ADAPTIVE DUAL-THRESHOLD NEURAL SIGNAL COMPRESSION SUITABLE FOR IMPLANTABLE RECORDING

*Russell Dodd, Bruce F. Cockburn**

University of Alberta
Electrical and Computer Engineering
Edmonton, AB T6G 2V4, Canada
email: rdodd|cockburn@ualberta.ca

Vincent Gaudet

University of Waterloo
Electrical and Computer Engineering
Waterloo, ON N2L 3G1, Canada
email: vcgaudet@uwaterloo.ca

ABSTRACT

This paper presents a digital architecture for neural signal compression using adaptive two-threshold spike detection and a nonlinear discrete wavelet coefficient selection scheme. The circuits and algorithms are described and compared with the state-of-the-art. The proposed 16-channel digital architecture is capable of neural data compression to 0.5% of the original raw data rate while consuming 21μW, with 30-kHz 8-bit sampling, in a 0.8-V 130-nm low-power IBM process.

1 Neural Interfaces

The pairing of implantable microelectrode arrays with micro-electronic devices allows researchers to interface with the nervous system in novel ways. Recent work [1], [2] has shown that by recording large populations of neurons, the resulting insights on firing patterns allow neuroprosthetic interfaces to be developed that restore motor and sensory function. The practicality of the interfaces in clinical practice depends on technological advances that overcome communication and power limitations and provide real-time operation. One of the main difficulties of real-time processing in prosthetic interfaces is the dynamic recording environment. In neural recordings from microelectrode arrays, the quality and quantity of neural signals depends on the position of the electrodes with respect to the active neurons. Over time the foreign body response and electrode movement can cause changes in signal-to-noise ratios (SNR) that result in the loss of neural sources. For neuroscience applications, real-time discrimination of single neural sources from all of the recorded signals and noise is critical. The technical challenge of discriminating neural sources is compounded by the constraints imposed by the low energy budget of battery-powered implanted devices.

1.1 Recording Systems

A generic implantable microelectronic system is illustrated in Figure 1. The device must respect several constraints that limit the potential processing complexity. The tissue surrounding the device can safely withstand only a small increase in temperature of up to 41.7 ± 0.9 C, which corresponds to a power dissipation of roughly 80 mW/cm^2 [3], [4]. In addition, the storage and wireless transmission of energy limits the average available operating power to 10s of milliwatts [5], [6]. Sampling rates of experimental recordings of neural action potentials are typically in the range of 15 kHz - 32 kHz with a pass band between 700 Hz and 7.5 kHz and 8-10 bits per sample. Therefore, on a single recording channel, we can expect a raw data rate of between 120 - 256 kbps. A microelectrode array may have up to 100 recording channels, which can impose a significant energy and data burden on the design of a very low-power implanted wireless recording device. If we consider a scenario in which the energy required to transmit information through tissue and air is greater than the energy required to compress the data in the implant, then adding signal processing prior to the radio frequency transmission can save power.

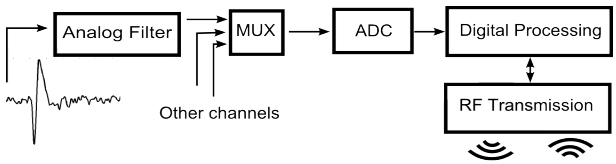

Fig. 1. Block diagram of generic neural recording system.

Fortunately, most of the useful neural information is believed to lie in the timing and membership of neural pulses [7]. A neural spike is typically 1-2 ms in length and has an amplitude of between 50-500 μV. Neural firing rates are limited by the ability of neural membranes to move ions and therefore neurons are unable to produce additional spikes within a refractory period. This refractory period is typically 1-2 ms after a spike. By capturing and processing only neural spikes, transmitted data rates can be reduced significantly.

For an implanted device to be useful for medical diagnosis and rehabilitation, spike detection, identification and

*Thanks to the AHFMR ProjectSMART Team for funding.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

classification must be performed adaptively in real-time. In previous work [8], we described a mixed signal system with single-threshold detection and compression through multi-level discrete wavelet transform decomposition. In this paper, we focus on new adaptive processing techniques intended for spike sorting that target a low-power implantable sensor. This paper proposes dual-threshold spike detection combined with a wavelet-based compression method. Section 2 reviews the literature on spike detection and presents our new approach. Section 3 describes the wavelet transform coefficient selection method. We present test and simulation results on adaptive detection and wavelet compression with experimental recordings in Section 4 and draw conclusions in Section 5.

