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

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;
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.
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

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.
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: $x_k \in \mathbb{R}^M$ denotes a column vector containing the $M$ observed neural firing rates at time $k$, $\hat{x}_k$ denotes a vector containing the decoded prosthesis kinematic variables at time $k$, and $y_k \in \mathbb{R}^N$ 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 ($p_k$) and velocity ($v_k$) of a robotic arm at time $k$, then $x_k$ would be the vector concatenation of vectors $p_k$ and $v_k$. If $y_{ki}^T$ is the neural spiking activity of the $i$th neuron at time $k$, then $y_k^T = [y_{1i}^{T}~y_{2i}^{T}~\ldots~y_{Ni}^{T}]$ is the activity of all the recorded neurons at time $k$, where $[\cdot]$ denotes horizontal concatenation and $y^T$ denotes the transpose of vector $y$. For convenience, we also define matrices $X = [x_1 \ x_2 \ \ldots \ x_K]$ and $Y = [y_1 \ y_2 \ \ldots \ y_K]$, with $K$ denoting the number of observed time instances.

As this review focuses on spike-based decoders, the neural spiking activity $y_{ki}^T$ 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 $\Delta t$. The interval $\Delta t$ 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 $X$ and $Y$ is $T = K\Delta t$.

### 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
According to this view, the activity of the motor cortex, in which the neural activity of individual neurons represent 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: $y_{ik} = f(x_k)$. 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 $y_{ik} \propto \cos(\theta_k + \phi)$, where $\theta_k$ is the angle of reaching and $\phi$ 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 $d_i \in \mathbb{R}^M$. The contribution of each vector $d_i$ to the decoded kinematics $\hat{x}_k$ is linearly proportional to the firing rate of the neuron. Hence,

$$\hat{x}_k = \frac{c}{N} \sum_{i=1}^{N} \frac{y_{ik} - b_i}{\alpha^i} d_i$$

(1)

where $b_i$ is an offset term (typically the mean firing rate of neuron i) such that, when $y_{ik} < b_i$, the neuron contributes movement in the direction $-d_i$. The variable $\alpha^i$ 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 $z_k = (y_{ik} - b_i)/\alpha^i$ to be the normalized firing rates, then the PV algorithm can be written as $x_k = (c/N) \sum_{i=1}^{N} z_k^i d_i$. For convenience, we also define vector $z_k = [z_{k1}, z_{k2}, \ldots, z_{kN}]^T$ and matrix $Z = [z_1, z_2, \ldots, z_k]$. 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 $d_i$ 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: $d_{i Ole} = \sum_{j=1}^{N} R_{ij}^{-1} V_j$ for $R_{ij} = E[z_{ij}^T z_{ik}]$, the (i,j) entry of the correlation matrix of the normalized firing rates, and $V_j = E[z_{ij}^T x_k^T]$, 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 $d_i$ is uniform, then it can be shown that $d_{i Ole} = d_i$ so that the OLE and PV algorithms are equivalent. R and V are typically estimated by their time-averaged estimates $R = ZZ^T$ and $V = ZX^T$. Thus, OLE corresponds to a least squares solution. By defining $L_{ole} = R^{-1}V$, it is apparent that $L_{ole}$ minimizes the squared error of $X - (L_{ole})^T Z$. Here, the jth row of $L_{ole}$ corresponds to the preferred direction $d_{i Ole}$. In closed-loop control, the kinematics can be decoded by calculating $\hat{x}_k = (L_{ole})^T z_k$. Typically, a constant bias term $b_{ole}$ is also included so that $\hat{x}_k = (L_{ole})^T z_k + b_{ole}$. Recently, a clinical demonstration [11] used a variant of OLE called indirect OLE [44], in which a linear tuning model $Z = B X$ was learned and subsequently used to infer $L_{ole}$ by setting $L_{ole} = (B^T)$ (e.g., [23], [24], and [44]) where $B$ denotes the pseudoinverse of $B$. 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, Fits 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. Fits throughput is further discussed in Supplementary Materials section 3.1 of the study by Gilja et al. [14]. Using Fits 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

<table>
<thead>
<tr>
<th>Study</th>
<th>Algorithm</th>
<th>Fitts throughput (bps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim et al., 2008</td>
<td>VKF</td>
<td>0.52</td>
</tr>
<tr>
<td>Taylor et al., 2007</td>
<td>PV</td>
<td>0.33</td>
</tr>
<tr>
<td>Gilja et al., 2012</td>
<td>VF</td>
<td>0.69</td>
</tr>
<tr>
<td>Ganguly et al., 2009</td>
<td>WF</td>
<td>0.95</td>
</tr>
<tr>
<td>Sussillo et al., 2012</td>
<td>ESN</td>
<td>1.45</td>
</tr>
<tr>
<td>Gilja et al., 2012</td>
<td>ReFF-KF</td>
<td>1.81</td>
</tr>
</tbody>
</table>

