Power Feasibility of Implantable Digital Spike Sorting Circuits for Neural Prosthetic Systems

Zachary S. Zumsteg, Caleb Kemere, *Student Member, IEEE*, Stephen O'Driscoll, *Student Member, IEEE*, Gopal Santhanam, Rizwan E. Ahmed, Krishna V. Shenoy, *Member, IEEE*, and Teresa H. Meng, *Fellow, IEEE*

***Abstract*—A new class of neural prosthetic systems aims to assist disabled patients by translating cortical neural activity into control signals for prosthetic devices. Based on the success of proof-of-concept systems in the laboratory, there is now considerable interest in increasing system performance and creating implantable electronics for use in clinical systems. A critical question that impacts system performance and the overall architecture of these systems is whether it is possible to identify the neural source of each action potential (spike sorting) in real-time and with low power. Low power is essential both for power supply considerations and heat dissipation in the brain. In this paper we report that state-of-the-art spike sorting algorithms are not only feasible using modern complementary metal oxide semiconductor very large scale integration processes, but may represent the best option for extracting large amounts of data in implantable neural prosthetic interfaces.**

***Index Terms*—analog-to-digital converter (ADC), brain–machine interfaces (BMI), low-power, neural prosthetics, spike sorting.**

I. Introduction

IMPLANTABLE arrays of hundreds of microelectrodes hold promise for fundamental neuroscience research and interfaces for patients with debilitating neuromuscular deficits. While modulations in the low frequency neuronal oscillations may contain useful information, the primary mechanism of information transmission in recordings of extracellular cortical signals is changes in the rate of pulse-like action potentials ("spikes") generated by individual neurons. Cortical electrode arrays are implanted neurosurgically, but the precise distance between each electrode tip and surrounding neurons is uncontrolled. Hence, implantable electrodes are manufactured with moderate impedances (e.g., a few hundred kiloohms) to ensure that at least one neuron is typically sensed. However, such electrodes often record action potentials from more than one neuron. We have shown previously that even for a simple classification task, performance can be increased by distinguishing the action potentials of different neurons detected on the same electrode [1]. Hence, an interface that discards this information will compromise system performance, at least to some extent [2].

Unfortunately, the transmission of this neural information out of the implanted site can be a difficult technical challenge. For neurophysiological studies, a single wire per electrode is bonded to a micro-connector mounted on the skull. However, for clinical applications with larger numbers of electrodes, the connectorization will become unmanageable despite miniaturization. Eventually a wireless interface will be the most practical alternative.

Herein lies the problem: while the fundamental information gathered by the array is the sequence of spike times for each neuron that can be sensed, the current approach of attempting to extract this information involves sampling the signal from each electrode at a rate of 10–30 kHz and transmitting the data to a signal processing system for analysis. As has been previously observed [3], for a sampling rate of 25 kHz, 12-bit digitization, and 96 electrodes, this yields an aggregate data rate of nearly 29 Mb/s. While this data rate is comparable to those achieved by the increasingly ubiquitous high-speed wireless network links, the power requirements for a clinical application necessitate battery lifetimes measured in years, as opposed to the hours that such high-bandwidth devices afford. Some form of bandwidth reduction is thus essential.

Two approaches for achieving this bandwidth reduction have been proposed. In [4], the waveform from each electrode is compressed using a lossy wavelet encoding scheme. A 30-fold data reduction is demonstrated by thresholding wavelet coefficients at the estimated noise level and using a lossless run-length coding. The shapes of the action potentials are preserved and postprocessing can be used to determine which neuron spiked for any given activity on an electrode. Still, the resultant data rate may be rather high in the context of a battery powered device, particularly in brain regions with high spike rates. Furthermore, the effect of the compression loss on the ability to distinguish spikes from different neurons is unknown.

Alternatively, the signal on each electrode could be reduced to simply the times at which it exceeded a threshold set at a multiple of a running estimate of its root mean square (rms) value [3]. Based on biophysical spiking constraints, the data rate per electrode (with two sensed neurons) would be a maximum 2 kb/s, 150 times smaller than the 300 kb/s raw signal. However,

Manuscript received April 22, 2005; accepted 6/24/05. This work of T. H. Meng and C. Kemere was supported by the MARCO Center for Circuit and System Solutions under Contract 2003-CT-888. The work of G. Santhanam was supported by the National Defense Science and Engineering Graduate (NDSEG) and National Science Foundation (NSF) fellowships. The work of K. V. Shenoy was supported by the NSF Center for Neuromorphic Systems Engineering at Caltech, ONR Adaptive Neural Systems, Whitaker Foundation, Center for Integrated Systems at Stanford, Sloan Foundation, and Burroughs Wellcome Fund Career Award in the Biomedical Sciences.

Z. S. Zumsteg, C. Kemere, S. O'Driscoll, G. Santhanam, and T. H. Meng are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: zzumsteg@stanford.edu; ckemere@stanford.edu; gopals@stanford.edu; stiofan@stanford.edu; thm@stanford.edu).

R. E. Ahmed is with the Systems Engineering Group, Qualcomm Inc., San Diego, CA 92121 USA.

K. V. Shenoy is with the Department of Electrical Engineering and the Neurosciences Program, Stanford University, Stanford, CA 94305 USA (e-mail:shenoy@stanford.edu).

