

Pulse-based signal compression for implanted neural recording systems

John G. Harris, José C. Principe, Justin C. Sanchez, Du Chen and Christy She
Computational NeuroEngineering Laboratory
University of Florida, Gainesville
Email:(harris, principe, justin, duchen, christy)@cnel.ufl.edu

Abstract—Today’s implanted neural systems are bound by tight constraints on power and communication bandwidth. Most conventional ADC-based approaches fall into two categories. Either they transmit all of the information at the Nyquist rate but are ultimately limited to only a handful of channels due to communication bandwidth constraints. Or they perform spike detection on the front-end which allows a scale up to 100 or more channels but prevents the use of spike sorting on the back-end. Spike sorting is an important step that provides a labeling to multiple neurons on each channel and further improves the accuracy of spike detection. In this paper we describe the pulse-based approach used in the FWIRE (Florida Wireless Implantable Recording Electrodes) project. A hardware spiking neuron on each channel is configured either to transmit pulses for full reconstruction on the back-end, or to transmit dramatically fewer pulses but still allow for spike sorting on the back-end. Spike sorting results show that the pulse-based spike sorting accuracy is competitive with conventional methods used in daily practice.

I. INTRODUCTION

There is a growing demand for wireless, low-power neural recording systems to amplify and transmit extracellular neural action potentials for either real-time processing or off-line analysis outside the body. A very important application that we are studying is that of brain-machine interfaces (BMIs) [1], which extracts information from neural recordings in real time with the goal of creating predictive models for hand movement and directly controlling a robotic device.

Current instrumentation technology and surgical procedures allow for recording from hundreds of electrodes at once, but the bottleneck is how to transfer the large bandwidth data streams without requiring the subject to be tethered with wires from the electrodes to a signal processing unit. Each channel may be sampled at 20 KHz with 12-bit samples leading to 240 Kbits/second bandwidth. Since a low-power, wireless link can transmit up to about 1Mbits/sec through the skin, only a handful of channels can be transmitted even if there are hundreds of channels available. From these considerations it is clear that signal compression is the key challenge for the future of these devices.

The difficulty in data reduction arises from the requirements of small size and low power for implanted circuitry. The size is constrained due to space on the subject and low power is necessary because of the difficulty of charging or changing implanted batteries and power dissipation over 80 mW/cm^2 has been reported to cause general tissue damage [2]. Many

research groups are attempting to build neural recording systems, and the overall power consumption, apart from the transmission, is manageable for hundreds of channels. Suitable integrated amplifiers dissipate on the order of 10 uW/channel when scaled to newer technologies [3] [4]. Successive approximation ADC technology can drop below 1 uW/channel for these sampling rates [5].

These problems are further compounded by that fact that in typical extracellular recordings a single electrode can record 3-5 neurons superimposed on the same signal. Thus, neuroscientists must rely on software strategies that must first detect the spikes and second perform spike sorting to classify the originating neuron for each spike. The sorting or classification process in the second step also serves to reduce the number of false spikes that are detected in the first step.

Previous solutions for these problems will be discussed in section 2. Section 3 then describes a novel pulse-based approach which allows full resolution of a select number of channels at an order of magnitude less bandwidth than conventional approaches. Section 4 describes a pulse-based feature extraction method which adds further dramatic data reduction while still allowing off-board spike sorting to be run on the pulse-based features. Section 5 concludes the paper by summarizing how the Florida Wireless Integrated Implantable Recording Electrodes (FWIRE) project [6] [7] is utilizing these pulse-based mechanisms.