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 (VFK), echo state networks (ESN), and a Kalman filter inspired from a control feedback approach (ReFF-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 bit rate, 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].
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 \( y'_k = f_j(x_k) \). 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 \( x_k = g(y_k, y_{k-1}, \ldots) \). 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 theory 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 \( y_1, \ldots, y_k \) is used to calculate a linear estimate of a signal \( x_k \), given by \( \hat{x}_k = \sum_{j=1}^{\infty} L^T_j y_j \), where \( L_j \in \mathbb{R}^{N \times M} \). In the Wiener–Kolmogorov filtering theory, the goal is to learn parameters \( L_1, \ldots, L_k \) such that the squared error in predicting \( x_k \) is minimized. A major distinction of the Wiener–Kolmogorov approach, in contrast to linear vector techniques, is the incorporation of neural history \( (y_{k-1}, y_{k-2}, \ldots) \) into the regression problem. In this section, we will describe the implementation of the Wiener filter by referring to the binned spike counts \( y_k \), but the Wiener filter could also be implemented using normalized spike counts \( z_k \).

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

\[
\hat{x}_k = \sum_{j=0}^{p-1} L^T_{j} y_{k-j}.
\]

(2)

(As in the OLE case, a constant bias term can also be included.) By defining \( L_W = [L_0 L_1 \ldots L_{p-1}] \), the vertical concatenation of the matrices \( L_0, L_1, \ldots, L_{p-1} \), the Wiener filter solution can be obtained by solving

\[
L_W = \left[ \begin{array}{ccc} R_{yy}(0) & R_{yy}(1) & \ldots & R_{yy}(p-1) \\ R_{yy}(1) & R_{yy}(0) & \ldots & R_{yy}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{yy}(p-1) & R_{yy}(p-2) & \ldots & R_{yy}(0) \end{array} \right]^{-1} \times \left[ \begin{array}{c} R_{yx}(0) \\ R_{yx}(1) \\ \vdots \\ R_{yx}(p-1) \end{array} \right]
\]

where \( R_{yy}(j) = E(y_j y_{k+j}^T) \) and \( R_{yx}(j) = E(y_j x_{k+j}^T) \) for all \( j = 0, \ldots, p-1 \). The index \( j \) refers to autocorrelations \( R_{yy}(j) \) or cross-correlations \( R_{yx}(j) \) 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 \( L_{\text{OLE}} = R_{xx}^{-1}(1) R_{xx}(0) \), where \( R_{xx}(0) \) and \( R_{xx}(0) \) are the normalized firing rate analogs of \( R_{yy}(0) \) and \( R_{yx}(0) \). The autocorrelations and cross-correlations are typically estimated by their time-averaged estimates. We let \( X_{[i,j]} \) denote \( [x_i x_{i+1} \ldots x_j] \) for \( i < j \), and define the following matrix:

\[
\hat{Y} = \left[ \begin{array}{cccc} y_p & y_{p+1} & \cdots & y_K \\ y_{p-1} & y_p & \cdots & y_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_1 & y_2 & \cdots & y_{K-p+1} \end{array} \right].
\]

Then, the time-averaged estimate of \( L_W \) can be calculated as \( L_W = (\hat{Y} \hat{Y}^T)^{-1} (\hat{Y} \hat{X}) \hat{K} \). 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 \( \tau \), then choosing \( p \) such that
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 \( L_j \) 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 the preferred directions associated with a neuron can be different for distinct time lags. While, at first, the Wiener filter may seem to be optimal. While, at first, the Wiener filter may seem to be optimal. While, at first, the Wiener filter may seem to be optimal.

A benefit in modeling a dynamical update law for the kinematic variables is the ability to enforce that the state and observation noise processes are typically continuous. A consequence of this is that the state and observation noise processes are typically continuous. A consequence of this is that the state and observation noise processes are typically continuous. A consequence of this is that the state and observation noise processes are typically continuous. A consequence of this is that the state and observation noise processes are typically continuous. A consequence of this is that the state and observation noise processes are typically continuous. A consequence of this is that the state and observation noise processes are typically continuous.

