Optimizing Signal Coding in Neural Interface System-On-a-Chip Modules

Karim G. Oweiss[1,*], David J. Anderson[**], Marios M. Papaefthymiou[**]

[*]ECE Dept., Michigan State University, East Lansing, USA

[**]EECS Dept., University of Michigan, Ann Arbor, USA

ABSTRACT

With the recent advances in mutlichannel microprobe neural recording technology, and the limitations of the associated processing and communication capabilities, optimizing the information transfer from multiunit neural recording devices becomes strongly motivated to better understand the underlying mechanisms of neural information processing in the central nervous system. We recently developped a framework aimed at processing neural signals with optimized computational complexity. In this work, we propose coding solutions aimed at implementing fast and efficient neural interface System-On-a-Chip (SOC) modules capable of exchanging neural information without compromising issues of communication bandwidth and signal fidelity. Preliminary results demonstrate that neural information processing and coding *on-chip* offers tremendous savings in communication costs compared to raw data transmission for *off-chip* analysis. Moreover, we demonstrate that bit rates in the order of 1bit/sample can be easily achieved with a 32 channel device and 25kHz sampling rate. Performance illustrations and experimental neural data examples are described in details.

Keywords: multichannel recording, array processing, wavelet transform, signal coding

I. Introduction

Signal and information management of neural recordings from many closely spaced recording and stimulation electrode arrays in the brain poses many challenges in excess of those seen by the current generation of neural prosthetic devices such as cochlear implants [1]. Among these are: the shear volume of the data, variable signal quality across members of the array and vulnerability to external Electro-Magnetic Interference (EMI) sources. In a static laboratory environment, with large bandwidth to the recording devices, data processing can be managed off-line using massive computing power, and human interaction at a rate slower than *real-time*. On the other hand, implantable/wearable and hopefully unobtrusive neural implant devices with high channel count interacting at the single neuron or cluster level of detail requires an *ultra* high bandwidth communication link. Therefore, pre-processing neural signals is sought to be implemented as close as possible to where the signal is acquired to reduce the computational and communication costs associated with neural implants[2]. This implies that recording and stimulation of multiple neural cell populations needs to be implemented with limited power, bandwidth and in an autonomous manner. Consequently, there is an intrinsic need to develop neural interface *SOC* modules that

---

[1] *Corresponding Author*

[2] Since neural signals mostly consist of short rapid transients, a rough calculation of the *useful* time where a neural signal is detectable in the acquired data can be calculated to yield a range between 10% (spontaneous activity) to 25% (stimulus driven activity) of the total recording time. This roughly means that 75%-90% of the time, the telemetry system is used to transmit noise data not useful to the offline analysis.

optimize the information transfer from brain regions under investigation without compromising issues of bandwidth, detection, and classification accuracy.

Analysis of basic information rates from groups of nerve cells from a multichannel recording device has sparked the need for a coding solution that will allow the transfer of the information across the communication channel without significant loss. As an example, a 96 recording electrode device sampled at 25 kHz per channel leads to an aggregate rate 2.4Msamples/sec. At 12 bits/word, the bit rate is nearly 29Mbit/sec, which is well beyond the reach of present day biotelemetry developments. Nevertheless, transmission of 1-2 Mbit/second over a telemetry link from an implant to an extracutaneous device is feasible with the current technology [2,3], which yields a telemetry channel capacity at one bit per sample. This requires a smart way of compressing the basic information in the neural data to adhere to this capacity limit. Basic information in multiunit neural data consists of spike event timing and spike waveform classification result. To transmit this information within the available channel capacity, the coder may require large amounts of computation and probably time delays. Accordingly, an appropriate compromise among power, required bandwidth, computational load, memory, and delay time is needed to yield the best reliable design with highest signal fidelity. These parameters are linked in a very complex way and an optimum solution is beyond the scope of this effort. Nevertheless, some facts listed below can relax some of the parameter constraints to ease the design approach:

1. 32 physical channels from a closely-spaced device is a manageable number of interconnects and signal conditioning components in the volume over the tissue being considered [4].
2. A neural prosthesis system can tolerate delays of *several* milliseconds [5].
3. Several hundred thousand transistors can be operational in a package the size of a standard 14mm burr hole used by neurosurgeons.
4. 25kHz samples/sec quantized to 8 bits is a usual and logical sampling scheme if scaling is controlled.

