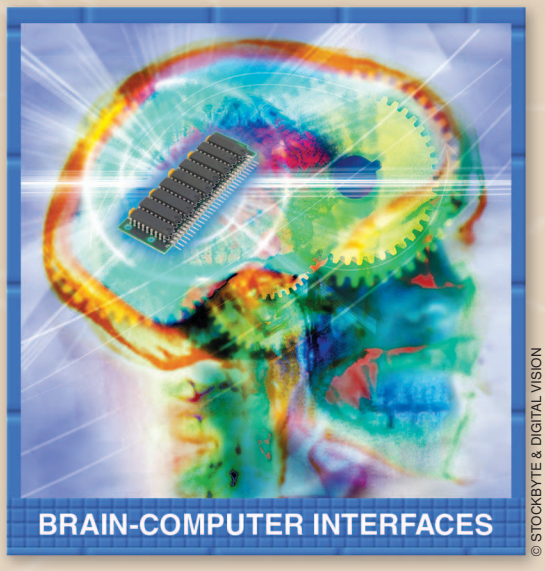
Michael D. Linderman, Gopal Santhanam, Caleb T. Kemere, Vikash Gilja, Stephen O'Driscoll, Byron M. Yu, Afsheen Afshar, Stephen I. Ryu, Krishna V. Shenoy, and Teresa H. Meng



Signal Processing Challenges for Neural Prostheses

A review of state-of-the-art systems

Cortically controlled prostheses are able to translate neural activity from the cerebral cortex into control signals for guiding computer cursors or prosthetic limbs. While both noninvasive and invasive electrode techniques can be used to measure neural activity, the latter promises considerably higher levels of performance and therefore functionality to patients. The process of translating analog voltages recorded at the electrode tip into control signals for the prosthesis requires sophisticated signal acquisition and processing techniques. In this article we briefly review the current state-of-the-art in invasive, electrode-based neural prosthetic systems, with particular attention to the advanced signal processing algorithms that enable that performance. Improving prosthetic performance is only

Digital Object Identifier 10.1109/MSP.2007.909016

1053-5888/08/$25.00©2008IEEE

part of the challenge, however. A clinically viable prosthetic system will need to be more robust and autonomous and, unlike existing approaches that depend on multiple computers and specialized recording units, must be implemented in a compact, implantable prosthetic processor (IPP). In this article we summarize recent results which indicate that state-of-the-art prosthetic systems can be implemented in an IPP using current semiconductor technology, and the challenges that face signal processing engineers in improving prosthetic performance, autonomy and robustness within the restrictive constraints of the IPP.

INTRODUCTION

An emerging class of prostheses aims to provide control of paralyzed upper limbs, prosthetic arms, and computers by translating cortical neural activity into control signals. A number of research groups have now demonstrated that monkeys and humans can learn to move computer cursors and robotic arms to various locations simply by activating the neural populations that participate in natural arms movements [1]–[7]. These compelling proof-of-concept laboratory demonstration systems motivate the development of clinically viable, electrode-based neural prosthetic systems that exhibit the level of cortical control needed for many everyday behaviors. The process of translating analog voltages recorded at the electrode tip into control signals for a prosthesis requires sophisticated signal acquisition and processing techniques. The challenge then to signal processing engineers is twofold: develop neural signal processing algorithms that achieve the maximum possible prosthetic performance and do so in a clinically viable manner.

Neural prosthetic systems are only clinically viable when the anticipated quality of life improvement outweighs the potential risks. Noninvasive techniques [8] are attractive due to their reduced surgical risk (and well studied, see other articles in this issue), however, invasive, electrode-based techniques have become a major research thrust, as they offer high signal quality and thus the potential for increased performance relative to noninvasive approaches. For example, the current state-of-the-art electrode-based system in our laboratory achieves an information transfer rate of 6.5 b/s [7], many fold higher than previously reported invasive and noninvasive systems. The tradeoffs for invasive approaches, however, are increased surgical risk and high cost. As a result, at present, chronic electrode-based prosthetic systems are a long-term approach, with near-term applications potentially limited to only the most severely disabled patients. The transition from research to widespread use will require improving the performance-risk-cost balance by increasing overall prosthetic performance and reducing surgical risk and device cost through system integration.

The prosthesis performance cited above is made possible through high-quality neural signal measurement and advanced signal processing methods, in particular uncompromising real-time action potential identification (spike sorting) [9] and probabilistic movement decoding algorithms (in particular, a special case of [10]). These techniques are differentiated from other approaches by their ability to extract more unique neurons, more accurately, in the spike identification process, and incorporate more, and make better use of, neural activity in the decoding process. It is anticipated that >10 b/s systems are achievable with further improvements in neural measurement and signal processing methods [11].

Equipment intensive, laboratory-based experiments, like those cited above, in which a restrained subject performs a highly controlled task under supervision by a trained researcher are a powerful experimental platform but not necessarily the best approximation of a clinical environment. Clinical systems cannot be reliant on trained operators and external control; they must be autonomous and capable of identifying patient intent, specifically whether neural activity actually corresponds to an intended movement, using that neural activity alone. Furthermore, prosthetic systems must provide these capabilities continuously (24 h/day, everyday) and robustly, not just during the short, discrete daily recording periods used in current experimental protocols. Spike sorting algorithms that utilize unsupervised learning reduce the need for a trained operator and offer the potential for robust, adaptive algorithms which respond autonomously to changes in the neural recordings. Similarly decoding algorithms with autonomous neural state detection (movement intended or not) eliminate the need for external cues to identify time periods with relevant neural activity.

Prosthetic systems need to reduce surgical risk and device cost by enabling system integration and eliminating chronic transcutaneous connectors. The goal is a fully integrated prosthetic system, where electrodes, digital post-processing and wireless telemetry comprise a single implantable unit that will provide state-of-the-art performance in a self-contained package with reduced physical footprint and no chronic tissue openings [12]. In such an approach, however, signal acquisition and processing must be performed within a very restrictive power budget. The transmission of neural information out of the electrode implantation site is a key challenge. While the required bandwidth is within the capability of current wireless links, the power consumption of such a link is prohibitive. Some form of bandwidth reduction is essential. There are a number of approaches to achieve this reduction; however, many utilize lossy compression and thus can potentially reduce prosthetic performance. Our goal is to not sacrifice any prosthesis performance; thus we wish to implement the same high performance signal processing algorithms we use in the laboratory, which can reduce the required bandwidth by a factor of $\sim 10^6$, in the implantable system, while meeting power constraints.

The combination of strict power constraints, aggressive performance goals, and robustness and autonomy requirements present a difficult design challenge to signal processing engineers, and new algorithms and implementations will be needed. In this article, we provide a brief introduction to chronic electrode-based neural prosthetic systems, with particular attention to the digital post-processing algorithms in current state-of-the-art laboratory based systems. We will show that such algorithms along with the relevant signal acquisition hardware are

in principle realizable as an IPP in current CMOS technology within power consumption constraints [13], [14]. Better performing, more robust algorithms are possible, and, as we will present, will be required to transition from laboratory-based systems to an IPP operating continuously, and autonomously, in a clinical setting.