Digital Object Identifier 10.1109/TNSRE.2005.854307

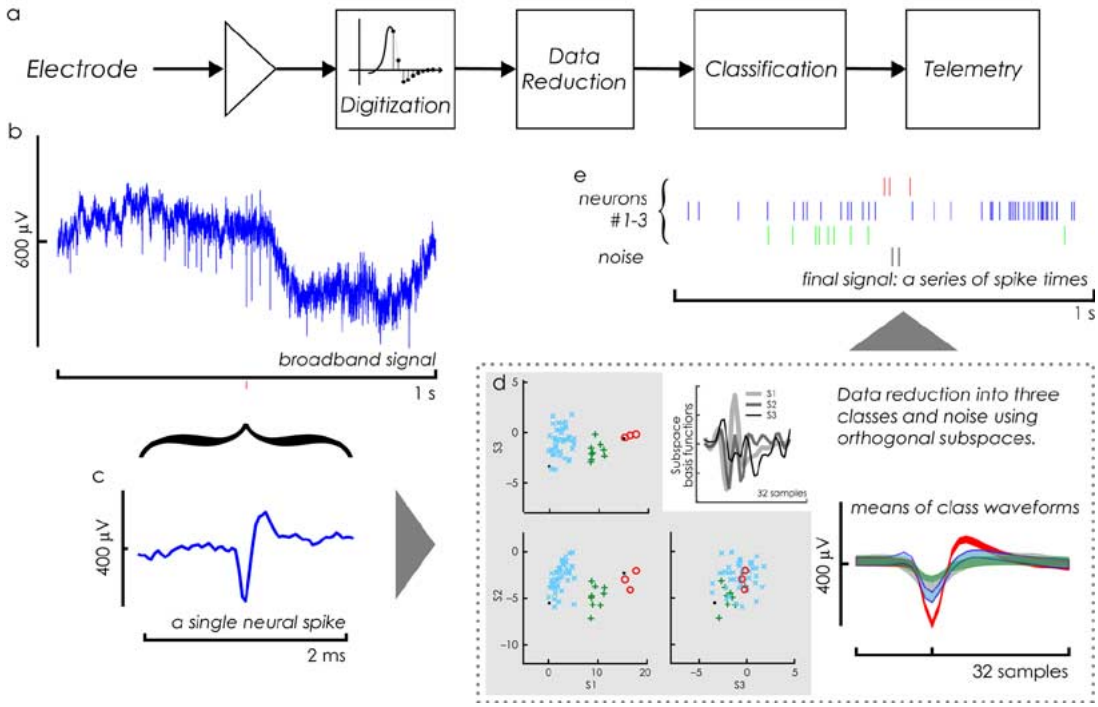


Fig. 1. Extraction of neural signals. a: General block diagram of data extraction from cortical neural recordings for a prosthetic interface: Broad-band signal [(b) 1 s of data; (c) 2 ms, showing a spike] recorded on electrode is first digitally sampled. Then, a feature extraction process reduces the dimensionality of the data [(d) spike waveforms in an optimized three dimensional subspace are easy to distinguish]. In this reduced signal space, the activity of individual neurons can be differentiated from each other and from background noise. Optimally, only the spiking times of neurons (e) are finally transmitted from the device to the downstream system which decodes neural activity into control signals for a prosthetic device.

the cost of such data reduction is that the information about which neuron produced which spike has been lost.

In this paper, we investigate the feasibility of a third alternative, namely implementing digital signal processing on the interface silicon so that spike sorting operations can be executed prior to transmitting data from the implant site. This approach achieves nearly the same transmission bandwidth floor of the threshold detector while achieving the neuron-by-neuron discrimination of the compression approach. We choose a modern digital spike sorting algorithm and demonstrate that the number of computational operations required, and hence the energy consumed when using standard CMOS VLSI, make an implantable spike sorting front-end realistic. This result holds for arrays with very large numbers of electrodes. Placing such a device into the context of other suggested techniques, this implies that aggressive bandwidth reduction, comparable to that achieved by merely recording threshold crossings, is possible without loss of neural information at a power consumption suitable for implantable devices.

II. Methodology

A. Spike Sorting Methodology

Certain features are common to nearly all spike sorting algorithms, as shown in Fig. 1. The signal from each electrode must be impedance (down) converted, filtered and amplified (Fig. 1(a), triangle). Additionally, because the fundamental data desired is the timing of action potentials, some form of conversion, whether a single comparator or more complex, from the analog waveform to a digital signal is required. Importantly, increased complexity in digitization can enable more information to be extracted from the signal. As waveforms are typically sampled at a rate much higher than the spiking rates of neurons, data reduction typically follows digitization. Then, detected spikes are classified into feature-derived categories corresponding to individual neurons or multiunit activity. Finally, the time and neural identity of each detect spike must be transmitted out.

B. Spike Sorting Algorithms Selected

In this paper, in order to demonstrate the feasibility of high quality real-time spike sorting in implanted hardware, we describe the implementation of what we believe to be both one of the best and most computationally intensive spike sorting algorithms available. We intentionally sought a state-of-the-art spike sorting algorithm, which is uncompromising in spike sorting quality and relies on principled machine learning techniques, to help assure that our power estimates would not be overly optimistic. Moreover, we routinely and productively use the algorithm presented to perform real-time neural prosthetic system experiments [1], [5], [6].

Fig. 2 depicts the spike sorting system developed by Sahani [7]. For real time classification, the data must first be preconditioned, using a high-pass filter (HPF) to eliminate the low-frequency local field potential (LFP). Spike times are then identified using a threshold identified in training. Spike event waveforms, small, $\sim$ 1 ms windows of data around a threshold-crossing event, are aligned to their exact peaks through an interpolation technique. These events are then projected into a noise-whitened robust principal components analysis (PCA) space, as