The following sections will address how to optimize the processing architecture to deliver both the timing and waveform information within the one bit per sample objective.

II. Neural Signal conditionning

Due to the nature of the neural signal environment, some conditioning needs to be performed before coding the signal can take place. A recent array processing framework relying on the compaction property of the Discrete wavelet Transform (DWT) [6] has shown great premise in overcoming many challenges in processing neural signals with simple computational blocks [7]. In this context, it was shown that spatial filtering implemented either in the time domain [8] or the multiresolution domain [9,10] greatly reduces the correlated noise usually observed in closely spaced electrode site recordings. Therefore, a key processing block for signal conditioning is a spatial filter that performs *prewhitening* of the neural data to aggregate the energy spread across many physical channels to a few neural channels that can be individually filtered and coded in subsequent processing stages. The block diagram in Figure 1 shows the architecture of the signal processing and coding system considered. The schema utilizes standard components that have been implemented on either the Michigan merged micro machining and CMOS processes [4], a mixed signal chip or standard digital electronics.

To gain some insight on the theoretical side, the multichannel neural data can be modeled using the additive noise array model expressed as

$$Y = X + Z \tag{1}$$

where $Y \in \Re^{MxN}$ denotes N time samples of an array of M physical microprobe channels, $X \in \Re^{M \times N}$ the *noise-free* matrix of observations to be detected (linear mixture of multiple neural sources) and $Z \in \Re^{MxN}$ denotes a zero mean additive Gaussian noise component with a nontrivial spatial covariance matrix $R_Z \in \Re^{M \times M}$. The Maximum Likelihood (ML) estimate of the spatial covariance matrix of Y is expressed as

$$R_Y = E[Y \ Y^T] \cong \frac{1}{N\text{-}1} \sum_{n=1}^{N} y[n]y^T[n] \tag{3}$$

where $y[n] \in \Re^{Mx1}$ denotes the nth snapshot of the array. Using Eigen-Value Decomposition (EVD) [11], R_Y can be spectrally factored to yield

$$R_Y = U_Y.D_Y.U_Y^T \tag{4}$$

where the orthonormal *MxM* matrix U_Y comprises the eigenvectors spanning the column space of R_Y, and the *MxM* diagonal matrix D_Y contains the rank ordered eigenvalues of R_Y, $\delta_1 > \delta_2 > ... > \delta_M$. The matrix U_Y corresponds to the spatial filter that prewhitens the data matrix Y, i.e.,

$$\widetilde{Y} = U_Y^T.Y \tag{5}$$

Figure 1- The salient features of the architecture are alternate cross channel processes (site selection, spatial filtering, and block coding) and per channel processes (amplification, scaling, A/D and digital filtering). Depending the complexity of the overall system chosen, more or less of the blocks will exist.

consists the matrix of *focused* physical channels that spatially compacts the neural signal energy into a fewer number of channels. These are subsequently subband filtered using the DWT operator W_j in the jth subband as

$$\widetilde{Y}_j = W_j\{\widetilde{Y}\} \quad j = 0,1,...,J \tag{6}$$

where J denotes the total number of subbands in the DWT stage [10]. Equation (6) yields a temporal energy compaction into a few sparse coefficients. The thresholding property of the DWT enables discarding small coefficients, presumably attributed to noise and very weak signals, thus reducing the amount of information that will be pipelined to the coder at the last stage as illustrated in Fig.1. Denoting the thresholding matrix operator by H, the thresholded wavelet expansion can be expressed as

$$\overline{Y}_j = H\{\widetilde{Y}_j\} \quad j = 0,1,...,J \tag{7}$$

Equation (7) describes the prewhitened, thresholded wavelet coefficients that need to be coded for transcutaneous transmission from the neural interface *SOC*.

III. Coder Optimization

At this stage, two major alternatives exist for compressing the neural data for transcutaneous transmission:

