

# A Systems Approach for Real-Time Data Compression in Advanced Brain-Machine Interfaces

K. Oweiss<sup>1</sup>, K. Thomson<sup>1</sup>, D. Anderson<sup>2</sup>

<sup>1</sup>Electrical and Computer Eng. Department, Michigan State University, East Lansing, MI, USA

<sup>2</sup>Electrical Eng. & Computer Sc. Department, University of Michigan, Ann Arbor, MI, USA

*Abstract* — Advanced array processing techniques are becoming an indispensable requirement for integrating the rapid developments in wireless high-density electronic interfaces to the Central Nervous System (CNS) with computational neuroscience. This work aims at describing a systems approach for data compression to enable real-time transmission of high volumes of neural data acquired by implantable microelectrode arrays to extra-cutaneous devices. We show that the tradeoff between transmission bit rate and processing complexity requires a smart coding mechanism to yield a fast and efficient neural interface capable of transmitting the information from the CNS in real-time without compromising issues of communication bandwidth and signal fidelity. The results presented demonstrate that on-chip coding offers tremendous savings in communication costs compared to raw data transmission for off-chip analysis. Performance illustrations and experimental neural data examples are described in details.

## I. INTRODUCTION

Largely motivated by the rapid advances in *micro* and *nano* fabrication technology, microelectrode array devices are becoming more feasible to design, fabricate and utilize for recording *ensemble* activity of small populations of neurons in the central nervous system [1]. Better understanding of the computational principles underlying the dynamic interactions between large neuronal aggregates involved in processing and storing information will consequently be more feasible [2]. Nevertheless, the success of this technology is persistently mitigated by the limitation of the associated communication and signal processing technologies [3]. Currently, there is a broad gap between the volume of the data acquired by the microelectrode array and the capacity of communication between the sensor array and the outside world. As an example, a 96-electrode array being sampled at 25 kHz per channel yields an aggregate rate of 2.4 Msamples/sec. At 12 bits/word, the bit rate is nearly 29 Mbit/sec, which is well beyond the reach of present day biotelemetry developments [1, 4]. On the other hand, telemetry channel capacity of 1-2 Mbit/second from an implant to an extracutaneous device is feasible with the current technology [3], which places the required data rate at one bit per sample [5]. While this may seem to be a radical compression ratio, one may consider that current *Hi-Fi* music compression techniques can achieve less than 3 bits per sample using modern transform coding methods [6]. Characteristics of the neural signal that make it vulnerable to compression are that neural action potentials –

or *spikes*– have a sparse base of temporal support leaving many samples devoted only to noise and signal strengths above the noise baseline depending on the underlying recording conditions.

Basic information in multi-source neural data consists of detected spike event timing and spike waveform classification result. For implantable neural recording systems, the main objective is to establish a *real-time* data stream from the device while preserving the ability to accomplish the detection and classification tasks with highest fidelity. To transmit this information within the available channel capacity, the coder may require large amounts of computation and consequently time delays. The following sections will address two functions necessary to deliver both the timing and waveform information within the one bit per sample objective.

## II. NEURAL SIGNAL CONDITIONING

Due to the nature of the extracellular surroundings, large correlation is observed among noise processes across adjacent electrode channels if the array is closely spaced. Some conditioning needs to be performed to strip this redundancy before coding the signal can take place. Two steps need to be performed to minimize this redundancy: First, the data is transformed using the discrete wavelet transform (DWT) [7]. This helps compact the signal energy in very few large coefficients and spreading out the noise power in many small coefficients. Second, the data is spatially *whitened* using a spatial filter derived from a singular value decomposition (SVD) of the spatial covariance matrix. This filter aggregates the energy spread across many physical channels to a few principal channels that can be individually filtered and coded in subsequent processing stages [8].