II. CURRENT APPROACHES

The most popular solution to the bandwidth problem is to select and transmit as many fully sampled channels as possible (typically only one channel at a time out of the dozens available). A simplified spike detection algorithm is concurrently run on the remaining channels and either the spike times or the binned spike times are transmitted. These spike-detected channels are extremely low bandwidth since neurons fire at a maximum of 200 Hz or so, and even 5 superimposed neurons on a channel would provide a worst case 10 Kbits/sec/channel binary sequence to transmit. There have been systems of this type built or described by several groups including Michigan [8] and Utah [9]. Having at least one high resolution channel allows the user to sequentially set the spike detection parameters which observing the full resolution channel output. Binning the spikes, say by summing

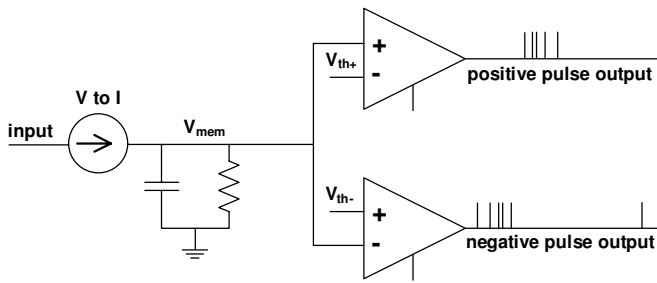


Fig. 1. Block Diagram of biphasic leaky integrate-and-fire neuron

the number of spikes over each 100ms window, further reduces the bandwidth.

However, there are several problems with transmitting spike-detected channels. First and most importantly, spike sorting can no longer be used to classify the originating neurons on each channel—this information is forever lost to the detriment of the full system performance. Spike sorting can only be run on the few full resolution channels transmitted. Second, on the remaining channels, spike detection must be run with low-power analog circuitry, which is difficult to achieve accurately. Finally, the spikes detection process is even more unreliable without spike sorting so there will be even more false detections and more missed spikes.

The alternative is to perform some sort of data reduction on all the channels and hope that transmitting the compressed signal will still allow enough information for spike sorting to be accurate. This sort of neural signal data reduction is a classical problem in neuroscience with many proposed algorithms in the literature. Popular data reduction methods include spike detection followed by different options to reduce the data. One option is to wirelessly transmit a clip of the raw waveform surrounding the spike for spike sorting outside the subject where power and size constraints are less stringent [10]. Another option is to extract and send the features themselves [11], but accurately computing and transmitting the necessary features at low-power is problematic. Both options suffer from complex circuitry and prohibitive power consumption for implantation with 100's of channels.

III. PULSE-BASED SIGNAL REPRESENTATION

Our lab previously proposed to encode the neural signal with a biphasic pulse train, which is amenable to low-power wireless transmission, to be reconstructed on the back-end and traditional spike sorting applied [12]. After amplification, a hardware integrate-and-fire neuron is used to convert a continuous-time signal into an asynchronous pulse train. A simplified version of the circuit is shown in Fig. 1. If the output of the integrator, $y(t)$ reaches the positive threshold of the comparator, θ , the output of the comparator raises and resets the integrator after a short delay, τ , in the feedback loop. Similarly, if the output of the integrator $y(t)$, reaches the negative threshold, $-\theta$, the output of the comparator drops and also resets the integrator. The leak term is some positive value

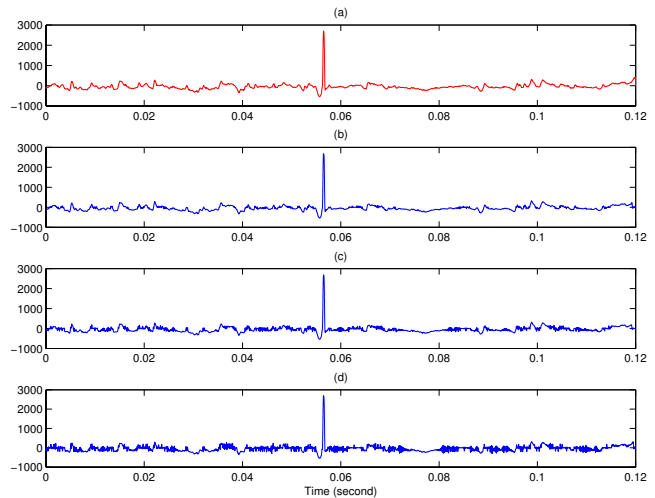


Fig. 2. Top plot shows original rat recording sampled at 25 kHz. Next 3 plots show reconstructed signals at 17.8Kpulses/sec, 9.2Kspikes/sec and 6.1Kspikes/sec