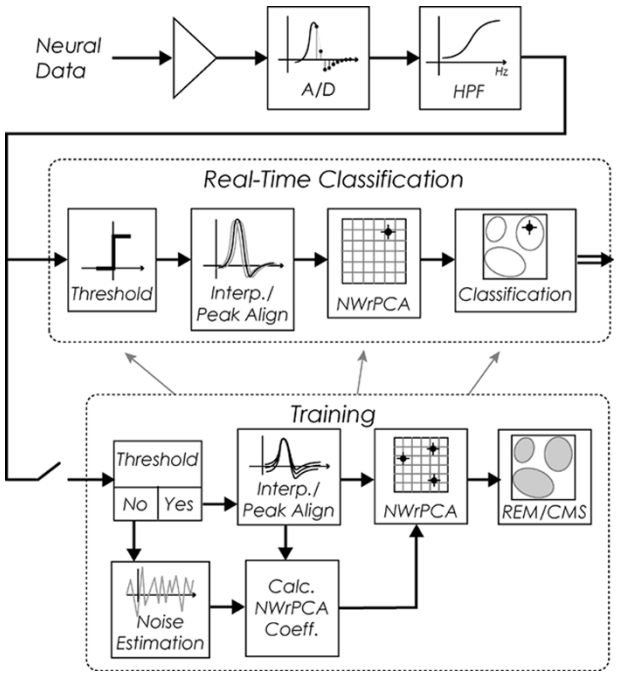


Fig. 2. Block diagram of the Sahani algorithm. Processing associated with one electrode is illustrated, such processing must be performed for all electrodes.

will be explained in the next paragraph, followed by a maximum *a posteriori* (MAP) classification for assignment to a particular neuron or background-noise class.

Parameter training requires a computational pathway separate from real-time classification. For training, a 1 min segment of the preconditioned data from a given electrode is used. The rms value of this data is calculated and the threshold used for event identification is set to three times this value. Using this threshold, spike events are identified and aligned as in real-time classification. In parallel, segments of signal which do not exceed threshold are used to estimate noise characteristics. This characterization of the background noise enables the projection of the spike event waveforms into a robust noise-whitened principal component analysis (NWrPCA) space. During training, relaxation expectation-maximization (REM) and cascading model selection (CMS) are then 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.

III. Power Estimates

In this section, we address the power consumption of the two major computational elements of a spike sorting system: analog to digital conversion (ADC) and digital training and classification.

A. ADC

The advanced spike sorting algorithms we will describe necessitate that the spike detection circuit transform the signal from the analog to the digital domain. On our electrode array (100-electrode silicon array, Cyberkinetics, Inc., Foxborough, MA), we observe a mean dynamic range of signal amplitude of 647 μV. With a mean rms noise level of 11 μV, this corresponds to a dynamic range to noise ratio (DNR) of about 35 dB. Given this level, a conservative requirement for analog-to-digital conversion (ADC) resolution would be 8 bits. Hence, the digitization process consumes a significant portion of the total power of the proposed system. As a result, in this section, we argue that, for the benchmark system of 100 channels, each with a bandwidth of 30 kHz and full-scale input-signal voltage equal to the ADC supply voltage, 100 μW is an achievable upper bound on ADC power consumption.

Many low-power ADC designs have been implemented in the past decade. For example, the successive approximation ADC of [8] could be arrayed, yielding a 100-channel ADC array with 7-bit resolution at 5 kHz bandwidth consuming 310 μW. Scaling this design using (1), as will be described in the next paragraph, we might expect a 100 channel, 30-kHz, 7-bit ADC in a 0.13-μm technology to consume 650 μW.

However, a theoretical consideration of analog to digital conversion suggests that significant reductions in energy consumption are possible. The power consumption of low frequency and moderate resolution converters (which typically use a successive approximation design), is constrained by V_{TH} (transistor threshold voltage) mismatch. If we size the components such that change in charge corresponding to the least significant bit is α times greater than the expected variation due to mismatch, as shown in [9], the energy per step size for each pair of matching critical transistors is bounded below by

$$\frac{\text{Power}}{2^N \text{BW}} = \pi\, \alpha\, q \sqrt{W L x_d N_a} \tag{1}$$

where N is the number bits, BW is the bandwidth, W, L, N_a, and x_d are process parameters (transistor width, channel length, total dopant concentration, and depletion depth), and q is the electron charge. Thus, the overall ADC energy per step size is bounded by

$$\frac{\text{Power}}{2^N \text{BW}} = N_{MCT} N_C \pi\, \alpha\, q \sqrt{W L x_d N_a} \tag{2}$$

where N_{MCT} is the number of pairs of matching critical transistors per comparator and N_C is the number of comparators.

As a relevant example, let us chose $\alpha = 10$, and assume $W = L = x_d = 0.13$ μm, $N_a = 10^{16}$ cm^{-3}, one comparator with 20 matching critical transistors, and that the ADC has a resolution of 8 bits at 30 kHz. This yields an energy per step size lower bound of

$$\frac{\text{Power}}{2^N \text{BW}} = N_{MCT} N_C \left(25 \times 10^{-6}\ \text{pJ}\right) = 5 \times 10^{-4}\ \text{pJ} \tag{3}$$

or

$$\text{Power} = 4 \times 10^{-4}\ \mu\text{W}. \tag{4}$$

So a lower bound on the ADC power consumption is about 0.4 nW, or 40 nW for a 100 electrode array.

The four orders of magnitude difference between this value and the result in [8] is largely due to assumptions in the derivation of the bound, including issues such as parasitic capacitance, process variation, process-voltage-temperature corners, and complications due to the required high magnitude of the

input signal. Recent development in low-power ADC design has leveraged the extremely power-efficient digital circuit, as will be discussed in the next section, to "aid" the analog design [10]. As a result, the power consumption of ADCs is expected to be reduced by an order of magnitude using these digital calibration and compensation techniques. Hence, a converter consuming close to 1 μW with 8-bit resolution at 30 kHz, or 100 μW for 100 channels, should be achievable.