1- Transmission of the processed version of the raw data given in (7) using a lossy compression scheme. This is contingent upon having a large bit rate to transmit event timing with full waveform data that can be classified *off-chip*,
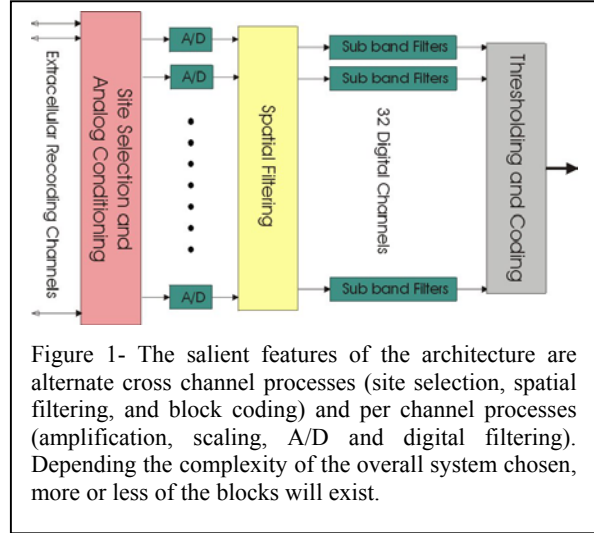
2- Detection of neural events in the processed data in (6), and transmitting only the event times across the channel with some side information about classification [12]. This is contingent upon sophisticated detection and classification algorithms *on-chip,* and yields very low transmission rates.

For the purpose of this paper, the former was chosen because the range of compromise is greater, that is, it offers the transmission of evoked potentials in parallel with the spike data with no special handling and it can take better advantage of information distributed across channels.

IV. Methods and Performance

A 32 channel device requires a spatial filter U_Y with 25×10^6 multiplication per second for prewhitening the energy of distributed spike to a single neural channel. This intensity of computation requires partitioning the 32x32 matrix down to a more manageable load. As an example, choosing eight 4x4 spatial filters reduces the computation to only 3.2×10^6 multiplications per second. This could provide local prewhitening over electrodes that are closely spaced but would not provide any global effects such as artifact reduction or evoked potential extraction at the sensor level. These signals could be derived with a simple row partition and canceled out of all the channels by subtraction. This process could be accomplished by analog referencing as well. This is out of the scope of this paper and is currently under investigation. As far as wavelet basis choice, we have gained experience processing neural data with several wavelet families and have determined that for this compression problem, the symlet4 family [6] carried down 4 decomposition levels provides satisfactory results and has a modest computational cost of about 2.2×10^6 multiplications/sec.

Waveform coding

Figure 2 below shows more details of the coding algorithm. The data are blocked by 16 samples per channel. That will yield one A4 sample/channel/block. The D1 band is disregarded because at a sample rate of 25kHz the spike train is over sampled leaving only noise in the D1 band. There are three stages to the coding process:

Figure 2- 16 samples from each of 32 channels are encoded using a wavelet decomposition, hard denoising algorithm [9] and run length coding. The resulting compression transmits data at just above 1 bit per sample.

1. First the wavelet coefficients are grouped across channels in bands with the A4 and D4 bands being combined because they are small and have similar statistics.
2. Next the coefficients are thresholded at levels determined by the noise power on each band [9].
3. Lastly the zero run lengths and the remaining coefficients are coded respectively by count and sign/magnitude with a simple binary quantizer.

The coding efficiency achieved is just slightly more than one bit per sample, which is well within the channel capacity. The minimum delay is .64 msec. which is also well within the tolerance for neuroprosthetic devices. Increasing the buffer length to more blocks of 16 samples would tend to average the bit rate over modulated signals bringing the maximum down and slightly increasing the
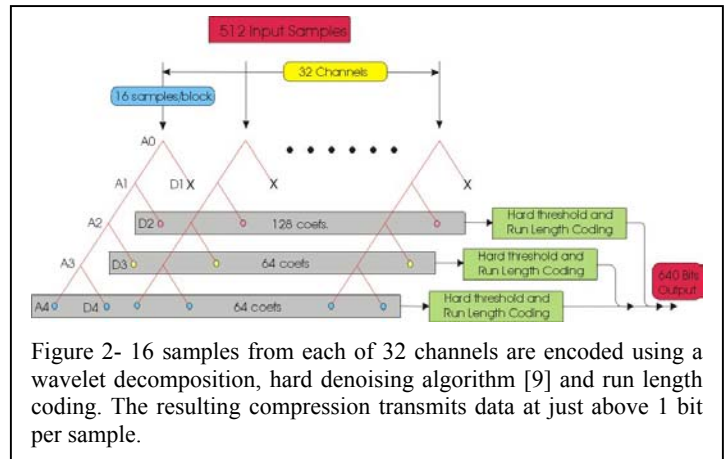
run length code efficiency. Typical input output of the coding system is shown below in figure 3. Reduction in wavelet domain vs. time domain is about 30:1 for these data. That reduction ratio holds the bit rate for 32 channels to below 1Mbit/sec.

