Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Hardware Considerations of a Spatial Filter for Decorrelating High-density Multielectrode Neural Recordings

Kyle E. Thomson, Karim G. Oweiss*, *Member, IEEE*

Abstract— **High density implantable microelectrode arrays record large amounts of highly correlated data, which causes large strains on limited bandwidth telemetry systems. Previous work has shown that the use of a spatial filter can significantly reduce the number of channels that must be transmitted to adequately represent the data. However, the limitations on power and size for an implantable neuroprosthetic device impose significant limitations on the computational complexity of the spatial filter. We assess the performance of the floating point operations of spatial filtering and show that it can be approximated to integers with negligible losses to signal fidelity, thus reducing the computational complexity.**

I. Introduction

High density microelectrode arrays (MEAs) are becoming increasingly popular for recording large ensembles of neurons in the brain. The current generation of neuroprosthetic devices that utilize MEAs is capable of simultaneously recording and decoding neural information from hundreds of neurons. Paving the way, however, for fully implantable high density MEAs requires more sophisticated signal processing to be implemented at the front-end to enable only the useful information to be transmitted. The current generation of these devices is only capable of transmitting a handful of channels in real-time due to telemetry bandwidth restrictions. Increasing the number of channels for a fixed bandwidth amounts to an increased in latency and compromises the real time performance.

A viable solution to reduce bandwidth requirements is to detect neuronal spikes prior to telemetry. Spike detection contains a large portion of the neural information sought, but may not yield an ultimate solution to the latency problem when the number of channels is considerably large, save that spike detection is a complex task when neural nonstationarity is encountered.

Previous work [1] has shown that the data recorded has considerable temporal and spatial redundancy. This correlation can be exploited to reduce the required bandwidth. By implementing a spatial whitening filter, the number of channels that need to be transmitted can be considerably reduced. An optimal solution for whitening the recorded data requires floating point operations.

Manuscript received April 24, 2006. This work was supported by NIH grant number NS047516-01A2. Kyle E. Thomson and Karim G. Oweiss are with the Electrical and Computer Engineering Department at Michigan State University. (koweiss@msu.edu)

While this operation can be managed with state-of-the-art DSP chips, a custom designed integrated circuit that conforms to the size and power constraint of an implantable device is better suited. To enable implementing the whitening process in ASIC, integer fixed point operations is highly desired because of the amount of memory, multiplications and additions are considerably reduced.

This paper describes the whitening process of recorded neural data and the various types of approximations introduced to enable real-time hardware implementation. We further assess the impact of these approximations on signal integrity.

II. Theory

A. Array Model

The observation model for the multielectrode data assumes that P signal sources impinge on an array of M electrodes within the recording interval $T = [t_1,...,t_N]$. The observation matrix is expressed as

$$\boldsymbol{Y} = \boldsymbol{A}\,\boldsymbol{S} + \boldsymbol{Z} \tag{1}$$

where $\boldsymbol{A} \in \Re^{M\times P}$ denotes the *mixing matrix* that expresses the array response to P neural sources, $\boldsymbol{S} \in \Re^{P\times N}$ denotes the signal matrix, and $\boldsymbol{Z} \in \Re^{M\times N}$ denotes a zero-mean additive noise component with an *arbitrary* spatial and temporal correlation [2].

B. Whitening Process

The redundancy caused due to observing the signals with multiple closely spaced electrodes can be reduced using advanced signal processing techniques. A spatial filter decorrelates the data by aggregating the signal energy spread across many physical channels to a few principal channels that can be further filtered and compressed. This process is known as whitening the data, while the inverse process will be referred to as coloring the data.