2 Spike Detection

Spike detection is the single most effective way of reducing the amount of neural data to be processed. Triggered sampling is a common method used for spike detection. Its performance depends on the type of trigger and the SNR of the channel. The performance of a spike detection method can be measured by looking at its complexity or energy consumption, and by its false detection and true positive detection rates. In channels with a high SNR, a simple and effective trigger is the absolute value threshold [9]. In [10] the authors compare the spike detection performance of an absolute value threshold, a nonlinear energy operator, and a matched filter. In [9] the authors extend the comparison to include a stationary wavelet transform product. In both studies a single absolute value threshold is found to have good spike detection performance with low complexity; however, in [9] the authors advocate for the use of a nonlinear energy operator because of superior probability of detection for an acceptable increase in computational complexity.

The neural recording environment can change over time and static threshold crossings can give a large number of false detections. Therefore a spike detection method must be adaptable to the channel SNR. By utilizing noise estimation a threshold can be adjusted or a different spike detection method enabled. In [11], depending on the channel SNR, either an absolute value threshold or a stationary wavelet transform is used. References [12–17] describe various methods that utilize noise estimation to set appropriate threshold levels. A single-level threshold can be adjusted based on the type of noise expected in the channel. In [15] the threshold is based on the standard deviation of Gaussian-shaped noise. Alternatively, the threshold can be based on a small deviation away from the mean signal and then adjusted depending on feedback from spike classification [12]. A single threshold based on the root mean square (RMS) amplitude can be applied to both the peaks and troughs, similar to absolute value threshold detection, as described in [13].

2.1 Adaptive Two-Threshold Spike Detection

To reduce the number of false detections in a single-valued threshold detector, we adopted a two-threshold spike detector. One threshold is used for the leading peak and another for the subsequent trough. We enhanced the application of a two-threshold adaptive spike detector, similar to the concepts discussed in [14] and [16], but used an arithmetic scheme for the adaptive noise estimate from [14] and applied it to the two-threshold time window from [16] without the low-pass filter on the trough. An adaptive two-threshold detection scheme can detect a peak and/or trough and does not require them to have the same amplitudes. Additionally, a two-threshold scheme can require that both a peak and corresponding trough occur within a specified short time span, for example < 1 ms. The peak and trough thresholds are calculated based on an estimate of the noise envelope. Given a neural signal $s[n]$, broken into a set of windows of $w_k = s[n : n + W - 1]$, with $K \geq 0$ and W samples per window, the proposed new algorithm for the positive noise envelope calculation is:

Initialize the envelope estimate E_e to $Max_0 = max(w_0)$.
  if $Max_i > E_e$ then
    if $Max_i < (E_e +$ HIGH) then
      $E_e = E_e +$ INCREMENT
  else
    if $Max_i < (E_e -$ HIGH) then
      $E_e = Max_i$
    else
      $E_e = E_e -$ INCREMENT

The value of HIGH is chosen to ensure that windows with spikes are ignored, and INCREMENT is used to set the new envelope to be either a step up or down from the previous value, similar to the method in [12]. The thresholds are based on the envelope estimate plus a small offset, usually smaller than the value present in HIGH. In the case where spikes are frequently occurring in the amplitude space between noise and the HIGH offset, a reset is available for inactive periods since this kind of spiking may raise the envelope estimate and need to be quickly re-adjusted. The two-threshold detector can use positive or negative threshold only operation, but for recording tri-phasic neural signals, using the two thresholds can reduce the number of false detections, as shown in Section IV.

3 Waveform Compression

Once the possible spikes are detected they can either be transmitted directly in "raw" form or processed further to compress the transmitted spike information. The extra processing can try to capture the shape of the waveform more efficiently, or may extract the significant features necessary for spike classification (spike sorting). The extraction and mapping of features for spike sorting can be computation-

ally expensive and therefore careful consideration of the energy costs of the extra processing versus transmission is needed. It will be advantageous to reduce the amount of data by transmitting only the features required by the external classification method. In [16] the authors use PCA on peak amplitudes, trough amplitudes, and the width of the peaks. This sorter can achieve 90% correct classification. In [15] they use an offline expectation-maximization algorithm and in [18] they use wavelet footprints as features in PCA. The authors in [19] encode the peak and trough of spikes using a first-order piece-wise linear representation and clustering based on Euclidean distance. In [9] PCA, the discrete wavelet transform, the discrete derivative, and the integral transform are compared. The authors advocate discrete derivatives as a good choice due to a PCA-like classification performance at a lower computational cost. We use the wavelet transform.