B. Digital Power Estimation Technique

We estimated the power requirements of spike sorting algorithm described in Section II-B by recasting the operations performed to simple instructions that can be implemented in hardware. A detailed analysis of the algorithms was carried out and approximate figures for the number of operations (specifically adds and multiplies) required for each task were obtained. Operation counts for some complex linear algebra functions used in the algorithms, like matrix decompositions, were taken from standard texts on numerical linear algebra [11]. Operation counts were then translated to power using the figure 1 mW/GOPS [12]. This figure is used as the standard power consumption per operation for ASICs implemented in 0.13-μm CMOS technology. Finally, to approximate power usage from memory accesses, we simply double the power from instruction execution [13]. The figures should be taken as an "order of magnitude" indication. However, we believe that these figures are indicative of the power consumption, and thus achieve the objective of showing that these systems can be implemented in an implantable neural prosthetic.

While we take the Sahani algorithm as a benchmark against which other spike sorting algorithms can be measured, we also consider a vastly simplified algorithm for comparing power consumption. In training, we eliminated many of the key innovations of the algorithm, noise whitening, background event distributions, and cascading model selection, and clustered in traditional PCA space using the common K-means algorithm with three clusters. In real-time classification, a minimum, Euclidean distance metric is used to classify events into the K-means/PCA clusters. We refer to this simplified implementation as the K, means/PCA spike sorting algorithm.

C. Training Power Consumption

We assumed that training need be conducted only once every 12 h for each electrode, based on our rough estimate of the relative stability of signals on our electrode array [1], implying that for a 100-electrode array the training process for each electrode can be allocated 432 s. Furthermore, immediately prior to each electrode's training period, a 1-min segment of its recent, filtered neural data is stored in memory.

To obtain power estimates, it is simplest to fix some stochastic training parameters. As the number of operations for training depends on the number of threshold-crossings occurring within the data set, we assumed that on average the training data set contains 2000 such incidents. In addition, due to the fact that clustering and statistical fitting algorithms are iterative by nature, we assumed that convergence would be achieved within a maximum of 20 iterations.

TABLE I
OPERATION ESTIMATES FOR ALGORITHM TRAINING

Step	Sahani	K–means/PCA
RMS Calculations	2.40E+02	2.40E+02
Spike Alignment	2.67E+08	2.67E+08
Noise Covariance	2.56E+05	—
PCA	1.42E+07	3.22E+06
Model Training	3.29E+08	1.62E+07
Total	6.10E+08	2.86E+08

The number of required operations per electrode for various parts of the Sahani and K-means/PCA training algorithms are listed in Table I. We see that training using the Sahani algorithm requires approximately 6.10×10^8 operations per electrode. With our previous assumption of 432 s of training time per electrode, we can convert to a power number by the following expression:

$$\text{Power} = \frac{\text{total ops}}{\text{training time}} * \frac{1\ \text{mW}}{10^9\ \frac{\text{ops}}{\text{s}}}. \tag{5}$$

In order to account for memory accesses, we double the resulting power number. Therefore, if performed over 12 h, for 100 electrodes, the total power requirements for training the Sahani algorithm is approximately 2.8 μW, for K-means/PCA, approximately 1.3 μW.

D. Real-Time Classification Power Consumption

The classification process itself contributes relatively little to the overall power consumption of real time spike sorting. Most of the real-time computational burden is dominated by the high-pass filter, thresholding, and the interpolation for peak alignment of events which cross threshold.

We will assume that an IIR high-pass filter consisting of two second order sections is used. The coefficients we used in our simulations are the same as those of the 250-Hz-cutoff filter of the commercially available Cerebus rack-mounted spike sorting system manufactured by Cyberkinetics, Inc. Also, a 30-kHz sampling rate is assumed. The IIR filter necessitates 6.3×10^5 ops/s/electrode. Digital thresholding contributes 3×10^4 ops/s/electrode. Combined, these sections should contribute about 1.32 μW/electrode.

Let us assume a "worst–case classification complexity scenario," in which there are a total of 50 threshold crossing events per second, coming from up to five neurons. In this case, peak alignment (interpolating by a factor of 32 times), can be expected to consume about 6.6×10^6 ops/s/electrode, corresponding to 13 μW/electrode. Following this, maximum *a posteriori* classification requires about 4.8×10^4 ops/electrode/s, corresponding to 0.096 μW/electrode. A simplified classification, using only the minimum Euclidean distance to a cluster (i.e., the traditional technique used in concert with K-means/PCA), requires 1.3×10^4 ops/s/electrode. This corresponds to 0.026 μW/electrode.

In data from a typical day of recording on our array, we observe a mean of 20–40 threshold crossings per second per electrode. Hence, for neurons similar to those of the cortical region

in which our array is implanted (dorsal premotor cortex), the actual number of computations and related power consumption for real-time classification would be less than that described above.

As will be seen in the following section, it may be possible to avoid interpolation with little degradation of performance. If interpolation is eliminated, the biggest hurdle to be overcome in real time classification is the design of a power efficient high-pass filter for hundreds or thousands of simultaneous channels. The problem is made more difficult by the fact that the LFP is in the 0.5–100-Hz frequency range, while much of the signal power is concentrated in the 1000–3000-Hz range. With a sampling frequency of 30 kHz, the necessary transition band is somewhat steep. Also, the amplitude of the LFP can be as large as the amplitude of the most prominent spike waveform, so the stopband attenuation must be fairly significant. An analog filter with a bandpass characteristic equivalent to combining the high-pass filter we require with the anti-aliasing filter used for the ADC is presented in [14]. Interestingly, it consumes approximately 1 μW, a figure comparable to our all-digital approach. While digital processing benefits from process scaling, analog filters remain a challenge, as the large capacitors and resistors they require remain chip-area intensive. Hence, innovative solutions to the filtering problem may well involve novel digital processing.

IV. Power and Algorithm Performance

In order for an implantable spike sorting system to be a useful tool, it is important that it detect and classify spikes as accurately as a nonimplanted real-time spike sorting system. The purpose of this section is twofold: first, to briefly demonstrate what we believe to be the excellent performance of the Sahani algorithm relative to simpler real-time spike sorting techniques, and second, to address the impact of a few design tradeoffs on spike sorting performance.