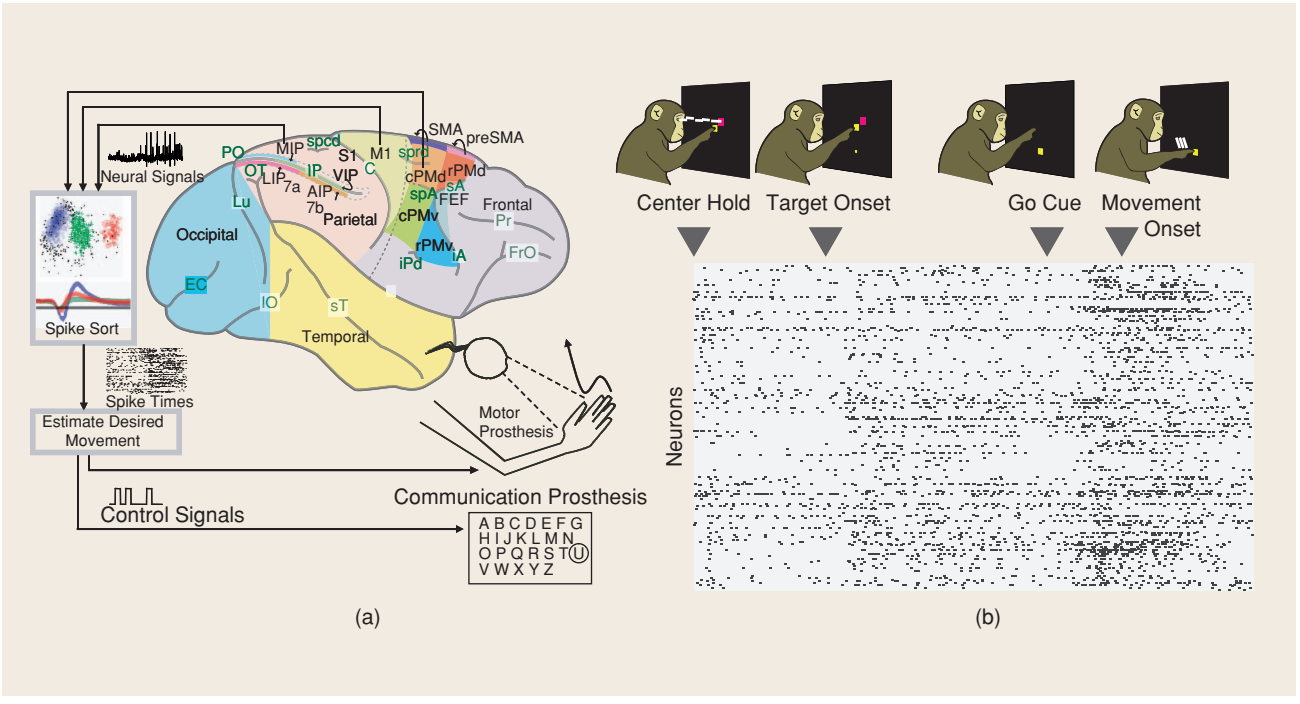
CHRONIC ELECTRODE-BASED NEURAL PROSTHESES

The basic architecture of motor and communication prostheses are shown in Figure 1(a). Motor prostheses aim to provide neural control of the paralyzed limb, while communication prostheses aim to provide a communication channel equivalent to "typing" on a computer. The relationship between a movement and the neural response (tuning) is used to design estimation (decoding) algorithms to infer the desired movement from only the neural activity, a sample of which is shown in Figure 1(b). The system can then generate control signals appropriate for continuously guiding a paralyzed or prosthetic arm through space (motor prosthesis) or positioning a computer cursor on the desired letter on a keyboard (communication prosthesis). Two types of neural spike activity, plan and movement, are well suited for driving prosthetic systems plan activity, present from soon after the reach target is identified until just after the movement begins, is tuned for the target of the movement. (In the context of this article, neural activity is almost exclusively spiking activity recorded in the motor (M1) and dorsal premotor (PMd) cortex. While the local field potential (LFP) has been shown to predict movement direction in other cortical regions [16], its role in M1 and PMd remains unclear.) This goal information can be decoded to drive communication prostheses, which only need to estimate the movement endpoint [5], [7]. Motor prostheses must incorporate movement activity since the goal is to recreate the desired movement. Plan activity can play a role in motor prostheses, however, by providing a probabilistic prior, or target estimation, to constrain movement estimation [15], [10].

NEURAL PROSTHETIC SYSTEMS ARE ONLY CLINICALLY VIABLE WHEN THE ANTICIPATED QUALITY OF LIFE IMPROVEMENT OUTWEIGHS THE POTENTIAL RISKS.

ACHIEVING STATE-OF-THE-ART PERFORMANCE

As shown in Figure 1, there are four major components—neural signal acquisition, spike sorting, neural decoding and control signal generation—that can be engineered to improved prosthetic system performance. Control signals are considered to be part of the actuation system and are not discussed. Neural signal quality, and particularly the number of independent neurons that could possibly be observed is a function of the electrode technology, surgical placement and time since implantation, all topics beyond the scope this article. The two remaining blocks, spike sorting and neural decoding, are where significant improvements in performance can be realized through advanced signal processing algorithms, and thus will be the focus of the remainder of this article. In the following section we will describe the algorithms



[FIG1] (a) Concept sketch of cortically controlled motor and communication prostheses. Adapted from [11]. (b) Neural activity (spikes indicated by black dots) during typical instructed-delay reaching task. Adapted from [15].

used to achieve the 6.5 b/s transfer rate communication prostheses described in [7]. Although these are not the most advanced algorithms available, they are some of the highest performing in active use in a complete prosthetic system, and are excellent examples of the types of algorithms currently being developed.

