, Stanford, CA, USA.

Paul Nuyujukian (Member, IEEE) received the B.S. degree in cybernetics from the University of California Los Angeles, Los Angeles, CA, USA, in 2006. He is in the MSTP program at Stanford University, Stanford, CA, USA, where he received the M.S. and Ph.D. degrees in bioengineering in 2011 and 2012, respectively, and is currently pursuing the M.D. degree.

His research interests include the development and clinical translation of neural prosthetics.



Krishna V. Shenoy (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of California Irvine, Irvine, CA, USA, in 1990 and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1992 and 1995, respectively.

He was a Neurobiology Postdoctoral Fellow at California Institute of Technology (Caltech), Pasadena, CA, USA, from 1995 to 2001, and then joined Stanford University, Stanford, CA, USA, where he is currently a Professor in the Department of Electrical Engineering, the Department of Bioengineering, and the Department of Neurobiology, and in the Bio-X and Neurosciences Programs. He is also with the Stanford Neurosciences Institute. His research interests include computational motor neurophysiology and neural prosthetic system design. He is the Director of the Neural Prosthetic Systems Laboratory and Co-Director of the Neural Prosthetics Translational Laboratory at Stanford University.

Dr. Shenoy was a recipient of the 1996 Hertz Foundation Doctoral Thesis Prize, a Burroughs Wellcome Fund Career Award in the Biomedical Sciences, an Alfred P. Sloan Research Fellowship, a McKnight Endowment Fund in Neuroscience Technological Innovations in Neurosciences Award, the 2009 National Institutes of Health Director's Pioneer Award, the 2010 Stanford University Postdoctoral Mentoring Award, and the 2013 Distinguished Alumnus Award from the Henry Samueli School of Engineering at the University of California Irvine.