Event time coding

The usual reference for an event time is the time since the previous event time or *run length*. The coding of run lengths depends strongly on the statistics of the intervals, which are related directly to neural properties. Some neurons discharge with nearly regular intervals that are slowly rate modulated while others are Poisson
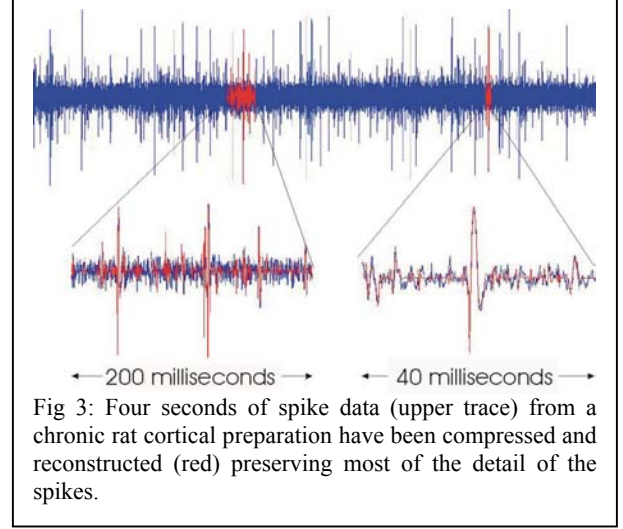

Fig 3: Four seconds of spike data (upper trace) from a chronic rat cortical preparation have been compressed and reconstructed (red) preserving most of the detail of the spikes.

distributed and can tolerate rapid rate modulation. A simple and efficient method for coding run lengths is to count the number items in the run with a finite register. When the register overflows, a zero value is assigned to that word, making it a marker and the counting starts in successive words
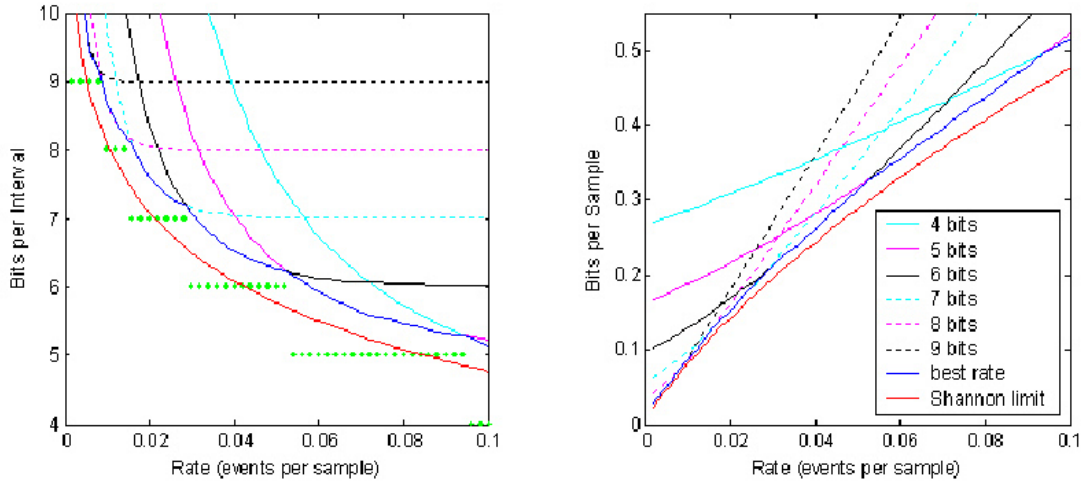

Figure 4- For exponential distributions of event intervals, Shannon theory predicts the optimum number of bits per interval is $\cong \log_2(e/\lambda)$. The run length algorithm with an optimum word length (best rate) exceeds the Shannon limit [13] by nearly a bit per interval while if the word length is badly chosen, the error grows. The data are shown for both bits/event and bits/sample. For the event rate ranges experienced for neural data, the overall bit rate is under .2 bits sample.

until the run length is reached. There is an optimum word length for the counter and that can be determined easily assuming the exponential distribution of run lengths. Figure 4 gives two views of bit rates of exponentially distributed intervals. The fundamental limit of information in bits per interval as a function of event rate is $r = -(\log_2(e^\lambda - 1) + \log_2(e^{-\lambda})/(1 - e^{-\lambda}))$ or $\cong \log_2(e/\lambda)$ [13]. It is not surprising that the optimum word length for the coder is closely related to $\log_2(average\ run\ length)$. This makes it easy to calculate the best word length and the overall bit rate needed for placing each event at its correct time location given an event rate. For rates below .002 to .02, word

lengths 7, 8, or 9 bits are very satisfactory using less than .2 bits per data sample for the intervals leaving $\geq .8$ per sample for the numerical value of the data. The algorithm complexity is very low.