### A. Mathematical Model

For  $N$  time samples of an array of  $M$  channels, the spatial filter computation is computationally intense if  $M$  is significantly large. If the observations  $\mathbf{Y} \in \mathfrak{R}^{M \times N}$  represent a mixture of signal sources  $\mathbf{X} \in \mathfrak{R}^{M \times N}$  and a zero mean additive Gaussian noise component  $\mathbf{Z} \in \mathfrak{R}^{M \times N}$  with a nontrivial spatial covariance matrix  $\mathbf{R}_Z \in \mathfrak{R}^{M \times M}$ , then the array model can be expressed as

$$\mathbf{Y} = \mathbf{X} + \mathbf{Z} \quad (1)$$

The Maximum Likelihood (ML) estimate of the spatial covariance of  $\mathbf{Y}$  can be expressed as

$$\mathbf{R}_Y = E[\mathbf{Y}\mathbf{Y}^T] \cong \frac{1}{N-1} \sum_{n=1}^N \mathbf{y}[n]\mathbf{y}^T[n] \quad (2)$$

which can be spectrally factored to yield

$$\mathbf{R}_Y = \mathbf{U}_Y \mathbf{D}_Y \mathbf{U}_Y^T \quad (3)$$

where the orthonormal  $M \times M$  matrix  $\mathbf{U}_Y$  is the required spatial filter that comprises the eigenvectors spanning the column space of  $\mathbf{R}_Y$ , and the  $M \times M$  diagonal matrix  $\mathbf{D}_Y$  contains the rank ordered eigenvalues of  $\mathbf{R}_Y$ ,  $\delta_1 > \delta_2 > \dots > \delta_M$ . The *whitened* observation matrix can thus be expressed as

$$\tilde{\mathbf{Y}} = \mathbf{U}_Y^T \mathbf{Y} \quad (4)$$

and consists the matrix of principal channels that compacts the neural signal energy into fewer than  $M$  channels (The exact number is determined from the diagonal matrix of eigenvalues,  $\mathbf{D}_Y$ ). These are subsequently subband filtered using the DWT operator  $W_j$  in the  $j^{\text{th}}$  subband as

$$\tilde{\mathbf{Y}}_j = W_j \{\tilde{\mathbf{Y}}\} \quad j=0,1,\dots,J \quad (5)$$

where  $J$  denotes the total number of subbands in the DWT decomposition [7]. Equation (5) 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. Denoting the thresholding matrix operator by  $H$ , the thresholded wavelet expansion can be expressed as

$$\bar{\mathbf{Y}}_j = H\{\tilde{\mathbf{Y}}_j\} \quad j=0,1,\dots,J \quad (6)$$

Equation (6) describes the prewhitened, thresholded wavelet coefficients that need to be coded for transcutaneous transmission from the neural interface as illustrated in Fig. 1.

### B. Bit Rate reduction

Aside from direct transmission of all the data points, there are three methods commonly used in analysis systems that could and in some cases have been used to reduce the bit rate out of a recording implant. These are: 1) transmission of event timing with side information on which class of event was detected; 2) transmit minimal descriptive coefficients from a transform such as principle components or; 3) transmit clips of waveforms surrounding a threshold point such as a spike leading edge. Each system of coding has its advantages and disadvantage. Table I summarizes the virtues of each of these methods. The tradeoff between rate and complexity for these coding methods is clearly seen from the table. As the amount of arithmetic grows, the data rate that must be transmitted diminishes.

TABLE I  
DIFFERENT METHODS FOR CODING MULTICHANNEL NEURAL DATA FROM  
IMPLANTABLE MICROELECTRODE ARRAYS

Method	Advantages	Disadvantages	No. Bits	No. Mult.
<b>Event timing with side information on classification</b>	Very low transmission rates	Need for sophisticated detection and classification algorithms	48	120k
<b>Transform code times w/ coefficients</b>	Low transmission rate w/o need for classification	Detection required and Computation of transform maybe costly but with simple logical flow	512	8k
<b>Event timing w/ full waveform data on each event</b>	Waveforms must be detected on implant but can be classified using off implant processor	Very large bit rate for waveform transmission	1464	None

The intermediate case from Table I is the transform coder with high compaction. The principle is that a data stream is converted to another format which may have as many values but in a form that represents the data with high energy coefficients extending above the noise level. The DWT has two large advantages: It is computationally very efficient and it tends to code neural waveforms very efficiently. The result is the same number of samples as the original signal arranged in bands with only very significant ones above the noise.