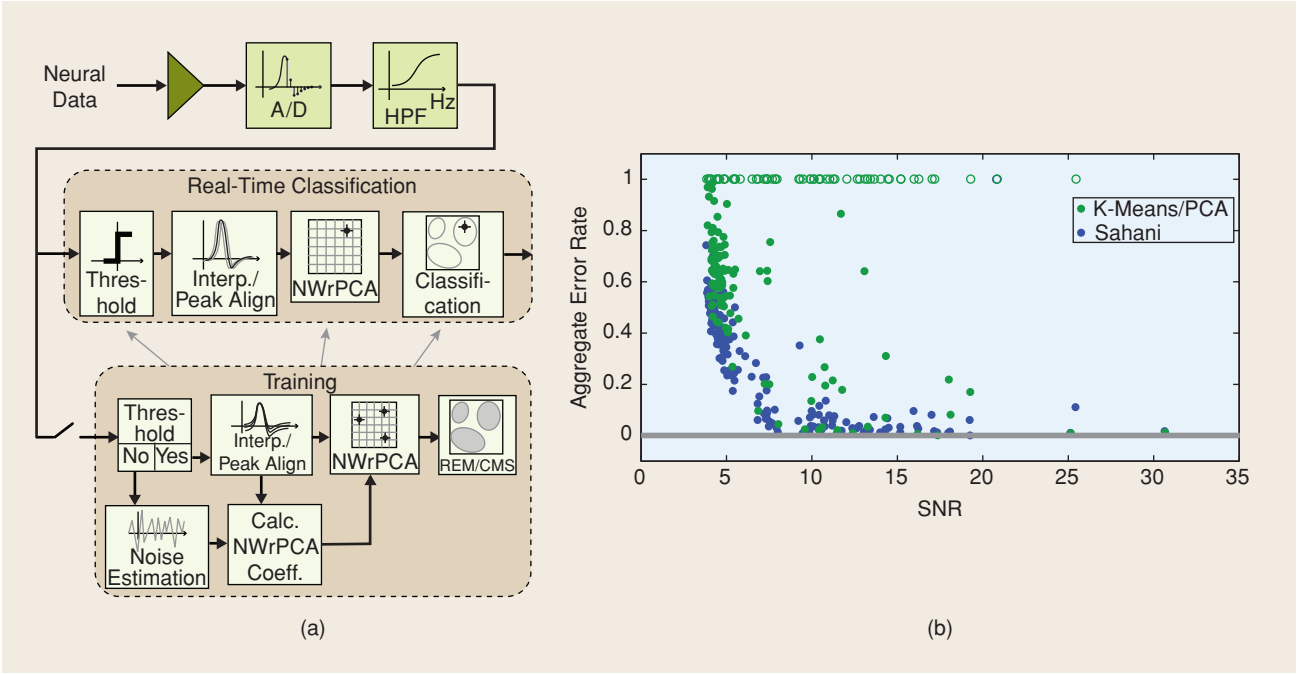
SPIKE SORTING

The neural signal captured by an extracellular electrode contains the overlapping, noisy measurement of action potentials (spikes) from several nearby neurons. Automatic or semi-automatic techniques for disambiguating between unique neurons, often termed "spike sorting," have been heavily investigated (see [17] for a review and [18] for a description of more recent methods). For a given channel, each neuron is assumed to produce a unique and consistent spike waveform (100–400 μV peak-to-peak, 1 ms in duration), which is then corrupted by noise. Spike-sorting techniques attempt to identify and classify these distinct spike shapes. Most algorithms consist of two phases, a training phase, when the sort parameters are set using a subset of the neural recording, and a classification phase, during which all identified spikes are assigned to originating neurons using previously determined parameters.

Although numerous automated spike sorters have been developed, they are only beginning to come into widespread use among electrophysiology researchers. The benefit of a manual, or semimanual approach is direct control of the spike identification; the tradeoffs are long sorting times and inconsistency (measured average false positive and false negative rates for manual sorting are 23 and 30% respectively [19]). In contrast, automated sorters are faster, more consistent (for a given data set), often more sensitive and accurate (i.e., able to extract more unique neurons, more accurately) and, most importantly for clinical applications, able to operate without human intervention. We describe one particular automated spike sorting algorithm, the Sahani algorithm [9], which is in daily use in our laboratory, and is an excellent example of a large class of statistically rigorous sorters.

It must be noted that while some form of spike sorting is employed in most systems, it is neither required nor universally used. In [20], the authors describe a threshold based spike identification methodology that does not attempt to disambiguate between multiple neurons recorded on a single electrode. Implemented with a simple analog circuit, this approach has very low power consumption and significantly reduces telemetry bandwidth, both important when developing an IPP. However, this approach conflates all the neurons recorded on a given channel to a single neuron. Two neurons observed on the same channel are not guaranteed, however, to have identical response properties. Therefore when unique neurons are not disambiguated information is irrecoverably lost, artificially limiting overall system performance.

Figure 2(a) shows the block diagram for the Sahani sorting algorithm. After digitization the broadband neural signal (sampled at 30 kHz) is high-pass filtered (cut off at 250 Hz) to remove the low frequency components (termed LFP) and expose the spikes. The rms of the filtered signal is computed, and a threshold of 3× the rms voltage is used to identify spike events. These events are "snipped" from the signal waveform, forming a set of ~1 ms (32 sample) spike snippets. Segments that do not exceed the threshold are also collected and used to estimate the background noise process. The characterization of



[FIG2] Sahani spike sorting algorithm (a) Block diagram showing signal flow for Sahani spike sorting algorithm. Both training and real time classification paths are shown. (b) Aggregate error rate versus neuron signal-to-noise ratio (SNR) for Sahani algorithm (blue) and K-means/PCA based sorting (green). Adapted from [14].

the background noise enables the projection of the spike waveforms into a robust noise-whitened principal components (NWrPCA) space. During training, relaxation expectation-maximization (REM) and cascading model selection (CMS) are used to cluster the data and fit the clusters to a mixture model. The mixture model represents the prior probability of observing each neuron identified as well as the probability of threshold crossings corresponding to noise rather than neural activity.

While computationally complex, the Sahani algorithm (and other similar algorithms) offer significant performance improvement. In comparing performance with a much simpler K-means/PCA-based technique, two critical aspects are made apparent. By using cascading model selection, the Sahani algorithm can typically determine the correct number of neurons autonomously, a crucial feature for an unsupervised spike sorter. Furthermore, the mixture model approach provides a well-founded technique for rejecting threshold-crossing events that do not actually correspond to neural spikes. Figure 2(b) shows the aggregate error rate of the Sahani algorithm (blue) and a K-means/PCA based sorting approach (green) for a synthetically generated data set [14]. The aggregate error rate is defined as the sum of false positives and false negatives divided by the total number of spikes generated. The hollow circles indicate missed neurons or error rates greater than 100%. The median aggregate classification error rate (false positive and false negative) for the Sahani algorithm is 3.7%. In contrast, the K-means/PCA approach, which assumed three unique neuron per electrode, misclassified many neurons entirely. Even if these misclassified neurons are removed, the aggregate median error rate is 20%. (See [14] for description of synthetic data generation and the K-means/PCA sorting methodology.)

PROSTHETIC SYSTEMS NEED TO REDUCE SURGICAL RISK AND DEVICE COST BY ENABLING SYSTEM INTEGRATION AND ELIMINATING CHRONIC TRANSCUTANEOUS CONNECTORS.

NEURAL DECODING

The intended movement can be estimated from the neural activity (as identified by the spike sorter) using parameterized models. Examples of decoding algorithms currently in use include population vectors [3] and linear filters, [2], [4], [6]. Both of these decoders assume a linear relationship between the neural activity and intended movement. The linear algorithms are effective, and attractive due to their low latency and simple implementations, but more accurate movement estimation can be obtained using recursive Bayesian decoders [21]–[23]. Unlike the linear decoders, the probabilistic methods allow for nonlinear relationships between the neural activity and the intended movement and provide confidence regions for the movement estimates. And although the probabilistic decoders tend to be more complex than the linear methods, the latency can be kept low by combining the results of many simple probabilistic decoders running in parallel [10].