to filter out noise. The leak value sets the cutoff frequency for the low-pass filter formed with the integrator. The timing of two consecutive pulses must satisfy the following equation:

$$\int_{t_i+\tau}^{t_{i+1}} x(\Delta) e^{\frac{\Delta-t_{i+1}}{RC}} d\Delta = \theta_i \quad (1)$$

where $\theta_i \in \{-\theta, \theta\}$ and C is related to the integration capacitor and the R is related to the leak value.

Reconstruction algorithms are discussed elsewhere [12]–[14]. A sufficient condition for theoretically perfect reconstruction resembles the conventional Nyquist sampling assumption, namely that the reciprocal of the maximum inter-pulse interval be less than half the signal bandwidth. This constraint implies that perfect reconstruction of a 10KHz bandwidth signal requires at least 20Kpulses/second. Enhancing the neuron with a suitable refractory period, an adaptive firing rate or other schemes can reduce the required bandwidth further [15]. Fig. 2 (top) shows a plot of an original rat recording sampled at 25 kHz. The next plot shows the reconstructed signal using a 17.8 KHz pulse rate, creating a signal virtually indistinguishable from the original.

One of the appealing characteristics of the pulse sampling method is its ability to distribute more pulses in the areas where the signal amplitude is larger. Intuitively, this brings the ability to reduce data rates and impact differently the resolution of low and high amplitude signal features. Traditional sampling, without any further processing will affect equally the data dynamic range. Fig. 2 also depicts the compression effects for various IF pulse rates. The bottom three plots show the reconstructed signal using various rates achieved by varying the comparator threshold, i.e. the minimal time between two consecutive pulses. Pulse rates are 17.8KHz, 9.2KHz and 6.1KHz. As the pulse rate decreases, the distortion in the signal is seen primarily in the noise region and not in the spike waveform. Fig 3 shows a zoom in of the spike waveform

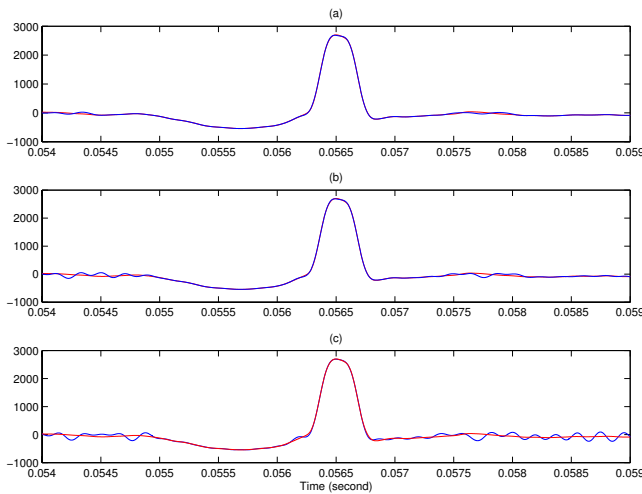


Fig. 3. Zoomed version of the spike for spike rates of 17.8Kpulses/sec, 9.2Kspikes/sec and 6.1Kspikes/sec

to show that the reconstruction is preserved in the neural spike regions. In this example, the 6.1 Kspikes/sec bandwidth compares very favorably with the 240 Kbits/sec needed by conventional systems.