In the example of Fig. 1, 1000 samples in the original data are represented by 32 samples or 512 bits for the transformed data. If the three significant waveforms in the 1000 samples were transmitted, the total number of samples would be about 180 (or 1440 bits). Adding the bits needed for interval

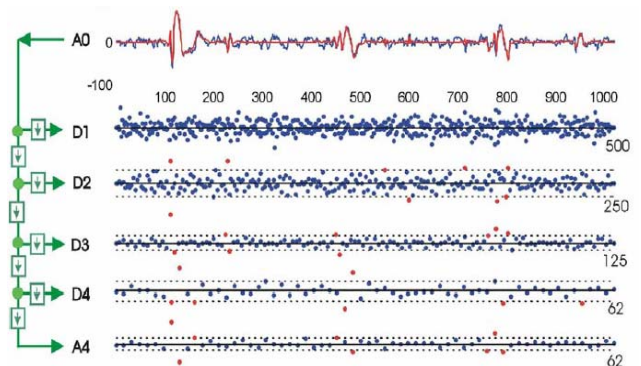


Fig. 1: The lower five traces of different sample rates represent the original blue signal labeled A0 in the top trace. The total number of samples in the five bands is the same as the number in the top trace. The largest transform samples (in red) can be used to approximate the original signal shown in red in the top trace.

transmission of 32x8 for the transform coder and 3x8 for the waveform clips, the total for the transform coder would be 512 and the total for the waveform clips is 1464 bits.

### C. Transform Computation

The transform coding yields high compaction of the signal energy and is simple to implement in hardware using the *lifting* scheme [9], which permits significant reduction in the internal memory requirements. The lifting method, is based on factorizing the wavelet filters  $h$  and  $g$  [7] into  $N$  lifting steps to obtain filters  $S_n$  and  $T_n$ . In practice, this is implemented by splitting the data into *even* and *odd* samples and applying the  $S_n$  and  $T_n$  filters consecutively, yielding an *in-place* computation. The last step is a multiplication by a scaling factor  $K$ . The outcome at an arbitrary decomposition level  $j$  is obtained as the approximation  $a_j$  and detail  $d_j$ . The former is fed back for the next level of DWT decomposition, whereas the latter is stored. When using integer wavelet transform, the last scaling steps can be eliminated.

### D. Transform Coding

For a fully integrated and functional system, the transformation and coding have to be implemented under the constraint of real-time operation. Therefore, it is necessary to interleave the computation of all levels of the DWT decomposition as described in [10]. The coefficients will have to be coded using a scheme suitable for streaming data. The traditional way of transform coding is to code the most significant magnitudes and the significance map which contains the location of these magnitudes [6]. However, for streaming applications, this methodology is inapplicable because the entire coefficient data is not available to construct the significance map. Run length encoding (RLE) is an alternative scheme for streaming data applications for which the significance map is not available. It encodes the redundancy of the zeros in a given sparse sequence. An example is illustrated in Table II.

TABLE II  
RUN-LENGTH ENCODED SEQUENCES AND THE RESULTING COMPRESSION RATIO

Sequence	RLE Result	Ratio
{5,4,0,0,3,2,0,0,1,4,2}	{5,4,0,3,3,2,0,2,1,4}	10/12
{0,0,0,0,0,5,0,0,0,0}	{0,6,5,0,5}	5/12
{5,0,5,0,5,0,5}	{5,0,1,5,0,1,5,0,1,5}	10/7

The interleaved computation of the DWT requires some sacrifice in the maximum achievable compression. In single channel RLE (*horizontal* encoding), small data segments may be stored and encoded to achieve *close to* real-time operation. An alternative is to *vertically* encode multiple channels, as they are streamlined out of the DWT real-time computation stage. In the limit, the compression gain would be maximized if the channels are pre-whitened and the number of channels is large.

## III. METHODS AND PERFORMANCE

The data used in this work was acquired with a 16-electrode 2-dimensional array that is arranged in subarrays of 4

electrodes (4 tetrodes). Accordingly, correlation is maximum among subarrays of 4 adjacent channels. We compared the compression gain of RLE of a single channel to the RLE of multiple channels to assess the performance when vertically encoding the transform coefficients. We'll denote by  $L$  the number of channels used for the vertical RLE scheme. A 4-bit value was appended to the stream, as a zero count. If the number of consecutive zeros reaches the limit, the counter resets and the number 16 is appended. If an additional consecutive zero exists in the signal, the process repeats independent of the first 16 zeros. This scheme was used for both the multichannel and single channel encoding. For a total of 700 frames, each is 128 samples long. The results were quantified against an uncompressed string of data. With five levels of decomposition performed, 124 data segments resulted. With 10-bit precision, 5120 bits were to be compressed. The compression ratio  $C$  is defined as

$$C = \frac{L \times 1280}{B} \quad (7)$$

where  $B$  is the number of bits required after RLE,  $L$  is the number of channels being processed simultaneously (also from which the spatial filter was computed). Larger compression ratio implies smaller bandwidth required for transmission.