3.1 Wavelet Transform

The discrete wavelet transform (DWT) can be considered a combination of a transform and subband decomposition [20]. The transform uses a half-band low-pass approximation filter and a half-band high-pass detail filter. The filter outputs are decimated by two at each level l of the transformation without losing any information. As noted in [11], the DWT suffers from shift variance, while the stationary wavelet transform (SWT) does not. The SWT utilizes the same filters as the DWT, but does not use decimation at each level. As proposed in [11], using the SWT in low-SNR regimes can help identify spike detections that triggered sampling might otherwise miss; however, we propose to use the DWT to compress the signal for all detections since our peak locations are determined by the two-threshold detection algorithm.

The performance of a transform for lossy compression can be measured with the mean squared error (MSE) of the reconstruction, shown in Eq. 1, and by the number of coefficients needed to represent the shape.

$$MSE = \frac{1}{N}\sum_{N}(x[n] - \hat{x}[n])^2 \tag{1}$$

As shown in [21], the symlet4 basis is an excellent wavelet basis for representing neural spikes. Using this basis we can apply coefficient thresholding and the knowledge of the most important coefficients within the windows of triggered samples to further compress neural signals.

4 System Architecture

A simplified system architecture is shown in Figure 2. The spike detection can be performed using positive-only, negative-only, or dual-threshold detection. A delay buffer is necessary to preserve the samples seen just before the first threshold crossing. The combination of a system-wide controller and channel buffer allows the spikes to be sent to the wavelet engine on an activity basis. The detected spikes are compressed by selecting K coefficients from a 3-level wavelet engine. The 3-level DWT implemented here uses a finite impulse response (FIR) filter structure [8]. The coefficients that are above the threshold have known locations within the window, so only those amplitudes need to be transmitted. The sample locations are chosen based on a training phase or previously recorded signals. The K largest coefficients from each spike are found for the observed spikes. The K coefficient locations that occur the most often are selected for transmission. The value of K is chosen to ensure reconstructions of triggered detections with a desired mean squared error (MSE). A K value of 10 can yield an average MSE of 50 over all of the channels, which creates visually accurate reconstructions of spikes. The coefficients selected are stored in a programmable register and can be provided on a per-channel basis.

5 Simulations and Results

This section contains simulation results and tested silicon results using experimentally recorded neural signals from the dorsal root ganglion of a feline and simulated data sets as described in [22]. The fabricated 2x2 mm chip uses the low-power, high-V_t, 130-nm IBM process. It contains a test channel and a separate 16-channel system. The wavelet compression is used in the next generation of the chip.

5.1 Two-Threshold Triggered Sampling

Table 1 shows simulation results on the data sets from [22]. The two-threshold detector's operational mode is set to positive-only threshold in column 'Pos', negative-only threshold in column 'Neg', and operates regularly in column 'Both'. The column 'Thr' indicates the single threshold of $4\sigma_N$ as described in [22]. The adaptive dual-threshold detector improves upon the static single-threshold detector in spike trains that have strong tri-phasic pulses by significantly reducing the number of false detections by 77%, which reduces the power consumption of the implanted device and improves the compression performance. However, the disadvantage in this version is that spikes with large positive peaks, but with very small negative peaks that are below the estimated threshold, are missed.

Table 1. Comparison of spike detection methods

	Missed Detections				False Detections			
Data	Pos.	Neg.	Both	Thr.	Pos.	Neg.	Both	Thr.
Sim1	0	60	1	0	1351	204	462	1340
Sim2	0	324	4	0	1279	246	419	1246
Sim3	0	700	179	0	1340	127	399	1328
Sim4	0	548	327	0	1290	114	383	1281
Sim5	0	4	0	0	1296	148	427	1279

Table 2 compares the performance of the two-threshold detector with the single threshold on a experimental recording with scaled noise. The test signal was created by piecing