Decorrelation in space can be optimally achieved [2] by first estimating the sample spatial covariance matrix of the data as

$$\hat{\boldsymbol{R}}_{\boldsymbol{Y}} = \frac{1}{N\text{-}1}\sum_{n=1}^{N} \mathrm{Y}[n]\mathrm{Y}^{T}[n] \tag{2}$$

where $\mathrm{Y}[n] \in \Re^{M\times 1}$ is the array snapshot at time sample n, and N denotes the length of time interval from which the filter is estimated. $\mathrm{Y}[n]$ is assumed to be zero mean since

the distance from zero to the actual mean is negligible in the neural environment. This covariance can be further decomposed using singular value decomposition (SVD) to yield

$$\hat{\boldsymbol{R}}_Y = \mathrm{U}_O \mathrm{D}_O \mathrm{U}_O^T = \sum_{m=1}^{M} \lambda_m \mathrm{u}_m \mathrm{u}_m^T \quad (3)$$

where λ_m denotes the m^{th} eigenvalue corresponding to the m^{th} diagonal entry in and $\mathrm{U}_O = [\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_M] \in \Re^{M \times M}$ comprises the eigenvectors spanning the column space of $\boldsymbol{R}_Y$. The whitening process is performed by the following matrix multiplication

$$\tilde{\mathrm{Y}} = \mathrm{U}_O^T \mathrm{Y} \quad (4)$$

The matrix U_O is the optimal whitening matrix and its entries are generally represented by floating point numbers.

C. Low Rank Approximation

One way to reduce the number of transmitted channels in $\tilde{\mathrm{Y}}$ is to use low rank approximation (LRA) [2] to shrink the sum in (3) to only the principal channels comprising most of the signal energy. This is guaranteed if $P \le M$ and the P signal sources have orthogonal waveforms. However, in a typical neural recording experiment, both conditions are often violated. Spike waveforms from distinct neurons are generally not orthogonal amongst each other [3][4], and typically the number of single units recorded exceeds the number of electrode channels [5]. Nevertheless, the principal role of the spatial processing stage is not to sort out the distinct signal sources, but rather reduce the *overall* signal energy impinging on the M channel array to a much smaller number of principal channels, say $Q << M$, that can be encoded and transmitted, thus reducing the required bandwidth.

Performing a LRA [2] on (4) is equivalent to zeroing out the M-Q rightmost columns of U_O. This is denoted as $\dot{U}_O = [\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_Q, 0, \ldots, 0_M] \in \Re^{M \times M}$. If the data is whitened using $\dot{U}_O$, the mean square error between the original data and the reconstructed data is defined as

$$e_{LRA}(Q) = \left(Y - \dot{U}_O \dot{U}_O^T Y\right)^2 \quad (5)$$

This error decays to zero when Q → M.

D. Integer Approximation

The hardware constraints imposed by implantability requirements relies on the minimization of chip size and power consumption. Typically, the computations being undertaken on chip require the elements be processed and stored in quantized integer format. Since the whitening process defined in Eq. 4 is a floating point process, the coefficients of U_O must be *approximated* to integers to reduce the hardware complexity of the whitening process. This creates a hardware-friendly whitening filter, which will be referred to as U_{HF}.

Quantizing the filter amounts to expressing the real numbers in U_O by their quantized values as

$$U_{HF}(i,j) = \frac{floor\left(2^B \cdot U_O(i,j)\right)}{2^B} \quad (6)$$

where B is the number of bits for quantization. It should be noted that the smaller B is, the more size and power efficient the hardware implementation is. Thus, a trade off arises between the resulting error from selecting a suboptimal B and the system performance. The following example illustrates the process of quantizing the spatial filter. A sample U_O is approximated to U_{HF} with B=3 and B=5 bits.

$$\begin{bmatrix} .9501 & .4460 \\ .2311 & .6068 \end{bmatrix} \overset{B=3}{\cong} \begin{bmatrix} \frac{8}{8} & \frac{4}{8} \\ \frac{2}{8} & \frac{5}{8} \end{bmatrix} \overset{B=5}{\cong} \begin{bmatrix} \frac{30}{32} & \frac{14}{32} \\ \frac{7}{32} & \frac{19}{32} \end{bmatrix}$$

It is obvious from the first row of the example that the error between quantizing to 3 bits versus 5 bits can be substantial. The error from approximating the entries of U_O by the hardware-friendly U_{HF} is defined as

$$e_{IA}(B) = \left(Y - U_{HF} U_{HF}^T Y\right)^2 \quad (7)$$

This error occurs due to the loss of orthonormality of the columns of U_O when representing U_O as integers.