TABLE III  
COMPRESSION RATIO OF SINGLE AND 4-CHANNEL DATA SETS.

Set	Mean		Median		St. Dev.	
	Single	Multi	Single	Multi	Single	Multi
1-4	5.19	6.57	5.14	6.44	0.58	1.11
5-8	4.75	5.84	4.76	5.81	0.77	1.39
9-12	5.51	7.32	5.38	6.94	0.78	1.78
13-16	4.12	4.74	4.01	4.47	0.81	1.40

From an experimental viewpoint, electrode channels belonging to tetrode 4 (set 13-16) had the most spatially correlated neural activity as evident from the correlation structure in Fig.2. Therefore, pre-whitening this set amounts to the closest performance to the single channel RLE because most of the energy is segregated in one principal channel to be transmitted, the rest does not survive the denoising threshold. When the data is least correlated (tetrode 3, set 9-12), the performance of the multi-channel RLE coder improves significantly compared to the single channel case. This is illustrated clearly in the bottom panel of Fig.2 for the corresponding shift in the mean of the distribution to the right implying more reduction in bandwidth.

In Fig. 3, we increased  $L$  to 8 channels. It is clear that the lack of correlation among tetrode 3 (set 9-12) was compensated for by the higher correlation among tetrode 4 (set 13-16), so that the net gain in performance was negligible compared to channels 1-8 (less correlated) as evident from table IV. In the bottom panel of Fig. 3, the correlation structure among all 16 channels is plotted as well as the performance. The net gain for  $L=16$  of the multichannel RLE

compared to single channel is also negligible, but the advantage is that streaming is achievable in this case.

#### IV. CONCLUSION

We presented a systems approach for reducing the data throughput in high-density implantable neuroprosthetic devices for Brain-Machine Interface applications. We demonstrated that transform domain processing is much more attractive for a number of reasons: First, energy compaction is largely achievable permitting significant noise suppression; Second, ease of computation using an interleaved structure for streaming data in real-time; Third, a substantial reduction in bandwidth compared to single channel for least correlated neural data and/or high-density electrode arrays. The spatial filtering mechanism has shown significant benefits. If detection is feasible *on-chip* then more compression can be achieved (1:32) by transmitting the significance map of the transform. If not, then vertical RLE is an acceptable alternative to keep real-time streaming operation possible. Thus, high signal fidelity and maximal neural yield require optimal utilization of the limited resources available *on-chip*.

#### REFERENCES

- [1] K. Wise *et al.*, "Wireless Implantable Microsystems: High-Density Electronic Interfaces to the Nervous System," *Proc. of the IEEE*, Vol.: 92-1, pp: 76-97, Jan. 2004
- [2] M. Nicolelis, "Actions from Thoughts," *Nature*, vol. 409, pp.403-407, January 2001
- [3] C. Bossetti, J. Carmena, M. Nicolelis, and P. Wolf, "Transmission Latencies in a Telemetry-Linked Brain Machine-Interface," *IEEE Trans. BME*, vol.51, No.6: 919-924, June 2004
- [4] S. Takeuchi, I. Shimoyama, "An RF-telemetry System with Shape Memory Alloy Microelectrodes for Neural Recording of Freely Moving Insects," *IEEE Conf. on Microtech. in Med. & Bio.*, pp. 491-496, October 2000
- [5] K. Oweiss, D. Anderson, M. Papaefthymiou, "Optimizing Signal Coding in Neural Interface System-on-a-Chip Modules," *IEEE Conf. on Eng. in Med. & Bio.*, pp. 2016-2019, September 2003
- [6] V. Goyal, "Theoretical Foundations of Transform Coding," *IEEE Signal Processing Mag.*, pp. 9-21, September 2001.
- [7] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 2<sup>nd</sup> edition, pp. 413: 1999.
- [8] K. Oweiss, D. Anderson, "A new technique for blind source separation using subband subspace analysis in correlated multichannel signal environments," *IEEE-ICASSP*, pp. 2813-2816, May 2001.
- [9] Y. Suhail, K.G. Oweiss, "A Reduced Complexity Integer Lifting Wavelet Based Module for Real-Time Processing in Implantable Neural Interface Devices," *IEEE Int. Conf. on Eng. in Med. and Bio.*, pp.4552-4555, September 2004
- [10] K. Thomson, Y. Suhail, and K. Oweiss "A Scalable Architecture for Streaming Neural Information from Implantable Multichannel Neuroprosthetic Devices," *IEEE Int. Conf. On Circuits & Systems*, May 2005, to appear
- [11] K. Oweiss *et al.* "A Scalable Wavelet Transform VLSI Architecture for Real-Time Neural Signal Processing in Implantable Multichannel Neuroprosthetic Devices," *IEEE Trans. On Circuits & Systems*, in review