IV. PULSE-BASED FEATURE EXTRACTION

We exploit the exact same leaky integrate-and-fire circuit shown in Fig. 1 to perform feature extraction [16]. If the leaky factor and threshold are appropriately set, then few or no pulses will fire in between neural spikes. At locations of neural spikes, many pulses will be generated and act as a spike “signature.” A template matching type algorithm can then be run off-board in lieu of conventional spike sorting to classify each spike. Full details of the pulse-based spike sorting algorithm are given elsewhere [16].

Instead of transmitting the raw neural waveform, the pulse-based feature extraction circuit encodes information about each spike in a biphasic pulse train. This greatly reduces the bandwidth required to transmit the spike trains especially because spike occurrences are sparse within neural data, while the pulse communication offers lower power transmission options. The encoding scheme uses pulses based on the area per time threshold of the waveform to represent the spike while the noise is mostly disregarded. Only the spikes and their time within the spike train contain information so not transmitting information about the noise saves power without any drop in system performance. Fig. 4 shows sorting error (described fully in [16]) versus minimum bandwidth at three different SNRs. This plots show the inverse relationship between bandwidth and sorting error.

The pulse-based feature extractor algorithm was tested with neural recordings from Bionic’s 128-channel hardware neural signal simulator. The use of a neural signal simulator allows the ground truths, the time of each spike and which neuron it came from, to be known. The neural simulator outputs a repeated 11 s pattern of spikes from three different action

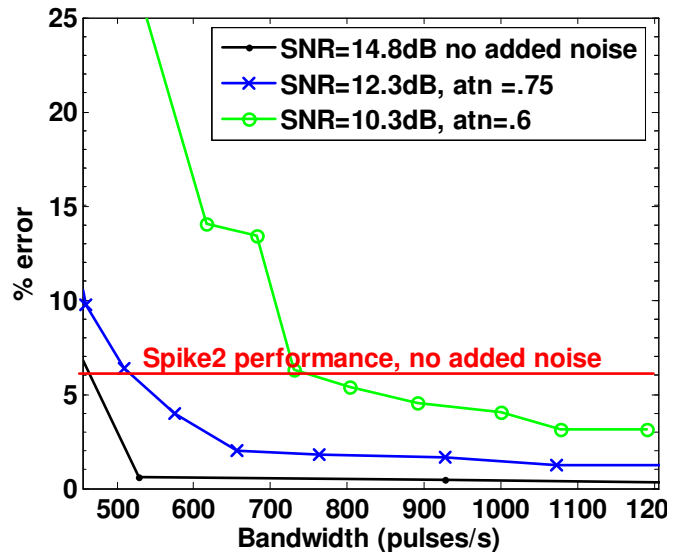


Fig. 4. Spike sorting error as a function of bandwidth for three SNRs.

potentials with amplitudes of $100 \mu\text{V}$ - $150 \mu\text{V}$ and a width of 1 ms. The interspike interval is 1s for 10s and then reduces to 10 ms for 1 s of burst firing. To increase the number of neurons on one channel the reference was chosen as another channel instead of ground. The referenced channel was carefully chosen to be a 5 ms delayed version of the first channel. In this manner, the simulated neural signal contains spikes from six different neurons with no superimposed spikes which are not addressed in this work since they are problematic for all spike sorting algorithms.

The UF bioamplifier [4], with a gain of 100, was used to amplify the neural simulator output. The amplified signal was then passed through our hardware leaky integrate and fire chip [16] and finally digitized at 24.4 KHz and 34.6 s were captured with a digital logic analyzer. The average spike firing rate for the data set is 19 Hz. The signal’s SNR is 14.8dB; a portion of the signal during bursting with all six neural spikes is shown in Fig. 5(A).

Spike2, a popular commercial program, which can spike sort offline, is used as a comparison to the feature extractor’s spike sorting performance. *Spike2* first performs general event selection by capturing windows around events that cross user defined thresholds. Then, spike sorting is performed with a combination of template matching and a PCA based cluster cutting. This process requires the user to select many parameters during the template setup such as the number of templates and allowable variation within the template. *Spike2* provides the user with an interactive visual display to assist in setting the spike sorting parameters. The parameters were set by an expert in the field with the same procedures used in typical experiments.