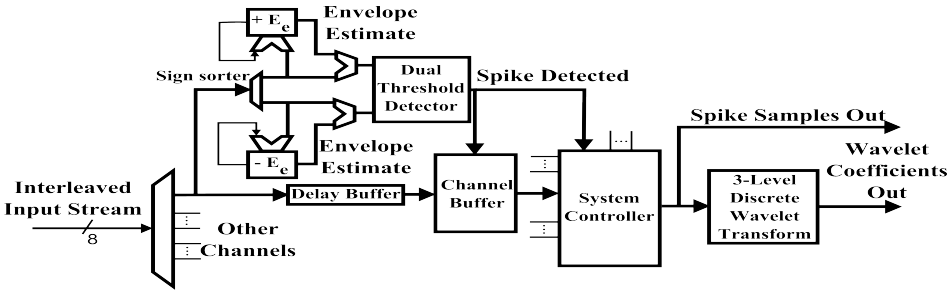

Fig. 2. Proposed system architecture with dual threshold spike detection and 3-level wavelet compression.

together recorded spikes and noise with the spikes having a mean of -0.57 μV and a standard deviation of 63.9 μV over 650 spikes. The noise is scaled by a factor A. At A=1, the noisy segments have an average of -0.007 μV and standard deviation of 8.04 μV. The number of false detections in the dual-threshold detector is reduced by 50% in the three scenarios; as well the number of true positives found is increased compared to the single threshold.

Table 2. Detection comparison with noise scaled recording

	True Positive %		False Positive %	
A	Dual	$4\sigma_N$	Dual	$4\sigma_N$
1	99.0	98.6	1.3	2.7
2	89.1	82.0	7.2	14.4
3	81.2	74.3	14.6	21.8

The power dissipation of single channel detector is shown in Table 3. The static power of high-V_t CMOS dominates the dynamic power. Future design iterations of the detectors should consider utilizing power gating or other technologies for mitigating leakage and static current in future iterations of the detector.

Table 3. Synthesized and measured power for one channel

	Synthesized	Measured	
VDD	1.08V	1.08V	0.79V
Static	53.18 μW	55.68 μW	7.97 μW
Dynamic	127 nW	189 nW	88 nW

5.2 Wavelet Compression

A triggered interval, depending on the sampling frequency and interval duration, can contain 15-60 samples. The compression method needs to accurately represent the waveform using fewer bits. A typical channel setting $K = 10$ resulted in a MSE of 47.6 for all detections. The resulting compression, considering a 48-sample window, is 21% of the original size. The reconstructions are shown in Figure 3.

The post-synthesis simulation results indicate a power dissipation of 454 μW, with 2.34 μW for dynamic power, at a clock frequency of 30 kHz with 8-bit samples at 1.2 V in 130-

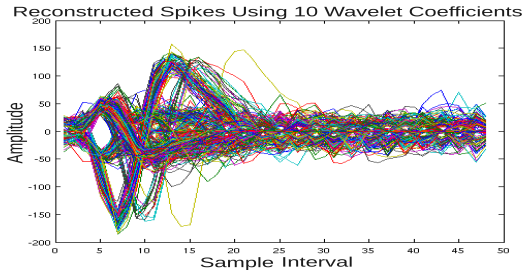

Fig. 3. A plot of reconstructed detections with $K = 10$.

nm low-power IBM technology. The pre-layout synthesized logic area is 250975 μm^2. Once again, the static power dissipation dominates as the main source of power dissipation. The 16-channel digital block, without wavelet compression, has a synthesized logic area of 0.8755 mm^2 with a power dissipation of 20.95 μW at 0.8V, which is comparable, on a per channel basis, to other work like [11] with an area of 0.082 mm^2 and power dissipation of 0.450 μW, and [19] with a complete channel area of 0.16 mm^2 and power of 3.1 μW. The digital system presented has an adjustable degree of compression and corresponding power cost as we can vary the number of wavelet coefficients and vary the thresholds to avoid more false detections in high SNR channels. In addition, the adaptive two-threshold scheme offers clear performance advantages over nonadaptive methods, so we are currently implementing an implantable prototype.

6 Conclusion

This paper presents new architectures for an adaptive real-time two-threshold neural spike detector and wavelet-based compression with an adjustable compression-accuracy trade-off. The test results indicate an implantable digital neural recording processor would dissipate an estimated 1.3 μW per recording channel while achieving 1.8% of the raw data rate using two-threshold triggered spike detection. For an additional 28 μW, per channel, a 0.4% of the raw data rate is obtained using a wavelet transform with nonlinear coefficient thresholding. The static power dissipation dominates the circuit performance and future versions will include techniques beyond utilizing a low-power, high-V_t, process to further reduce power dissipation.

7 References

[1] R.B. Stein, V.K. Mushahwar, "Reanimating limb movements after injury or disease," *Trends in Neuroscience*, vol. 28, pp. 518-524, 2005.