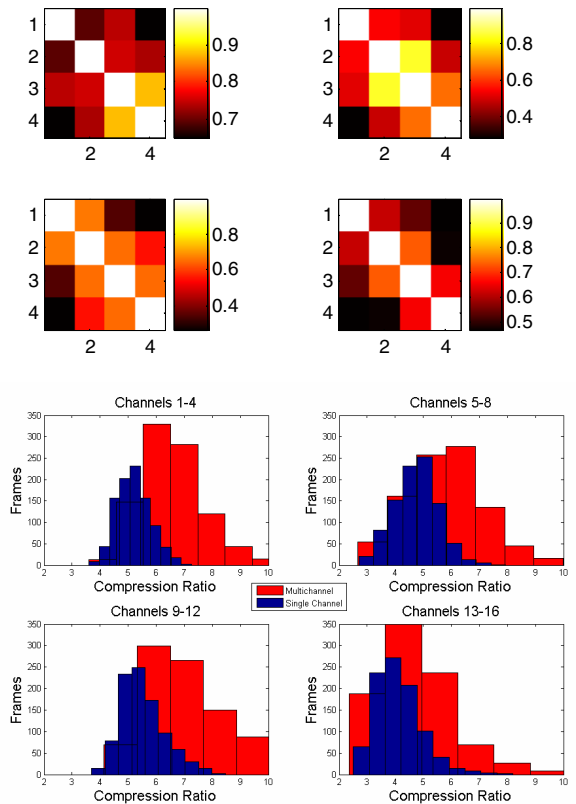


Fig. 2. Compression ratio histograms for sets of 4 channels.

TABLE IV  
COMPRESSION RATIOS OF SINGLE AND MULTICHANNEL RLE FOR L=8 AND 16 CHANNEL DATA

Set	Mean		Median		St. Dev.	
	Single	Multi	Single	Multi	Single	Multi
1-8	4.65	5.70	4.69	5.68	0.62	1.07
9-16	4.27	5.03	4.21	4.86	0.65	1.14
1-16	4.23	4.86	4.20	4.78	0.53	0.87

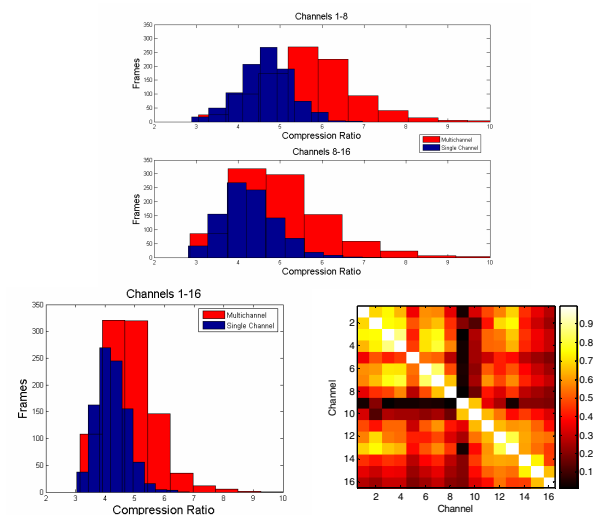


Fig. 3. Compression Ratio for RLE length of 8, and 16 channels and correlation structure for all 16 channels