As discussed previously, the plan activity reliably indicates the intended reach goal and can serve either as the primary source of information (for a communication prosthesis) or as a probabilistic prior constraining movement estimation [2]. Let z be a $q \times 1$ vector of spike counts across the q simultaneously recorded neurons in a prespecified time window (e.g., 100 ms) during the delay period preceding the reach. The distribution of spike counts (from training data) for each reach goal m can be fit to either a product of Gaussians or a product of Poissons. In both models the neurons are assumed to be independent given the reach goal.

For any test trial, the probability that the upcoming reach goal is m given the plan activity z can be computed by applying Bayes' rule

$$P(m|z) = \frac{P(z|m)P(m)}{P(z)} = \frac{P(z|m)}{\sum_{m'} P(z|m')} \qquad (1)$$

where $P(m)$ is assumed to be uniform. The most likely reach goal (*i.e.,* the one with the largest $P(m|z)$) is taken to be the decoded reach goal.

The accuracy of the goal decoder varies with the duration and placement of the time window in which spikes are counted, as well as the precise spike count model $P(z|m)$ that is used. The 6.5 b/s communication prosthetic performance cited earlier is achieved in large part by optimizing the configuration of the time window with respect to overall prosthetic performance [7]. Earlier placement with respect to target appearance and shorter duration enables more trials in a fixed unit of time, but with reduced accuracy, with the opposite true for longer duration. The maximum information transfer rate capacity (ITRC) actually occurs using eight targets at short window durations (~70 ms, corresponding to short trials), despite the relatively low single-trial accuracy at these durations (~70% versus ~90% achieved with long windows).

The results of goal decoding can be used to constrain trajectory estimation for motor prosthesis in several ways. Goal directed reaches are observed to be highly stereotyped and thus can be recreated with high accuracy using a set of canonical trajectories selected by a goal decoder [15]. Alternately, $P(m|z)$ can be incorporated into a probabilistic trajectory estimator in place of the otherwise uniform probability of a given target on a given trial [10].

BUILDING AN IPP

Current neural prosthetic systems, which use microconnectors to bring neural signals out of the body and a rack full of computers and specialized post-processing hardware, are not clinically viable in the long term. Although already in use for human clinical trials [6], these systems, all of which required skilled operators, will not scale for use outside laboratory settings. In

previous sections we described algorithms, like unsupervised spike sorting, which reduce the need for a trained operator. Equally as important is reducing the physical system footprint, cost, and surgical risk. By integrating the entire prosthetic system, or a large portion thereof, into a single unit with wireless telemetry and powering, the cost and size can be reduced and the transcutaneous connector, a potential source of infection, can be eliminated. To address these needs we propose an integrated IPP, which combines a variable precision analog-to-digital converter array (ADC) [13], a digital spike sorter [14], maximum-likelihood neural decoder [15], and a wireless data and power transceiver (an integrated analog front end with wireless transceiver is described in [12]).

The IPP described here is not the only approach for building an implantable neural prosthetic system. Other low power designs have been proposed. In [12] and [20], described earlier, the authors proposed analog comparator-based spike identification. However this system does not differentiate between different neurons recorded on the same channel, thereby reducing the amount of information extracted. In [24], the authors limited the number of channels, ADC resolution and sophistication of the spike sorting algorithm to reduce power consumption. In [25], a lossy wavelet encoding scheme is used to reduce the necessary data bandwidth (and thus reducing transmitter power). The shapes of the action potentials are preserved and post-processing can be used perform spike sorting. However the data bandwidth reduction is smaller than can be achieved with integrated sorting and decoding. Furthermore, the effect of the compression loss on the ability to distinguish spikes from different neurons is unknown.

In all the systems described above, the designers were forced to make tradeoffs to reduce power consumption. However, we argue that such concessions are not necessary. Instead, current laboratory class capabilities can be retained by implementing the digital post processing in hardware as part of the implantable system. Using a metric of 1 GOPS/mW, the sorting and decoding algorithms described previously, along with the ADC and amplifiers are estimated to consume less than 10 μW per channel [14], well below the limit for safe power dissipation into the brain (80 mW/cm^2 [26]). The ADC array is a large power consumer, and reducing its power dissipation is key to minimizing overall power consumption; the per channel power consumption is: ADC: 4.2 μW, digital filter and threshold: 1.32 μW, real-time sorting: .1 μW, spike-sorter training: 2.8 μW [14]. ADC power consumption can be reduced by 3.6$\times$ to 1.16 μW using a variable precision ADC array [13], which sets the optimal bit depth of each ADC (no higher than necessary to save power, but no lower than necessary to maintain spike sorting accuracy) using the results of the downstream spike sorter. The current focus is on designing more power-efficient implementations for the digital filtering and spike sorter training, efforts that would benefit from signal processing expertise. In the case of the spike sorter training, the solution might lie in similar forms synergy, such as the neural decoder feeding back to the spike sorter, an as of yet largely unexplored but potentially bountiful source of performance improvements.

MOTOR PROSTHESES AIM TO PROVIDE NEURAL CONTROL OF THE PARALYZED LIMB, WHILE COMMUNICATION PROSTHESES AIM TO PROVIDE A COMMUNICATION CHANNEL EQUIVALENT TO "TYPING" ON A COMPUTER.