To simplify the hardware to perform the spatial filtering, the denominator of U_{HF} can be moved outside of the matrix, leaving an integer multiplication. To perform the division, a simple hardware truncation is performed by dropping the least significant bits, and is equivalent to eq. 9.

$$\tilde{Y}' = floor\left(\frac{U_{HF}^T Y}{2^B}\right) \quad (8)$$

The error due to the loss of the decimal places will be defined as

$$e_{TR}(B) = \left(\tilde{Y}' - \tilde{Y}\right)^2 \quad (9)$$

The ideal signal processing situation when performing LRA is to perform the entire whitening process, and then determine the principal channels Q via thresholding. This approach is hardware intensive, as each sampling interval requires M^2 multiplications. Since it is expected that $Q << M$, if Q is determined in advance, only $M \times Q$ multiplications need to be performed. Reducing Q also reduces the amount of bandwidth required to transmit the neural data. Additionally, only $M \times Q$ coefficients of the spatial filter need to be stored in memory, further reducing

the amount of hardware required for spatial filtering. Thus, setting Q to a predetermined value for performing LRA in hardware is optimal. This will be referred to as fixed low rank approximation (FLRA). However, as Q is lowered, e_{LRA} increases. Additionally, if $\dot{U}_O$ is replaced with $\dot{U}_{HF}$, a low rank approximation of U_{HF}, then e_{LRA} increases as a function of e_{IA}. The error due to IA and LRA is defined as

$$e_{IA+LRA}(Q,B) = \left(Y - \dot{U}_{HF}\dot{U}_{HF}^{T}Y\right)^2 \qquad (10)$$

This error decays to e_{IA} when Q → M.

The total error is equal to the superposition of all of the individual errors incurred in the implementation of the spatial filter, i.e.

$$e_{IA+LRA+TR}(Q,B) = \left(Y - floor\left(\dot{U}_{HF}\, floor\left(\frac{\dot{U}_{HF}^{T}Y}{2^B}\right)\Big/ 2^b\right)\right)^2 \qquad (11)$$

This error is indicative of the hardware friendly spatial filter performance. In actual hardware implementation, the three types of errors are interdependent.

III. Results

The data generated for fig. 1 and fig. 2 is from a simulation of *P* random neural signals across *M* channels. Four spike templates were used to generate the neural spike trains. A different spatial covariance was generated for each simulation with parameters uniformly distributed random values between zero and one. All recordings were simulated in MATLAB. Each figure represents the mean of 100 trials.

Fig. 1 compares four of the errors listed in this paper as a function of the number of bits *B* used for quantizing U_{HF}. Note that e_{LRA} is not a function of *B,* and is included to show that the hardware performance approaches that of the floating point optimal whitening filter. This occurs *only* when the transmitted signal is colored by $\dot{U}_O{}^T$. Additionally, for lower bit values, the effect of truncation, illustrated by e_{TR}, is larger then when it is combined as with low rank approximation, as $e_{IA+LRA+TR}$. This interesting fact arises because the truncation of the whitening filter has a larger effect on smaller coefficients, and thus LRA minimizes the effect of e_{TR} for smaller levels of quantization.

Fig. 2 compares the use of LRA with $\dot{U}_O$ to $\dot{U}_{HF}$ quantized to 3, 4 and 5 bits. We notice that for the case B=3, the error decreases for Q/M < 0.3 for which the LRA error component dominates. For Q/M > 0.3, the error increases because the number of multiplications is increased and therefore is more influenced by the IA and TR errors. Increasing the quantization resolution to 4 bits helps reduce the IA and TR errors significantly up to Q/M < 0.7.