Measuring the accuracy of a spike sorting system requires a data set in which the time and correct classification of neural spikes are known *a priori*. In the case of single electrode recordings, it is possible to experimentally gather such data by simultaneously recording from intra- and extra-cellular electrodes. However, for microelectrode arrays, an equivalent technique has not been demonstrated. Thus, one must create a synthetic data set which closely resembles experimentally observed neural activity. Several techniques for generating synthetic data to test spike sorting algorithms have been reported ([15], [16]). In this work, we use a simple three–component model fit to data experimentally observed on our microelectrode array.

A. Generation of Realistic Synthetic Data

Neural activity recorded from acute or chronically implanted electrodes varies tremendously depending on the recording technique. For example, the location and type of neurons being observed, the level of activity in the surrounding tissue, and the impedance characteristics of the electrodes themselves can result in large differences in signal quality, even among neurons observed on different electrodes in a single-microelectrode array. To capture this variability, we generate synthetic data to correspond to data recorded on each of the electrodes in our 96 electrode array (chronically implanted in a rhesus macaque monkey, see [1] for details), and use the inherent variation in number and quality of neural signals to probe performance characteristics.

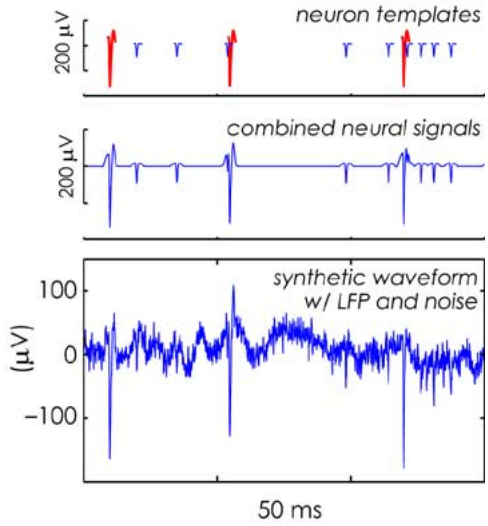


Fig. 3. Generating synthetic neural data. Mean waveform of each neuron detected on an electrode becomes a template for a synthetic neuron. Spike trains for each neuron are independently generated and added to the randomly generated LFP and colored noise waveforms. Two neurons shown above demonstrate that while spikes from a given neuron cannot overlap, those from different neurons can.

As depicted in Fig. 3, we modeled recorded signals as being composed of the sum of three components: LFP oscillations, colored noise, and action potentials. Thus, for a given electrode, we fit an autoregressive (AR) model to the LFP observed in a 2-min segment of experimental data. Additionally, we used a second AR model to emulate the colored noise remaining after the high pass filtering process. Finally, we used mean waveforms found by spike sorting the data as action potential templates. Using spline interpolation of the templates to introduce random jitter, we then placed spikes into the synthetic data. The number of spiking events for each neuron was proportional to the number observed in the experimental data. Spike times were chosen randomly, with the constraint that two spikes from a particular neuron could not occur within 1.2 ms; spikes from different neurons were allowed to overlap. A sample of recorded action potentials from one electrode of the array, and the synthetic spikes generated to simulate them are shown in Fig. 4.

Error rates were obtained by training the spike sorting algorithm on one set of synthetic data, and then classifying a second set of synthetic data generated using the same model. By averaging performance over synthetic data generated from each electrode on the array on a given day, we can evaluate the impact of performance tradeoffs on neural signals whose qualities (number of neurons, similarity of spike waveforms, etc.) vary in a way that is relevant to the environment of our microelectrode array.

B. Algorithm Performance Assessment

The most important characteristic of any spike sorting algorithm is its ability to accurately classify action potentials. How

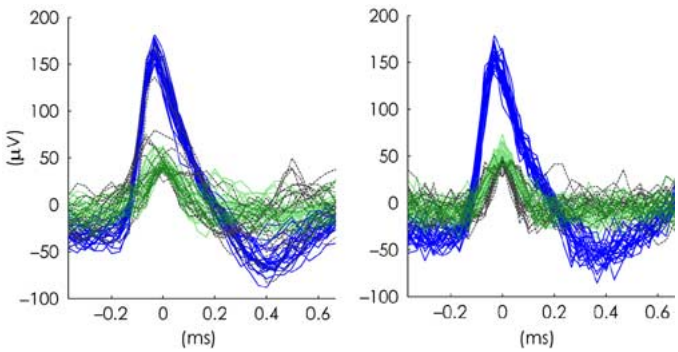


Fig. 4. Recorded and synthetic spikes. (Left) Sample of spikes recorded from a specific electrode, color coded by assignment to one of two neurons (solid lines) or noise (dotted lines). (Right) Synthetic spikes generated for the electrode. In the case of the recorded data (left), the color coding is generated by our initial spike sorting process. In the case of the synthetic data, the color coding corresponds to the known timing of generated spikes. SNR of the larger neuron is 15.2 and that of the smaller is 4.3.