Matlab was also used to simulate the pulse-based feature extractor and its spike sorter. The LIF was set with a threshold and leakage value such that its spike sorting error was

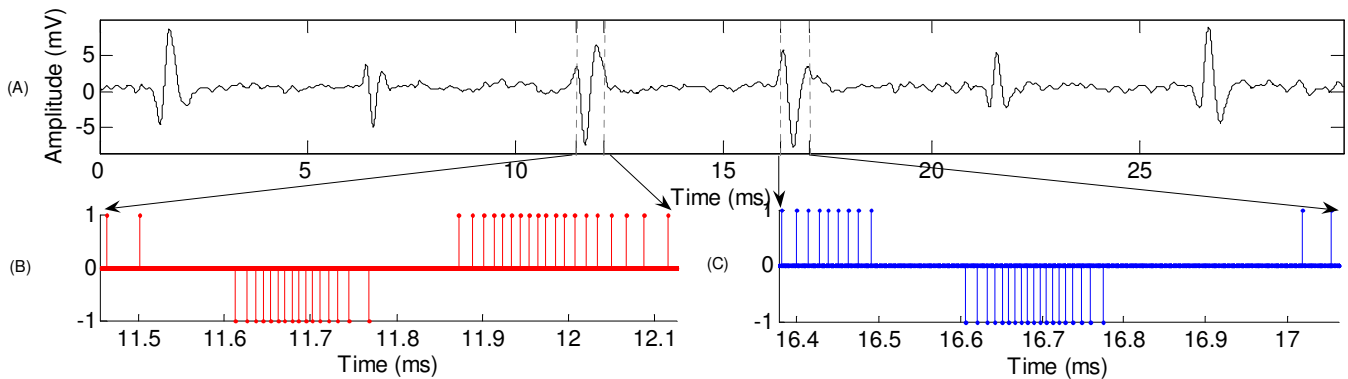


Fig. 5. (A) Neural simulator signal, after amplification, with all six neural spikes, one through six from left to right. The second row is the biphasic pulse train output from the LIF circuit with the bandwidth at 455 pulses/s. (B) Zoomed in spike three. (C) Zoomed in spike four.

similar to Spike2's which resulted in a bandwidth of 455 pulses/s. Fig. 5(B) shows examples of the biphasic output for spikes from two different neurons. The regions between spikes did not have any pulses. The biphasic output was then spike sorted. Overall at 455 pulses/s the feature extractor had 6.8% error compared with Spike2 which had 6.1% error. While maintaining a similar classification error to traditional sorting with Spike2, the feature extractor requires much less bandwidth with only 455 pulses/s compared to 300 Kbps for a traditional 25 KHz sampled signal at only 12-bits. 1 pulse/s is equivalent to 1bps. The pulse-based feature extractor can reduce its bandwidth even more if more sorting error can be tolerated or increase its bandwidth to lessen sorting errors. The two are inversely related. At 680 pulses/s the feature extractor actually outperforms Spike2 for this data set. More data simulations need to be performed across different SNRs and data sets to see if the trend continues, but one possible explanation is that the feature extractor preserves the important information in distinguishing between spikes while eliminating extraneous information. Extra information can make it more difficult for the neuroscientist to optimally set the spike sorting parameters in Spike2 making it harder to distinguish between spikes from different neurons.

V. CONCLUSION

The two pulse-based data compression mechanisms are being implemented in the FWIRE system. Since the hardware is identical for the two systems (signal compression and feature extraction), it is a simple matter to switch the circuit parameters to select one channel for high resolution transmission while still allowing the remaining channels to be feature extracted and spike sorted. Further experiments with real data are ongoing to determine the expected data rates more accurately.