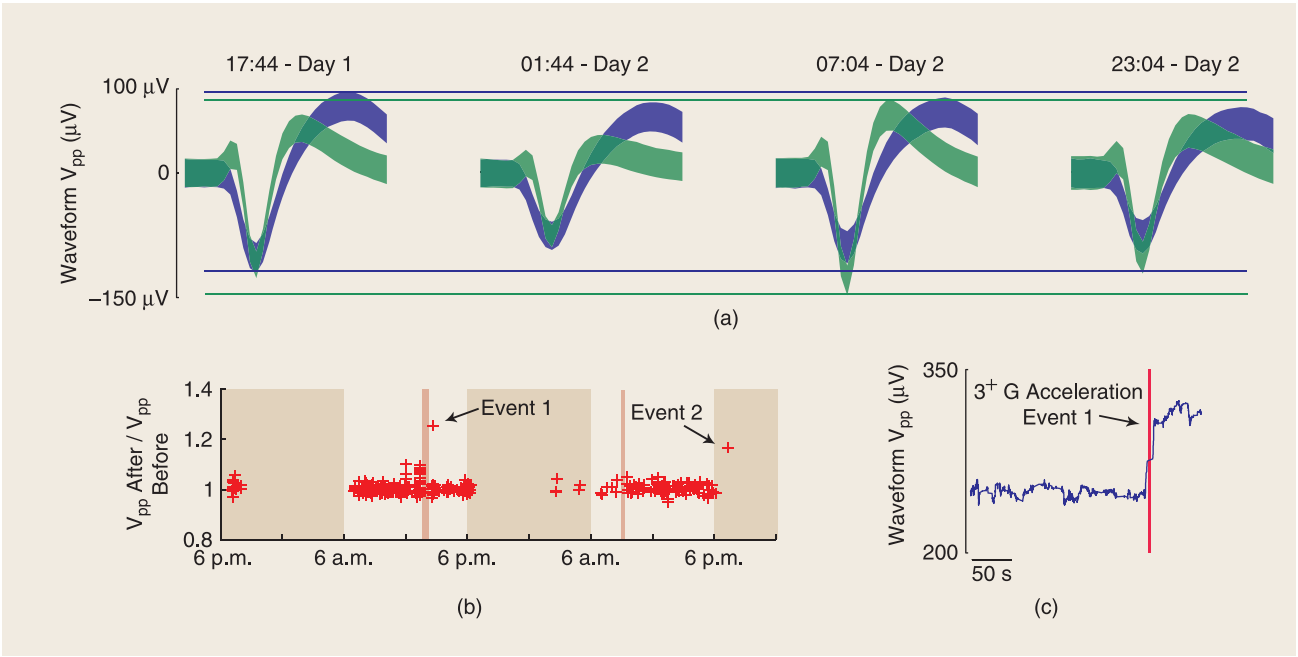
NEW CHALLENGES FOR NEURAL PROSTHESES

The spike sorting and decoding techniques described above are relatively mature and ready for research into low-power implementations. The transition from experimental to clinical settings, however, requires prosthetic systems to be more robust and autonomous, challenges not addressed by the established techniques. Although some solutions have been proposed, most are in their infancy, and thus the primary research focus is the development of principled algorithms. The following sections describe some of the signal processing research underway to provide robust spike sorting and autonomous prosthetic control and neural decoding.

CONTEXT DETECTION

A typical, daily, laboratory prosthetic session lasts two to three hours and is conducted by highly trained researchers. A clinical prosthetic system, however, will need to operate continuously (24 h/day, everyday) with minimal outside assistance. Reliable prosthetic performance across different behavioral contexts is imperative. Prosthetic systems will need to interpret the user's current behavioral context (i.e., sleeping versus active) so as to most efficiently use resources, by going into low-power sleep mode for example, and, perhaps, more importantly, so as to not generate undesired actions, such as arm movements while sleeping. Any approach that utilizes outside assistance to identify these various contexts will not be clinically viable due to cost and scalability, and systems that rely on the user to manually select modes might have basic technological problems (e.g., how does a user reliant on a prosthetic system for movement wake it up?). As neural prosthetic systems transition from the laboratory to the clinical setting, new types of information beyond just reaching control and discrete target selection will be required.

Although such topics are only beginning to be examined, there is strong evidence that these macro-behavioral contexts can, and must, be determined form the same neural activity used to drive the prosthetic systems. Using an autonomous, long-duration neural recording system for freely behaving primates that we developed, called HermesB, we recorded numerous neural channels nearly continuously over 54 h (5-min recording periods separated with 2.5-min break) [27]. With



[FIG3] Neural recording stability for freely behaving primate. (a) Spike waveforms of two neurons for selected five-min blocks across 54 h. Colored regions indicate 10–90th percentile in amplitude. Horizontal lines indicate maximum and minimum voltages for each neuron. (b) Local change in mean waveform amplitude ($V_{pp}^{after}/V_{pp}^{before}$) (red + symbols) for 200 snippets before and after 3G acceleration events. (c) Waveform V_{pp} moving average (over 200 snippets) centered about the time indicated for the 5-min block containing event 1 from panel (b). Time of event 1 is indicated with red vertical bar. In (b) the wide grey regions indicate night and the thin pink regions indicate "pit stops" when the monkey was taken form the home cage and placed in the primate chair to service the recording equipment. Adapted from [27].

the accelerometer built into the HermesB system, the recording blocks could be classified as active or inactive. We noticed that while the spiking statistics were not a good indicator of behavioral context, the LFP power in 5–25 Hz band was different in the two contexts. Using a simple LFP power threshold, >89% of 5-min blocks were correctly classified as active/inactive, suggesting that using a small subset of the neural information (just one channel in this case) we could accurately monitor the subject's behavioral context. Although a simple example, this type of analysis highlights the types of additional information that will be useful in developing responsive, intuitive clinical prostheses.

ROBUST SPIKE SORTING

Part of reducing outside assistance is making prosthetic systems more robust. In most experimental protocols the parameters used in spike sorting and decoding are regenerated at the beginning of the session and then assumed to remain constant for the duration of the session. Daily parameter regeneration is required to compensate for changes in the neural signals observed between recording sessions. However, the timescales at which these changes occur are much shorter than one day [27], suggesting that prosthetic systems will need to regenerate their parameters (termed retraining) more often. Using the long duration recordings made with the HermesB system we characterized the stability of neural recordings at the intermediate timescales (i.e., between discrete daily recording periods) inaccessible with traditional experimental protocols. We observed variations of up to 30% in mean waveform amplitude over periods ranging from 5 min to several hours, up to 5 μV change in background RMS voltage, and abrupt changes in waveform amplitude of up to 25% ($V_{pp}^{after}/V_{pp}^{before}$) in response to high acceleration movements of the subject's head.

Figure 3(a) shows the waveform amplitude for two neurons for selected 5-min blocks across 54 h recorded from a freely behaving subject. The lines of constant voltage provide a reference against which one can see the large changes in the waveform amplitude. These changes in action potential shape have been previously observed across once-daily recordings [28]. Here, the results gathered with HermesB from a freely behaving primate indicates substantial variation in spikes waveforms over intermediate timescales as well. Figure 3(b) shows the local change in waveform amplitude ($V_{pp}^{after}/V_{pp}^{before}$) when the head mounted accelerometer measured a > 3G acceleration event. Most of ~1,700 events show little or no change in waveform amplitude, however, two events (indicated by arrows) show much larger, abrupt changes in waveform amplitude. Figure 3(c) shows the waveform amplitude as a function of time for the 5-min block during which event 1 occurred. The red vertical line marks the the >3 G acceleration. The close alignment between the acceleration event and the step change in waveform amplitude strongly suggests that the relationship between the waveform variation and acceleration event is not coincidental. The profile is consistent with an abrupt change in array posi-

tion. The second event shows similar behavior but is not shown. Variations in waveform amplitude and background noise can both have adverse affects on spike sorting performance, either through the use of inappropriate threshold or outright misclassification. The spike identification threshold is typically set as a multiple of the rms noise (typically 3×), thus a 5 μV change in the RMS noise will translate into a much larger change in the threshold, potentially resulting in missed spikes, or an increase in noise generated, nonneural snippets. Spike classification is based on the waveform shape, and in particular the amplitude, and so changes in the waveform amplitude will result in misclassification. Using the types of variations shown in Figure 3, recorded neural waveforms were artificially perturbed (either to increase or decrease amplitude) prior to spike sorting and decoding. As would be expected, as the number of neurons and the extent of perturbation increases, the decoding accuracy in communication prosthetic experiments decreases. For the particular experimental data (taken from [7]) used in this simulation, the 30% variation in waveform amplitude observed with HermesB translated to a reduction in decoding accuracy from 92% to 62% (unpublished observation, V. Gilja).