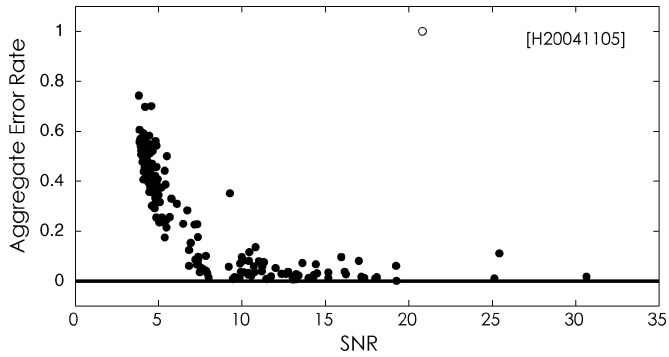


Fig. 5. SNR and classification error. Using a synthetic data set generated to match the noise and neural activity statistics of a real recording session, we evaluated the effect of spike SNR on the Sahani algorithm spike sorting error rate. Aggregate error rate is the sum of false positive and false negative classification errors normalized by the number of synthetic spikes generated. Hollow circle represents a neuron which was not successfully classified due to insufficient firing rate (0.3 spikes/s) during training (Data set H20041105).

well the algorithms perform this task depends on the characteristics of the signal, the number of neurons, and how different the waveforms are. We define the signal-to-noise ratio (SNR) of an action potential waveform as the ratio of the mean peak signal level to the standard deviation of the background noise measured using segments of data not containing spikes. That is

$$\mathrm{SNR} = \frac{\mathrm{E}\left[\max_t\{|s_k(t)|\}\right]}{\sigma} \tag{6}$$

where $s_k(t)$ is the waveform of the kth spike attributed to the neuron. During the recording session on which we modeled our synthetic data (H20 041 105), we classified, and thus generated synthetic data corresponding to 176 neural units on 95 electrodes (one electrode was unusable due to noise artifacts), with an average firing rate of 20 spikes/s per unit, and a mean SNR of 7.8 (18 dB).

In Fig. 5, the dots represent the aggregate error rate of spike sorting using the Sahani algorithm: the number of false positive and false negative classifications divided by the actual number of synthetic spikes generated for a given neuron. Error rates

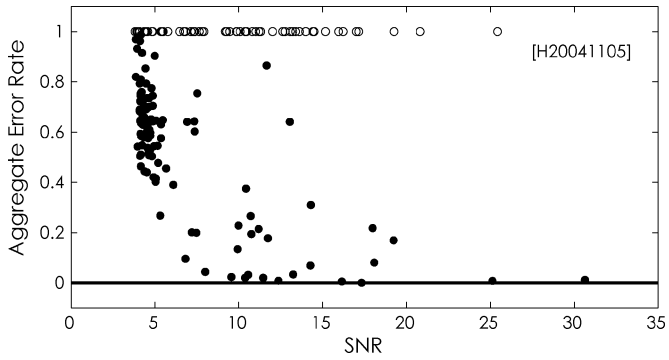


Fig. 6. K-means/PCA classification error. Performance of the K-means/PCA algorithm is shown using the format of Fig. 5. Missed neurons (51) or error rates of 100% or more (11) are depicted by hollow circles. Notice that in the case of SNR larger than 15, when a neuron is correctly identified, classification can be quite successful. Performance suffers greatly, however, for neurons with SNR in the range of 5 to 20.

were generated for each synthetic neuron generated as described above to imitate an actual neuron observed during the recording session. Notice that for SNR lower than about 7, the ability of the algorithm to correctly classify threshold-crossing events is severely hindered. Similar performance has been reported previously [15], [16]; intuitively, as shown in Fig. 4, it is quite difficult to visually distinguish between noise events and low SNR spikes. However, if we restrict ourselves to neurons with SNR greater than 7, "high SNR," the median false negative (i.e., missed spikes) classification error rate is 0.4% and the median false positive (i.e., misclassified spikes or noise) error rate is 3.0%. The median aggregate classification error rate for the high SNR neurons is 3.7%. While computationally intensive nonrealtime spike sorting systems may be able to achieve marginally better accuracy (for example, in cases in which errors are due to overlapping spikes), we have found in practice that the performance of the Sahani algorithm exceeds that of other real-time spike sorting algorithms available to us.

In fact, the Sahani algorithm offers many advantages over simpler real-time spike sorting algorithms. In comparing performance with the much simpler K-Means/PCA technique two critical aspects are made apparent. By using cascading model selection, the Sahani algorithm can typically determine the correct number of neurons on its own, which is a crucial feature of any unsupervised spike sorting algorithm. Furthermore, the mixing model approach provides a well founded technique for rejecting threshold-crossing events which do not actually correspond to neural spikes. As shown in Fig. 6, the lack of these two capabilities, as in the simple K-Means/PCA algorithm, is quite detrimental. Even at high SNR, the lack of automatic model selection results in many neurons being misclassified entirely, as denoted by the hollow circles. However, even if totally misclassified neurons, defined as missing more than half of the synthesized spikes, are excluded, for the remainder of the high SNR class, the aggregate median error rate rises to 20%. In addition to numerous noise-events, high SNR units are often present on the same electrode as other units (24 electrodes of 96). Thus, these deficiencies would also be a significant problem for a spike-sorting strategy which simply accepts all waveforms that cross threshold.

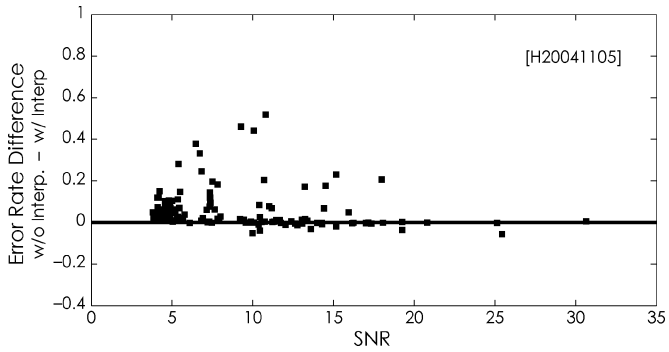


Fig. 7. Peak alignment and error rate. Using the synthetic data set, we evaluated the effect of peak alignment on the error rate of the Sahani algorithm. For neurons with SNR greater than 7, the median difference was 0.4%.

In situations in which many neurons are detectable by a given electrode, there is an additional advantage to using the Sahani algorithm. By projecting the data in the whitened noise principal component space, the Sahani algorithm maximizes the separability of the clusters. In other words, this projection has the greatest ratio of the average distance between clusters to the average spread of the data. It is, therefore, possible to separate clusters that would be indistinguishable in regular principal component analysis [7].