ACKNOWLEDGMENT

The authors gratefully acknowledge funding from NINDS (Grant #NS053561) and NSF (Grant #0541241), and also the work of numerous other students in the lab, in particular J. Xu, M. Rastogi and V. Garg.

REFERENCES

- [1] J. Carmena, M. Lebedev, R. Crist, J. O'Doherty, D. Santucci, D. Dimitrov, P. Patil, C. Henriquez, and M. Nicolelis, "Learning to control a brain-machine interface for reaching and grasping by primates," *PLoS Biology*, vol. 1, no. 2, pp. 193–208, Nov. 2003.
- [2] T. Seese, H. Harasaki, G. Saide, and C. Davies, "Characterization of tissue morphology, angiogenesis, and temperature in the adaptive response of muscle tissue to chronic heating," *Lab. Invest.*, vol. 78, no. 12, pp. 1553–1562, 1998.
- [3] R. R. Harrison and C. Charles, "A low-power low-noise cmos amplifier for neural recording applications," *IEEE Journal of Solid-State Circuits*, vol. 38(6), pp. 958–965, 2003.
- [4] D. Chen, J. G. Harris, and J. C. Principe, "A bio-amplifier with pulse output," in *Int'l. Conf. IEEE Engineering in Medicine and Biology Society*, San Francisco, California, 2004.
- [5] H. Yang and R. Sarpeshkar, "A bio-inspired ultra-energy-efficient analog-to-digital converter for biomedical applications," *IEEE Transactions on Circuits and Systems*, vol. 11(53), pp. 2349–2356, 2003.
- [6] J. Sanchez, J. Principe, T. Nishida, R. Bashirullah, J. Harris, and J. Fortes, "Technology and signal processing for brain-machine interfaces," in *IEEE Signal Processing Magazine*, 2007, to appear.
- [7] R. Bashirullah, J. G. Harris, J. C. Sanchez, T. Nishida, and J. C. Principe, "Florida wireless implantable recording electrodes (fwire) for brain machine interfaces," in *ISCAS*, IEEE, 2007, pp. 2084–2087.
- [8] A. Sodagar, K. Wise, and K. Najafi, "A fully integrated mixed-signal neural processor for implantable multichannel cortical recording," *IEEE Transactions on BME*, vol. 54(6), pp. 1075–1088, 2007.
- [9] P. Watkins, R. Kier, R. Lovejoy, D. Black, and R. Harrison, "Signal amplification, detection and transmission in a wireless 100-electrode neural recording system," in *IEEE ISCAS*, Kos, Greece, May 2006.
- [10] I. Obeid, "A wireless multichannel neural recording platform for real-time brain machine interface," Ph.D. dissertation, Duke University, Durham, NC, 2004.
- [11] T. Horiuchi, T. Swindell, D. Sander, and P. Abshier, "A low-power cmos neural amplifier with amplitude measurements for spike sorting," in *International Symposium on Circuits and Systems*, May 2004, pp. IV 29–32.
- [12] D. Chen, Y. Li, D. Xu, J. Harris, and J. Principe, "Asynchronous biphasic pulse signal coding and its cmos realization," in *IEEE ISCAS*, Kos, Greece, May 2006.
- [13] A. A. Lazar and L. T. Tóth, "Time encoding and perfect recovery of bandlimited signals," in *Proc. ICASSP '03*, 2003.
- [14] A. A. Lazar, "Time encoding with an integrated-and fire neuron with a refractory period," Department of Electrical Engineering, Columbia University, New York, NY, Tech. Rep. BNET 3-03, Oct. 2003.
- [15] J. Xu and J. Harris, "The time derivative neuron," in *International Symposium on Circuits and Systems*, Seattle, WA, May 2008, submitted.
- [16] C. L. Rogers, J. G. Harris, J. C. Principe, and J. C. Sanchez, "A pulse-based feature extractor for spike sorting neural signals," in *Int'l IEEE EMBS Conference on Neural Engineering*, Kohala Coast, HI, May 2007.