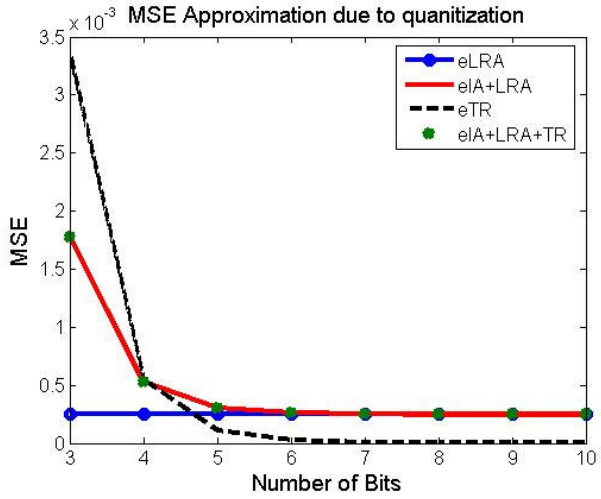

Fig. 1 A comparison between $e_{LRA}, e_{IA+LRA}, e_{TR}$ and $e_{IA+LRA+TR}$ as a function of *B*. *M=8, P=8, Q= 3.*

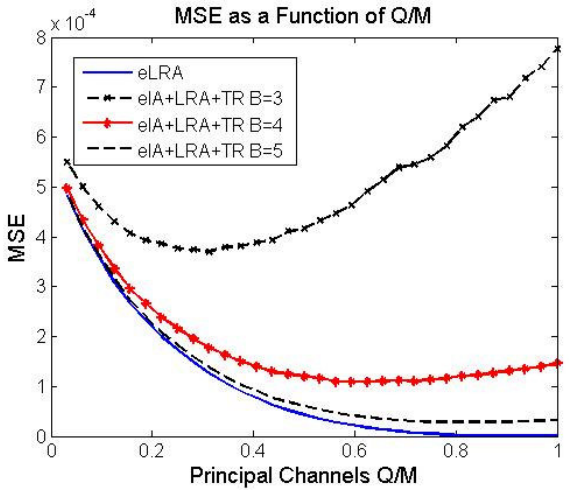

Fig. 2. A comparison of e_{LRA} and $e_{IA+LRA+TR}$ as a function of *Q/M* for B=3,4 and 5. *M=32, P=32.*

Lastly, a resolution of 5 bits stabilizes the error surface significantly and approaches the optimal floating point performance, which translates into substantial savings in computations and memory requirements.

Fig. 3 shows a comparison between the optimal spatial filter U_O and the quantized spatial filter U_{HF} with *B=3.* Fig. 3a is 100 ms of experimental data recorded from the Dorsal Cochlear Nucleus of an adult guinea pig across an 8-channel electrode array. The activity of seven distinct neurons can be discerned from the recorded data. Careful examination of the neural traces reveals that event '**A**' can be detected on all 8 channels, with various SNR, while event '**B**' can be reliably detected on the bottom 4 channels. The optimal spatial filtering case was obtained from the full 32 channels data covariance. In Fig. 3b, the energy of event '**A**' is completely contained in the first principal channel, while the energy of event '**B**' extends to the first three principal channels. Despite this mismatch, it is clear that the spatial whitening process has a large effect on reducing the total number of channels that need to be transmitted. For the

quantized filter case, illustrated by Fig. 3c, we can see that a large amount of energy is condensed into the first principal channel for event '**A**', however, energy also exists in all channels for this event. How much of this energy is actually needed for adequate signal representation depends on the acceptable loss to signal fidelity. Additionally, some of this energy may cause more error, as illustrated by Fig. 2.