One of the critical observations that was gleaned from our simulations is that the computationally intensive interpolation used for peak alignment during real-time classification offers modest benefit. As shown in Fig. 7, at higher SNR, there is sometimes a noticeable difference: the median increase in error rate for high SNR neurons 0.4% (median error rate increases to 4.3%). However, eliminating interpolation during real-time operation nearly halves real-time power consumption. This is one example where performance/power analyzes can unveil important design tradeoffs.

V. Discussion

We have shown that currently available spike sorting algorithms can be both reliable and power efficient. With 100 electrodes, an upper bound of the power consumption of our spike sorting algorithm (without interpolation during real-time operation) is about 150 μW. Also, we have shown that the one hundred 8-bit, 30-kHz ADC needed for digital spike sorting are expected to consume less than 100 μW of power. Thus, 250 μW is an achievable level of power consumption for an implantable, 100 electrode digital spike sorting circuit. Assuming heat dissipation over a 16 mm^2 chip, we have a power to area ratio of about 1.6 $\mathrm{mW/cm^2}$, which is well below the 80 $\mathrm{mW/cm^2}$ chronic heat dissipation threshold believed to cause tissue damage [17]. By way of comparison, for 100 electrodes, the all-analog approach of [3] would require 5.7 mW, and the wavelet compression technique of [4] 120 mW. While these alternative approaches may benefit from voltage scaling, their respective drawbacks, a loss of information and less than ideal data compression, remain significant when compared with an implantable spike sorting paradigm. Furthermore, we have not considered the requisite low-noise amplifier in this report as all approaches to spike sorting require their use, and because recent reports have demonstrated low power (< 1 μW per channel) and noise (~ 2 μV) levels [14], [18].

In addition, we have found that the high pass filter stage dominates power consumption. Alternative methods of multichannel filtering for spike sorting should be investigated to further reduce the power consumption and improve performance as the number of available electrodes expands.

VI. Conclusion

We have shown that digital spike sorting is feasible using currently available algorithms and technology. As a result of the efficiency of low-power digital CMOS, digital spike sorting can, in fact, be achieved without significant increases in power when compared with simpler techniques. Furthermore, by using a very effective spike sorting algorithm, we can achieve data bandwidths comparable to those of a simple threshold-detector, while retaining essentially all the information content of the original, high-bandwidth signal. Thus, as implantable electrode arrays grow in density, we believe that the on-chip spike sorting approach will be a viable and attractive solution for neural prosthetic interfaces.

Acknowledgment

The authors would like to thank A. Afshar, S. Ryu, and B. Yu for collecting neural data used in this study, and M. Howard for expert veterinary care. The authors would also like to thank M. Linderman for helpful comments. Finally, the authors are grateful to M. Sahani for his generosity and helpful advice.

References

[1] G. Santhanam, M. Sahani, S. I. Ryu, and K. V. Shenoy, "An extensible infrastructure for fully automated spike sorting during online experiments," in *Proc. 26th Annu. Conf. IEEE EMBS*, San Francisco, CA, Sep. 2004, pp. 4380–4384.

[2] J. Carmena, M. Lebedev, R. E. Crist, J. E. O'Doherty, D. M. Santucci, D. Dimitrov, P. Patil, C. S. Henriquez, and M. A. Nicolelis, "Learning to control a brain-machine interface for reaching and grasping by primates," *PLoS Biol.*, vol. 1, no. 2, pp. 193–208, Nov. 2003.

[3] R. R. Harrison, "A low-power integrated cicuit for adaptive detection of action potentials in noisy signals," in *Proc. 25th Ann. Conf. IEEE EMBS*, Cancun, Mexico, Sep. 2003, pp. 3325–3328.

[4] K. G. Oweiss, D. J. Anderson, and M. M. Papaefthymiou, "Optimizing signal coding in neural interface system-on-a-chip modules," in *Proc. 25th Annu. Conf. IEEE EMBS*, Cancun, Mexico, Sep. 2003, pp. 3325–3328.

[5] S. I. Ryu, G. Santhanam, B. M. Yu, and K. V. Shenoy, "High speed neural prosthetic icon positioning," in *Soc. Neurosci.*, 2004, Program 263.1.

[6] G. Santhanam, S. I. Ryu, B. M. Yu, and K. V. Shenoy, "High information transmission rates in a neural prosthetic system," in *Soc. Neurosci.*, 2004, Program 263.2.

[7] M. Sahani, "Latent variable models for neural data analysis," Ph.D. dissertation, California Inst. Tech., Pasadena, CA, 1999.

[8] M. D. Scott, B. E. Boser, and K. S. J. Pister, "An ultralow-energy ADC for smart dust," *IEEE J. Solid-State Circuits*, vol. 38, no. 7, pp. 1123–1129, Jul. 2003.

[9] M. Pelgrom, "Low power CMOS data conversion," in *Low-Voltage/Low-Power Integrated Circuits and Systems: Low Voltage Mixed Signal Circuits*, E. Sanchéz-Sinencio and A. G. Andreou, Eds. Piscataway, NJ: IEEE Press, 1999, pp. 432–438.

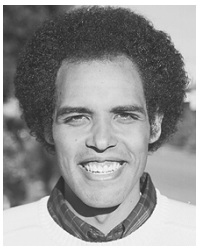
[10] B. Murmann and B. E. Boser, *Digitally Assisted Pipeline ADCs*. New York: Kluwer, 2004.

[11] G. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1983.

[12] A. P. Chandrakasan, S. Sheng, and R. W. Broderson, "Low power CMOS digital design," *IEEE Solid-State Circuits Soc. Quarterly Newsletter*, Apr. 2003.