Tolerance to some variations in neural recordings has already been incorporated into sorting algorithms. Firing rate dependent changes in spike shape can be addressed by incorporating firing statistics into the spike sorting algorithm [29] and changes in rms voltage can be addressed through adaptive thresholding [20]. Long time-scale variations, however, may require periodic retraining of the spike sorting parameters. There does not appear to be a consensus on exactly what retraining period is required. Experiments that use discrete daily recording periods typically only update once per day, but future prosthetic systems that operate continuously will likely need to retrain more regularly. If the necessary retraining interval is short enough, adaptive approaches, which link together otherwise independent retraining operations might be appropriate.

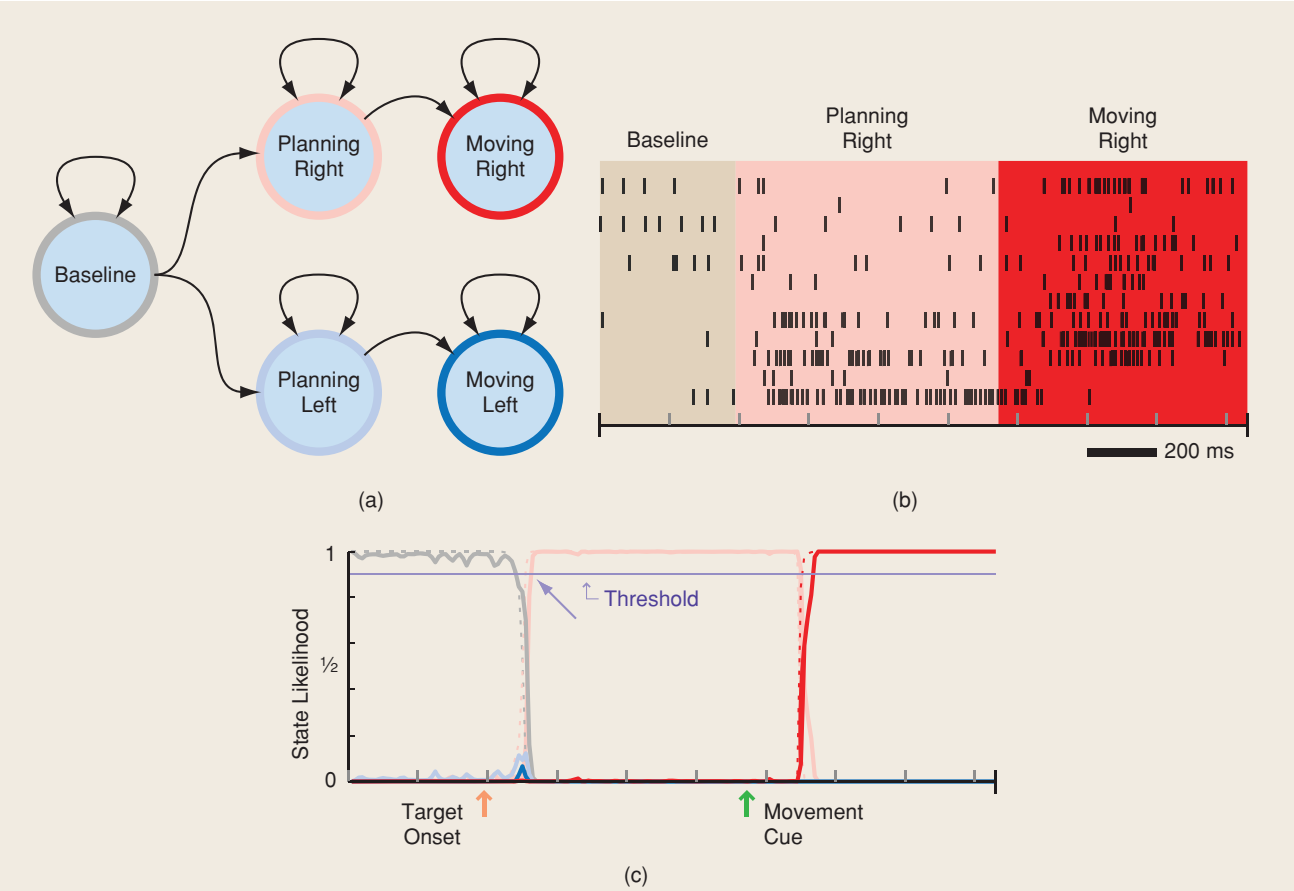
In [30], the authors propose an algorithm that divides the data into a set of short frames, sorts each frame independently, and then performs a second global clustering operation. Each frame (~1,000 spikes) is assumed stationary, allowing all variations to be addressed at the global level. In [31], the authors propose tracking changes in waveform shape by linking high-dimensional spike clusters between frames, essentially following the "crumbs" of shifting clusters. The tradeoffs for both approaches is the computational overhead of continually retraining and the large memory footprint of maintaining global information over long time scales. For these reasons it is unlikely that these algorithms could be successfully implemented on a clinical IPP. Instead, a truly adaptive approach, which continuously integrates and updates the parameters might be the best approach. A suitable algorithm would have an effective training interval short enough to track variations in waveform shape and background process, without the cost of discrete retraining, and long duration global waveform shape tracking.

THERE ARE FOUR MAJOR COMPONENTS—NEURAL SIGNAL ACQUISITION, SPIKE SORTING, NEURAL DECODING AND CONTROL SIGNAL GENERATION—THAT CAN BE ENGINEERED TO IMPROVED PROSTHETIC SYSTEM PERFORMANCE.

AUTONOMOUS DECODE CONTROL

The macro-behavioral introspection described previously can provide insight into the general behavioral context (e.g., awake versus sleeping) but is insufficiently precise for guiding the decoding of a particular movement. During each movement, motor cortical firing rates transition through a sequence of discrete and stable states, termed "epochs," such as baseline, prepare, and execute (the latter two corresponding to the plan and movement periods). Most prosthetic systems, including those described previously, require knowing when one or both of plan and movement activity is present. When the transition between epochs is not detected, and neural activity is incorrectly or imprecisely "labeled," decoding performance suffers. Current experimental protocols require human intervention to differentiate between epochs; to be useful outside the laboratory, however, prosthetic systems will need to determine the epoch autonomously. Researchers have begun to tackle this problem; a finite state machine (FSM)-based state detector was recently proposed in [32]. This approach uses a sliding-window maximum-likelihood state classifier coupled to a finite state machine (FSM) to estimate the current epoch.

The state transition can be modeled as a Markov process in which the hidden state transitions through three epochs identified above. Using the moment-by-moment a posteriori likelihoods of the states of the resultant hidden Markov model (HMM), the current epoch (baseline, preparatory, or execution) can be accurately, and autonomously, estimated [15]. Compared to the ad-hoc FSM-based approach, the HMM-based epoch estimator provides increased accuracy, and offers a principled approach that can leverage the existing body of work on HMMs; can more readily be extended to incorporate other classes of neural data, such as LFP; and can potentially adapt to the nonstationarities in neural recordings described previously [27]. A simplified, didactic version, of the HMM-based decoder using two targets and ten neurons is shown in Figure 4. During the baseline phase, a small number of states (drawn as one for simplicity) model the variation in background of neural activity. For each possible target, one state models plan activity and second models execution, or movement, activity. The HMM structure enables the inference of state likelihoods using a simple, well-known recursive computation in which the a priori estimate and the newest observation are used to compute an a



[FIG4] HMM regime detection. (a) Simple five-state reaching movement HMM. (b) An example of the neural activity for a rightward movement. Ten neurons, recorded simultaneously are shown. Black bars indicate spike times. (c) Time series of state likelihoods for each HMM state. The arrow depicts the estimated time of the beginning of the planning regime for the threshold value depicted. Adapted from [15].

posteriori estimate (abbreviated APL and defined specifically as $\mathcal{L}(i, t)$, the likelihood of being in state i at time t). Figure 4(c) depicts the estimated APL for this trial. In this example, the transitions form the baseline to plan to movement regimes of activity are quite apparent in the spikes, and as expected, the estimated state probabilities track these transitions accurately and closely.