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 \( y_k \) [31]. The dynamical model proposes that the kinematics of the prosthesis \( x_k \) are the state of a linear time-invariant dynamical system, while the neural activity \( y_k \) 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

\[
\begin{align*}
    x_{k+1} &= Ax_k + w_k \\
    y_k &= Cx_k + q_k
\end{align*}
\]

with \( w_k \sim \mathcal{N}(0, W) \) and \( q_k \sim \mathcal{N}(0, Q) \). Because sequences \( \{x_k\}_{k=1...K} \) and \( \{y_k\}_{k=1...K} \) are observed in the training set while \( w_k \) and \( q_k \) are zero mean terms, \( A \) and \( C \) can be learned via least squares regression:

\[
    A = X_{[2×K]}X_{[1×K]}^T\{X_{[1×K]}X_{[1×K]}^T\}^{-1} \quad \text{and} \quad C = YX^T\{XX^T\}^{-1}
\]

After learning \( A \) and \( C \), \( W \) is calculated as the sample covariance of the residuals \( X_{[2×K]} - AX_{[1×K]} \), while \( Q \) is analogously the sample covariance of the residuals \( Y - CX \). Given \( A \), \( W \), \( C \), \( Q \) as well as an initial state condition, \( x_0 \) (often set to be zero), the Kalman filter recursively estimates the current state \( x_k \), given the current neural observation \( y_k \), and the previous state estimate \( x_{k-1} \) [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 \( x_k = [p_k^T \; v_k^T]^T \), then the \( A \) matrix can be additionally designed such that the position obeys \( p_{k+1} = p_k + v_k\Delta t \). Further, the \( A \) 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).
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]

\[
\dot{x}_k = M_1 \dot{x}_{k-1} + M_2 y_k \tag{7}
\]

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:

\[
L_j = M_j \sigma^2_j
\]

for \(j = 0, \ldots, p\). 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 \(x_k = v_k\), e.g., [32]), matrix \(M_1\) is of the form \(M_1 \approx \alpha I\) with \(\alpha < 1\), since BMI training paradigms tend to sample kinematic velocities uniformly in all directions. Therefore, the velocity Kalman filter is
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 those of BMI systems using velocity Kalman filtering, as shown in Table 2.

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

\[ \tau \dot{s}_k = -s_k + J r_k + W_i y_k + W_f \hat{x}_k \]

where \( s_k \) is the hidden state of the recurrent network. The hidden units interact through matrix \( J \). The continuous variable \( r_k \) is the “instantaneous firing rate” and is defined as \( r_k = \tanh(s_k) \). Interesting dynamics arise in the network due to this nonlinear coupling. The inputs \( y_k \) enter the system through weights \( W_i \) while a linear readout of the kinematics \( \hat{x}_k = W_o r_k \) is fed back to the hidden units through feedback weights \( W_f \).

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 \( J \) matrix and limited supervised training of only the output weights \( W_o \). 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 \( \hat{x}_k \) (along with input \( y_k \)). Because learning is focused exclusively on \( W_o \), 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 \( W_o \) 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

\[ y_k \]

674 PROCEEDINGS OF THE IEEE | Vol. 102, No. 5, May 2014
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
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 $p_k$ are removed, indicating that no uncertainty is propagated to $p_k$. The ReFIT–KF algorithm also estimates the contribution of position to the neural activity by finding the matrix $C_p$ that minimizes the squared error $y_k - C_p p_k$. Given the assumption that there is no uncertainty in the decoded position $\hat{p}_k$, the ReFIT–KF algorithm subtracts the position contribution to the neural signal by calculating $\tilde{y}_k = y_k - C_p \hat{p}_k$. Subsequently, the position subtracted neural data $\tilde{y}_k$ 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 multunit 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...
have examined the LFP recorded during natural reaches
and shown that this signal contains considerable informa-
tion 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 deco-
ders that beneficially combine these signals would be a
major step toward increasing the clinical viability of BMI
systems.

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 wide-
spread BMI use, it would be useful for BMI systems to be
autonomous, capable of running for days without recal-
ibration 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 infrastruc-
ture. 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

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 wide-
spread BMI use, it would be useful for BMI systems to be
autonomous, capable of running for days without recal-
ibration 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 infrastruc-
ture. 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

Kao et al.: Information Systems Opportunities in Brain–Machine Interface Decoders

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].
brain and surrounding tissue. The limit for power dissipation set by the American Association of Medical Instrumentation is 10 mW within a 6 × 6 mm² 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 decoder 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.

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. Eisenee for administrative assistance.

REFERENCES

Kao et al.: Information Systems Opportunities in Brain-Machine Interface Decoders


[54] H. Jaeger and H. Haas, “Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication,”
Kao et al.: Information Systems Opportunities in Brain-Machine Interface Decoders


H. Miranda, V. Gilja, C. A. Chestek, Ph.D. degree in electrical engineering. 2010, where he is currently working toward the B.S. and M.S. degrees in electrical engineering from Stanford University, Stanford, CA, USA.

2010, Stanford, CA, USA. His research interests include algorithms for neural prosthetic control, neural dynamical systems modeling, and the development of clinically viable neural prosthetics.

Sergey D. Stavisky  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  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 Student 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.