[13] T. H. Meng, "Low-power signal processing system design for wireless applications," *IEEE Personal Commun. Mag.*, vol. 5, no. 3, pp. 20–31, Jun. 1998.
[14] T. Horiuchi, T. Swindell, D. Sander, and P. Abshire, "A low-power CMOS neural amplifier with amplitude measurements for spike sorting," in *Proc. 2004 IEEE Int. Symp. Circuits Syst.*, vol. 4, Vancouver, Canada, May 2004, pp. 29–32.
[15] 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, Jun. 2004.
[16] K. M. L. Menne, A. Folkers, T. Malina, R. Maex, and U. G. Hofmann, "Test of spike-sorting algorithms on the basis of simulated network data," *Neurocomputing*, vol. 44–46, pp. 1119–1126, Jun. 2002.
[17] T. M. Seese, H. Harasaki, G. M. Saidel, and C. R. Davies, "Characterization of tissue morphology, angiogenesis, and temperature in adaptive response of muscle tissue to chronic heating," *Lab Investigation*, vol. 78, no. 12, pp. 1553–1562, Dec. 1998.
[18] R. R. Harrison and C. Charles, "A low-power low-noise CMOS amplifier for neural recording applications," *IEEE Trans. Syst. Sci. Cybern.*, vol. 38, no. 6, pp. 958–965, Jun. 2003.

Zachary S. Zumsteg received the B.M. degree in music engineering from the University of Miami, Coral Gables, FL and the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 2005. He is currently working toward the M.D. degree at the University of California, Los Angeles.



Caleb Kemere (S'00) received the B.S. in electrical engineering with honors from the University of Maryland, College Park, in 1998 and the M.S. degree in electrical engineering, in 2000, from Stanford University, Stanford, CA, where he is currently working toward the Ph.D. degree in electrical engineering. His dissertation centers on optimally extracting information about target and trajectory from cortical neural activity recorded during goal-directed movements.

After arriving at Stanford, he worked in the Space Systems Development Laboratory and the Magnetic Resonance Systems Research Laboratory. Following a brief stint with Datapath Systems (now part of LSI Logic), San Jose, CA, he returned to Stanford University to pursue the Ph.D. degree.

Stephen O'Driscoll (S'01) was born in Cork, Ireland, in 1979. He 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, Stanford, CA, in 2005, where he is currently working toward the Ph.D. degree.

During the summers of 1999 and 2000, he was a Student Designer at Farran Technology, Ballincollig, Cork, Ireland, where he designed a 77-GHz radar front-end. From 2001 to 2003, he was an analog IC Design Engineer with Cypress Semiconductor, San Jose, CA, where he worked on the design of clock and data recovery PLLs. His current research focuses on low-power circuit design techniques for medical applications.

Mr. O'Driscoll was the recipient of 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.

Gopal Santhanam received the B.S. degree in electrical engineering and computer science and the B.A. degree in physics from the University of California, Berkeley, and the M.S. degree in electrical engineering, in 2002, from Stanford University, Stanford, CA, where he is currently working toward the Ph.D. degree.

His research involves 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.

Mr. Santhanam is the recipient of notable awards including the National Defense Science and Engineering Graduate fellowship and the National Science Foundation graduate fellowship.

Rizwan E. Ahmed received the B.S. degree in electrical engineering and physics from Virginia Commonwealth University, Richmond, in 2001, and the M.S. degree in electrical engineering from Stanford University, Stanford CA, in 2004.

He is now with the systems engineering group at Qualcomm Inc., San Diego, CA. His current interests are in position location techniques in wireless networks and novel communication systems.

Krishna V. Shenoy (S'87–M'01) 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 the Massachusetts Institute of Technology, Cambridge, in 1992 and 1995, respectively.

He was a Systems Neuroscience Postdoctoral Fellow in the Division of Biology, at California Institute of Technology, Pasadena, from 1995 to 2001. He joined the Faculty of Stanford University, Stanford, CA, in 2001 where he is an Assistant Professor in the Department of Electrical Engineering and Neurosciences Program. His current research activities include neurophysiological investigations of sensorimotor integration and coordination, neural prosthetic system design, and neural signal processing and electronics.

Dr. Shenoy is the recipient of awards and honors including the 1996 Hertz Foundation Doctoral Thesis Prize, a Burroughs Wellcome Fund Career Award in the Biomedical Sciences, the William George Hoover Faculty Scholar in Electrical Engineering at Stanford University, the Robert N. Noyce Family Scholar in the Stanford University School of Engineering, an Alfred P. Sloan Research Fellow and a Defense Science Research Council Fellow.

Teresa H. Meng (S'82–M'83–SM'93–F'99) received the Ph.D. degree in electrical engineering and computer science from the University of California, Berkeley, in 1988.

She is the Reid Weaver Dennis Professor of Electrical Engineering at Stanford University. Her research activities during the first 10 years at Stanford included low-power circuit and system design, video signal processing, and wireless communications. In 1999, she took leave from Stanford University and founded Atheros Communications, which delivers the core technology for high-performance wireless communication systems. She returned to Stanford University in 2000 to continue her research and teach. Her current research interests focus on circuit optimization, neural signal processing, and computation architectures for future scaled CMOS technology. She has given plenary talks at major conferences in the areas of signal processing and wireless communications. She is the author of one book, several book chapters, and over 200 technical articles in journals and conferences.

Dr. Meng has received many awards and honors for her research work at Stanford including an National Science Foundation Presidential Young Investigator Award, an Office of Naval Research Young Investigator Award, an IBM Faculty Development Award, a Best Paper Award from the IEEE Signal Processing Society, the Eli Jury Award from the University of California, Berkeley, and awards from AT&T, Okawa Foundation, and other industry and academic organizations. As a result of founding Atheros Communications, she was named one of the Top 10 Entrepreneurs in 2001 by Red Herring, Innovator of the Year in 2002 by MIT Sloan School eBA, and the CIO 20/20 Vision Award in 2002.