V. Conclusion

A design methodology for optimizing neural information coding for transmission in neural interface *SOC* modules has been proposed. This consists the first thrust towards the development of highly sophisticated implantable/wearable neural chips interfaced with small populations of neural cells. The main challenge in the design of the neural interface chips will be to achieve the required level of processing throughput with the minimum possible power dissipation and chip area. The area of the compression chip should not exceed $0.5cm^2$, one third of the approximately $1.5cm^2$ available on the platform. The chip will be capable of simultaneously processing 32 recording channels, each at the rate of 25KHz. To that end, the chip hardware should be operating at an approximate rate of 150MHz. Using 16-bit arithmetic, the estimated power dissipation of the proposed programmable chip is expected to be in the 40mW range, yielding a power density below $100mW/cm^2$. These issues will greatly benefit from the optimization methodology proposed in this work and are currently under investigation.

VI. References

[1]   T. E. Bell and K. D. Wise, and D. J. Anderson, "A Flexible Micromachined Electrode Array for a Cochlear Prosthesis," *Sensors and Actuators, A, 66,* pp. 63-69, April 1998

[2]   Wolpaw J. *et. al.*, "Brain-Computer Interface Technology: A Review of the First International Meeting," *IEEE Trans. on Rehab. Eng.*, vol.8, no.2, pp. 164-173, June 2000

[3]   Takeuchi, S. and Shimoyama, I. "An RF-telemetry System with Shape Memory Alloy Microelectrodes for Neural Recording of Freely Moving Insects," Proc. of the 1st Annual International IEEE-EMBS Special Topic Conference on Microtechnologies in Medicine & Biology, pp. 491-496, Oct. 12-14, Lyon, France, 2000

[4]   C. Kim and K. D. Wise, "A 64-Site Multiplexed Low-Profile Neural Probe with On-Chip CMOS Circuitry," *Digest 1994 IEEE Symposium on VLSI Circuits*, Honolulu, pp 23-27. June 1994

[5]   Anderson, D.J., Najafi, K., Tanghe, S. J., Evans, D.A., Levy, K.L., Hetke, J.F., Xue, K.L., Zappia, J.J. and Wise,K.D., "Batch-fabricated thin film electrodes for stimulation of the central auditory system," *IEEE Trans. on BME*, vol.36, pp. 693-704, 1989

[6]   S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 2nd edition, pp. 413: 1999.

[7]   Oweiss K.G., Anderson D.J., " A Unified Framework for Advancing Array Signal Processing Technology of Multichannel Microprobe Neural Recording Devices, " *Proc. of the 2nd IEEE Conf. on Microtechnology in Medicine and Biology*, pp. 245-250, May 2002

[8]   S. M. Bierer and D. J. Anderson, "Multi-channel Spike Detection and Sorting using an Array Processing Technique," Neurocomputing vol. 26-27, pp. 947-956, 1999

[9]   K. G. Oweiss, D. J. Anderson, "A New Approach to Array Denoising," *Proc. of the IEEE 34th Asilomar Conference on Signals, Systems and Computers*, pp. 1403-1407, November 2000.

[10]   K. G. Oweiss, D.J. Anderson, "A new technique for blind source separation using subband subspace analysis in correlated multichannel signal environments," *Proc. of ICASSP'2001*, pp. 2813-2816, May 2001

[11]   T.K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing,* Prentice Hall, New Jersey, 1st edition, 2000

[12]   K. G. Oweiss, D.J. Anderson, " A Multiresolution Generalized Maximum Likelihood Approach for the Detection of Unknown Transient Multichannel Signals in Colored Noise with Unknown Covariance*," Proc. of ICASSP'2002*, vol. 3, pp.2993-2996, May 2002

[13]   M. Cover and Joy A. Thomas, Cover & Thomas: *Elements of Information Theory*, John Wiley, 1991