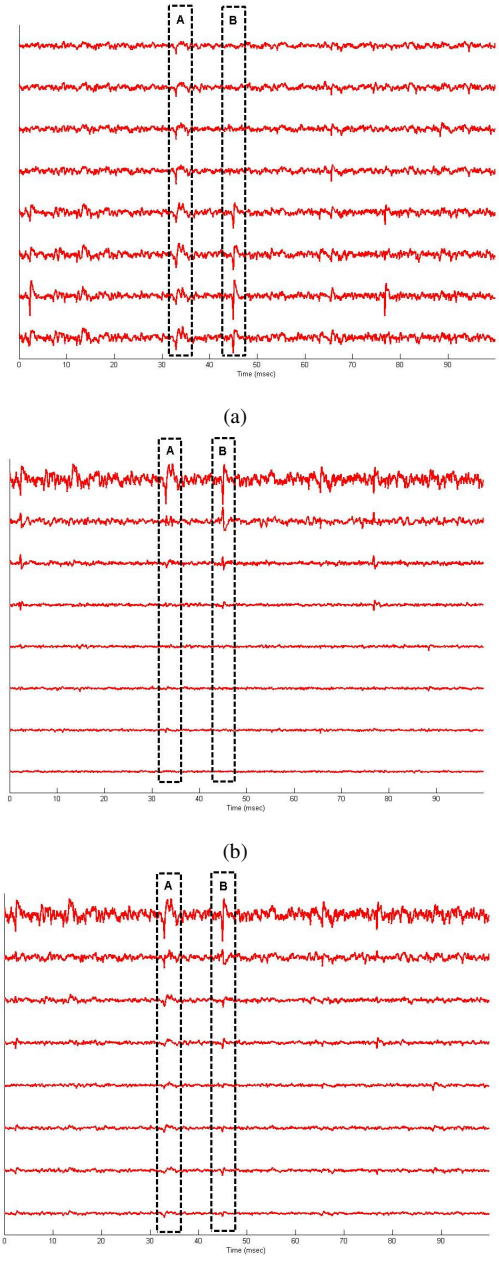

Fig. 3. (a) 100 ms of experimental data across an 8-channel electrode array before, (b) after optimal filtering with U_O, and (c) quantized filtering with U_{HF}, with *B=3*. The data in the top panel indicates that the number of channels from which data is to be transmitted has been effectively reduced from 8 to 3 for optimal filtering. Two events 'A' and 'B' are boxed to illustrate the additional energy that is spread across all channels due to integer approximation of U_O.

IV. Conclusion

Practical considerations in the implementation of a spatial filter for decorrelating highly redundant neural recordings have been discussed. It was shown that three types of errors are encountered that render the filter to be suboptimal. The effects of quantization and Fixed Low Rank Approximation on spatial filtering have been investigated. In the case where the electrode array is closely spaced such that large correlation is observed, quantization of the filter coefficients to 3 bits is acceptable for channel compression ratios of up to 30%. When the array is largely spaced such that no significant correlation is observed, quantization to 5 bits seems a reasonable choice compared to the optimal floating point representation. The proposed savings in integer approximation and FLRA were motivated by the need to efficiently process highly correlated neural recordings from high-density implantable neuroprosthetic devices. The anticipated reduction in chip area and power consumption without reducing the performance to a noticeable degree strongly suggests that the proposed savings are well suited for implementing in a spatial filter for next generation neuroprosthetic devices. Further research must be conducted to find the optimal system design and hardware layout for performing the spatial filter *in vivo.*

References

[1] Karim G. Oweiss, "A Systems Approach for Data Compression and Latemcy Reduction in Cortically Controlled Brain Machine Interfaces." *IEEE Transactions on Biomedical Engineering*, vol. 53, no.7, pp. 1364 – 1377 July 2006

[2] H. Van Trees, *Optimum Array Processing*, New York: John Wiley & Sons, 1st ed., 2002

[3] K. G. Oweiss, D. Anderson, "Spike Sorting: A novel shift and amplitude invariant technique," *J. Neurocomputing*, vol. 44-46, pp. 1133-1139, July 2002

[4] K.G. Oweiss, "Integrating temporal, spectral and spatial information for resolving muliunit extracellular neural signals," *IEEE Trans. on BME*, submitted.

[5] G. Buzsaki, "Large scale recording of neuronal ensembles," *Nat. Neurosci.*, (7):5, pp.446-451, May 2004