The state probabilities can be used in two ways. First, as described earlier, the epoch can be determined by combining the APL of activity regimes across goals. For plan activity,

$$\Pr(\text{preparatory regime} \mid \mathbf{n}_{0:t}) = \frac{\sum_{i \in \mathcal{P}} \mathcal{L}(i, t)}{\sum_j \mathcal{L}(j, t)}, \qquad (2)$$

where $\mathbf{n}_{0:t}$ is the neural activity up until time t and $\mathcal{P}$ represents plan states (in this case planning left or right). The probability of movement states can be found similarly. The time at which the probability crosses a predetermined threshold is an estimate of the time of transition between epochs. An autonomous neural prosthesis can be formed by combining this epoch estimation with the recursive Bayesian decoder described previously. The second use for the state probabilities is as a decoder and not just an epoch detector. At any moment during the trial, it is possible to estimate the target of the intended movement as the target whose combined preparatory and movement APL is highest.

CONCLUSIONS

The success of laboratory-based neural prosthetic systems provide proof of concept and motivate the continued development of clinical prostheses. Although the basic neuroscience research will always be ongoing, many of the obstacles facing the prosthetics community as it develops a clinically viable implantable prosthetic processor are primarily engineering challenges. In this article we identified some of these challenges, namely improving the robustness, autonomy and power efficiency of the prosthetics systems, along with potential solutions. The challenges are formidable, but familiar to the engineering community, and the field of neural prosthetics will benefit greatly from the early and continued involvement of

experts in signal acquisition, signal processing and analog and digital system design.

AUTHORS

Michael D. Linderman (mlinderm@stanford.edu) received the B.S. degree in engineering from Harvey Mudd College, Claremont, California, in 2003 and the M.S. degree in electrical engineering from Stanford University, where he is currently working toward the Ph.D. degree. His research interests are multiprocessor architectures and programming models for neural prosthetic systems and other challenging information processing problems. He received the National Defense Science and Engineering Graduate Fellowship and Stanford Graduate Fellowship. He is a student member of the IEEE.

Gopal Santhanam (gopals@stanford.edu) received the B.S. degree in electrical engineering and computer science and the B.A. degree in physics from the University of California, Berkeley. He later completed his M.S. and Ph.D. degrees in electrical engineering in 2002 and 2006, respectively, from Stanford University. His research involved neural prosthetics system design, neural signal processing, and embedded neural recording systems. He also has extensive industry experience through various consulting projects involving embedded systems. He received the National Defense Science and Engineering Graduate fellowship and the National Science Foundation graduate fellowship.

Caleb T. Kemere (ckemere@phy.ucsf.edu) received his B.S. in electrical engineering from the University of Maryland, College, Park in 1998. After receiving his M.S. in electrical engineering at Stanford University in 2000, he was with Datapath Systems (now part of LSI Logic, Inc), before returning to Stanford, receiving his Ph.D. in 2006. He is currently a Sloan-Schwartz Postdoctoral Fellow in computational neuroscience at the University of California, San Francisco. His research focuses on general hardware and signal processing problems related to multichannel neural interfaces. He currently uses multichannel recording and real-time optical stimulation to probe neural circuitry underlying the acquisition of spatial memory. He is a student member of the IEEE.

Vikash Gilja (vgilja@stanford.edu) received S.B. degrees in electrical engineering and computer science and brain and cognitive sciences in 2003 and the M.Eng. degree in electrical engineering and computer science in 2004 from the Massachusetts Institute of Technology, Cambridge. He is currently working toward the Ph.D. degree in computer science at Stanford University. At Stanford, he joined the Neural Prosthetics Laboratory. His research interests center around the design of practical and robust neural prosthetics systems. He received the National Defense Science and Engineering Graduate Fellowship and the National Science Foundation Graduate Fellowship.

Stephen O'Driscoll (stiofan@stanford.edu) received the B.E. degree in electrical engineering from University College Cork, in 2001 and the M.S. degree in electrical engineering from Stanford University in 2005, where he is currently working toward the Ph.D. degree. His current research interests focus on low-power analog circuit design and wireless power transfer. He received the IEE Graduate Prize (Ireland) 2001 and the Lu Stanford Graduate Fellowship in 2003. He represented Ireland at the 38th International Mathematical Olympiad in Argentina in 1997. He is a student member of the IEEE.

Byron M. Yu (byronyu@stanford.edu) received the B.S. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 2001. He received the M.S. and Ph.D. degrees in electrical engineering in 2003 and 2007, respectively, from Stanford University. He is currently a postdoctoral fellow jointly in electrical engineering and neuroscience at Stanford University and at the Gatsby Computational Neuroscience Unit, University College London. His research interests include statistical techniques for characterizing multi-channel neural data, their application to the study of neural processing, and decoding algorithms for neural prosthetic systems.

Afsheen Afshar (afsheen@stanford.edu) received a B.S.E. degree in electrical engineering and a certificate in engineering biology from Princeton University in 2002 and an M.S. in electrical engineering from Stanford University in 2005. He is currently working towards an M.D. and a Ph.D. degree in electrical engineering at Stanford University. His research involves improving neural prosthesis performance, developing better neural processing algorithms, and helping further the understanding of the Parkinsonian brain. His awards include the Medical Scientist Training Program (MSTP) Fellowship and the Bio-X Fellowship. He is a student member of the IEEE.

Stephen I. Ryu (seoulman@stanford.edu) received the B.S. and M.S. degrees in electrical engineering from Stanford University in 1994 and 1995, respectively. He received the M.D. degree from the University of California at San Diego in 1999 and completed neurosurgical residency and fellowship training at Stanford University in 2006. He is an assistant professor in the Department of Neurosurgery at Stanford University. His research interests include brain-machine interfaces, neural prosthetics, spine biomechanics, and stereotactic radiosurgery.

Krishna V. Shenoy (shenoy@stanford.edu) received the B.S. degree in electrical engineering from the University of California, Irvine, in 1990, and the S.M. and Ph.D. degrees in electrical engineering and computer science from MIT in 1992 and 1995. He was a neurobiology postdoctoral fellow at Caltech from 1995 to 2001 and then joined the Stanford University faculty where he is an assistant professor in the Department of Electrical Engineering and Neurosciences Program. His research interests include computational motor neurophysiolgy and neural prosthetic system design. His awards and honors include the 1996 Hertz Foundation Doctoral Thesis Prize, a Burroughs Wellcome Fund Career Award in the

Biomedical Sciences, an Alfred P. Sloan Research Fellowship, and a McKnight Endowment Fund in Neuroscience Technological Innovations in Neurosciences Award. He also serves on the Defense Science Research Council for DARPA. He is a Senior Member of the IEEE.