[2] Y.G. Gerasimenko, et al., "Novel and direct access to the human locomotor spinal circuitry," *J. of Neuroscience*, vol. 30, no. 1, pp. 3700–3708, Mar. 2010.

[3] R.R. Harrison, "A low-power low-noise CMOS amplifier for neural recording applications," *IEEE J. of Solid-State Circuits*, vol. 38, no. 6, June 2003.

[4] T.M. Seese, H. Harasaki, G.M. Saidel, and C.R. 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.

[5] K.M. Silay, C.D. Dehollain, M.D. Declercq, "Inductive power link for a wireless cortical implant with two-body packaging," *IEEE Sensors J.*, vol. 11, no. 11, pp. 2825–2833, Nov. 2011.

[6] H. Jiang, J.M. Zhang, S.S. Liou et al., "A high-power versatile wireless power transfer for biomedical implants," *IEEE Intl. Conf. on Engineering in Medicine and Biology*, pp. 6437–6440, Sept. 2010.

[7] M. Lewicki, "A review of methods for spike sorting: the detection and classification of neural action potentials," *Network: Computing Neural Systems*, vol. 9, pp. 53-78, 1998.

[8] R. Dodd, B. Crowley, V.G. Vaudet, V. Mushahwar, B.F. Cockburn, "Microelectronics for in-vivo neural recording," *Intl. Functional Electrical Stimulation Society (IFESS) Conference*, pp. 42–45, Sept. 2012.

[9] S. Gibson, J.W. Judy, D. Markovic, "Comparison of spike-sorting algorithms for future hardware implemementation," *IEEE Engineering in Medicine and Biology Society (EMBS) Conf.*, pp. 5015–5020, Aug. 2008.

[10] I. Obeid, P. Wolf, "Evaluation of spike-detection algorithms for a brain-machine interface application," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 905-911, June 2004.

[11] Y. Yang, A. Kamboh, A.J. Mason, "Adaptive threshold spike detection using stationary wavelet transform for neural recording implants," *IEEE Biomedical Circuits and Systems Conf. (BioCAS)*, pp. 9–12, Nov. 2010.

[12] C. Peng, P. Sabharwal, R. Bashirullah, "An adaptive neural spike detector with threshold-lock loop," *IEEE Intl. Symp. on Circuits and Systems (ISCAS)*, pp. 2133–2136, May 2009.

[13] S. Barati, A.M. Sodagar, "Discrete-time automatic spike detection circuit for neural recording implants," *Elects. Lett.*, vol. 47, no. 5, Mar. 2011.

[14] L. Traver, C. Tarin, P. Marti, N. Cardona, "Adaptive-threshold neural spike detection by noise-envelope tracking," *Electronics Letters*, vol. 43, no. 24, Nov. 2007.

[15] P.T. Watkins, G. Santhanam, K.V. Shenoy, R.R. Harrison, "Validation of adaptive threshold spike detector for neural recording," *IEEE Eng. in Medicine and Biology Soc. (EMBS) Conf.*, pp. 4079–4082, Sept. 2004.

[16] A. Bonfanti, T. Borghi, R. Gusmeroli, G. Zambra, A. Oliyink, L. Fadiga, A.S. Spinelli, G. Baranauskas, "A low-power integrated circuit for analog spike detection and sorting in neural prosthesis systems," *IEEE Biomedical Circuits and Systems Conf. BIOCAS*, pp. 257–260, Nov. 2008.

[17] T. Borghi, A. Bonfanti, G. Zambra, R. Gusmeroli, A.S. Spinelli, "A power-efficient analog integrated circuit for amplification and detection of neural signals," *Eng. in Medicine and Biology Soc. (EMBS)*, 2008.

[18] K.Y. Kwon, K. Oweiss, "Wavelet footprints for detection and sorting of extracellular neural action potentials," *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 609–612, May 2011.

[19] A. Rodriguez-Perez, et al., "A Low-Power Programmable Neural Spike Detection Channel With Embedded Calibration and Data Compression," *IEEE Trans. on Biomedical Circuits and Systems*, vol. 6, no. 2, pp. 87–100, Apr. 2012.

[20] M. Mandal, Multimedia Signals and Systems, Kluwer, 2003.

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

[22] J. Martinez, C. Pedreira, M.J. Ison, R.Q. Quiroga, "Realistic simulation of extracellular recordings," *J. of Neuroscience Methods*, vol. 184, pp. 285–293, 2009.