Teresa H. Meng (thm@stanford.edu) is the Reid Weaver Dennis Professor of Electrical Engineering at Stanford University. Her current research interests focus on neural signal processing, bioimplant technologies, and parallel computation architectures. In 1999, she took leave from Stanford and founded Atheros Communications, Inc. She returned to Stanford in 2000 to continue her research and teaching at the University. She is a member of National Academy of Engineers and Fellow of the IEEE. She received her Ph.D. in electrical engineering and computer science from the University of California at Berkeley in 1988.

REFERENCES

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

[2] M.D. Serruya, N.G. Hatsopoulos, L. Paninski, M.R. Fellows, and J.P. Donoghue, "Instant neural control of a movement signal," *Nature*, vol. 416, no. 6877, pp. 141–142, 2002.

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

[4] 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, p. E42, 2003.

[5] S. Musallam, B.D. Corneil, B. Greger, H. Scherberger, and R.A. Andersen, "Cognitive control signals for neural prosthetics," *Science*, vol. 305, no. 5681, pp. 258–262, 2004.

[6] 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, 2006.

[7] G. Santhanam, S.I. Ryu, B.M. Yu, A. Afshar, and K.V. Shenoy, "A high-performance brain-computer interface," *Nature*, vol. 442, no. 7099, pp. 195–198, 2006.

[8] J.R. Wolpaw and D.J. McFarland, "Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 51, pp. 17849–17854, 2004.

[9] M. Sahani, "Latent variable models for neural data analysis," Ph.D. dissertation, Dept. Computat. Neural Syst., Calif. Inst. Technol., 1999.

[10] B.M. Yu, C. Kemere, G. Santhanam, A. Afshar, S.I. Ryu, T.H. Meng, M. Sahani, and K.V. Shenoy, "Mixture of trajectory models for neural decoding of goal-directed movements," *J. Neurophys.*, vol. 97, no. 5, pp. 3763–3780, 2007.

[11] K.V. Shenoy, G. Santhanam, S.I. Ryu, A. Afshar, B.M. Yu, R.S. Kalmar, J.P. Cunningham, C.T. Kemere, A.P. Batista, M.M. Churchland, and T.H. Meng, "Increasing the performance of cortically-controlled prostheses," in *Proc. Conf. IEEE EMBS*, 2006, pp. 6652–6656.

[12] R.R. Harrison, P.T. Watkins, R.J. Kier, R.O. Lovejoy, D.J. Black, B. Greger, and F. Solzbacher, "A low-power integrated circuit for a wireless 100 electrode neural recording system," *IEEE J. Solid State Circuits*, vol. 42, no. 1, pp. 123–133, 2007.

[13] S. O'Driscoll, T. Meng, K. Shenoy, and C. Kemere, "Neurons to silicon: Implantable prosthesis processor," in *Proc. ISSCC*, 2006, pp. 552–553.

[14] Z.S. Zumsteg, C. Kemere, S.O'Driscoll, G. Santhanam, R.E. Ahmed, K.V. Shenoy, and T.H. Meng, "Power feasibility of implanatable digital spike sorting circuits for neural prosthetic systems," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 13, no. 3, pp. 272–279, 2005.

[15] C.K. Kemere, "Model-based decoding of neural signals for prosthetic interfaces," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., 2006.

[16] 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, 2005.

[17] M.S. Lewicki, "A review of methods for spike sorting: the detection and classification of neural action potentials," *Network*, vol. 9, no. 4, pp. R53–R78, 1998.

[18] S. Shoham, M.R. Fellows, and R.A. Normann, "Robust, automatic spike sorting using mixtures of multivariate t-distributions," *J. Neurosci. Methods*, vol. 127, no. 2, pp. 111–122, 2003.

[19] F. Wood, M.J. Black, C. Vargas-Irwin, M. Fellows, and J.P. Donoghue, "On the variability of manual spike sorting," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 912–918, 2004.

[20] P.T. Watkins, G. Santhanam, K.V. Shenoy, and R.R. Harrison, "Validation of adaptive threshold spike detector for neural recording," in *Proc. Conf. IEEE EMBS*, 2004, pp. 4079–4082.

[21] E.N. Brown, L.M. Frank, D. Tang, M.C. Quirk, and M.A. Wilson, "A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells," *J. Neurosci*, vol. 18, no. 18, pp. 7411–7425, 1998.

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

[23] W.Wu, Y. Gao, E. Bienenstock, J.P. Donoghue, and M.J. Black, "Bayesian population decoding of motor cortical activity using a kalman filter," *Neural Computat.*, vol. 18, no. 1, pp. 80–118, 2006.

[24] R.H. Olsson and K.D. Wise, "A three-dimensional neural recording microsystem with implantable data compression circuitry," *IEEE J. Solid State Circuits*, vol. 40, no. 12, pp. 2796–2804, 2005.

[25] K.G. Oweiss, "A systems approach for data compression and latency reduction in cortically controlled brain machine interfaces," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 7, pp. 1364–1377, 2006.

[26] T.M. Seese, H. Harasaki, G.M. Saidel, and C.R. Davies, "Characterization of tissue morphology, angiogensis, and temperature in adaptive response of muscle tissue to chronic heating," *Lab Investigation*, vol. 78, no. 12, pp. 1553–1562, 1998.

[27] G. Santhanam, et al., "HermesB: A neural recording system for freely behaving primates," *IEEE Trans. Biomedical Eng.*, to be published.

[28] S. Suner, M.R. Fellows, C. Vargas-Irwin, G.K. Nakata, and J.P. Donoghue, "Reliability of signals from a chronically implanted, silicon-based electrode array in non-human primate primary motor cortex," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 13, no. 4, pp. 524–541, 2005.

[29] C. Pouzat, M. Delescluse, P. Viot, and J. Diebolt, "Improved spike-sorting by modeling firing statistics and burst-dependent spike amplitude attenuation: A Markov chain Monte Carlo approach," *J. Neurophysiol.*, vol. 91, no. 6, pp. 2910–2928, 2004.

[30] A. Bar-Hillel, A. Spiro, and E. Stark, "Spike sorting: Bayesian clustering of non-stationary data," in *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press, 2005, pp. 105–112.

[31] R.K. Snider and A.B. Bonds, "Classification of non-stationary neural signals," *J. Neurosci. Methods*, vol. 84, no. 1–2, pp. 155–166, 1998.

[32] N. Achtman, A. Afshar, G. Santhanam, B.M. Yu, S.I. Ryu, and K.V. Shenoy, "Free-paced high-performance brain computer interfaces," *J. Neural Eng.*, vol. 4, no. 3, pp. 336–347, 2007. SP