Energy-Efficient Two-Stage Compressed Sensing Method for Implantable Neural Recordings

Yuanming Suo, Jie Zhang, Ralph Etienne-Cummings, Trac D. Tran
Electrical and Computer Engineering
Johns Hopkins University
Email: {ysuo1, jzhang41, retienne, trac @jhu.edu}

Sang Chin
Applied Physics Laboratory
Johns Hopkins University
Email: schin11@jhu.edu

***Abstract*— For *in-vivo* neuroscience experiments, implantable neural recording devices have been widely used to capture neural activity. With high acquisition rate, these devices require efficient on-chip compression methods to reduce power consumption for the subsequent wireless transmission. Recently, Compressed Sensing (CS) approaches have shown great potentials, but there exists the tradeoff between the complexity of the sensing circuit and its compression performance. To address this challenge, we proposed a two-stage CS method, including an on-chip sensing using random Bernoulli Matrix S and an off-chip sensing using Puffer transformation P. Our approach allows a simple circuit design and improves the reconstruction performance with the off-chip sensing. Moreover, we proposed to use measureed data as the sparsifying dictionary D. It delivers comparable reconstruction performance to the signal dependent dictionary and outperforms the standard basis. It also allows both D and P to be updated incrementally with reduced complexity. Experiments on simulation and real datasets show that the proposed approach can yield an average *SNDR* gain of more than 2 dB over other CS approaches.**

I. Introduction

Implantable neural recording devices, such as Multi-Electrode Arrays (MEA), are widely used in the field of Neuroscience to monitor the Neural Action Potentials (or spikes) of a designated brain area. For an MEA with hundreds of electrodes, the acquisition rate is on the order of megabytes per second, inducing power consumption in the mW range for traditional wireless transmission method, where the energy cost is on the order of nJ/bit [1]. To tackle the high acquisition rate issue, there are two categories of on-chip compression approaches, event based and transformation based compression. Spike detection belongs to the first category [2]. It detects the spikes by thresholding and transmits only the spikes. Its implementation requires a small layout area and low power consumption. However, the information in the segments without spikes, which might also be useful to the neuroscientists, is totally lost in this approach. For the transformation based approach, the wavelet transform is usually chosen for its high compression rate and good reconstruction quality [3]. The signal is projected into the wavelet domain and only a few significant coefficients are transmitted. However, its implementation requires dedicated DSPs for the wavelet transform and on-chip memories operating at a frequency above the signal's Nyquist rate.

The emerging field of Compressed Sensing (CS) has shown comparable compression and reconstruction performance to the transformation based approach with a simpler circuitry [1], [4]–[6]. The keys in the CS approaches are the design of sensing matrix and the sparsifying dictionary. Random Gaussian and Bernoulli matrices are theoretically proven to be good candidates for the sensing matrix [7]. Bernoulli is more favorable than Gaussian because less significant bits leads to simpler implementations [1], [5]. So there is tradeoff between the complexity of the sensing circuit and its compression rate. For the dictionary, researchers have adopted wavelets such as Daubechies-8 (db8) [4], [6]. Recently, it is shown that dictionaries learned from data perform better than wavelet basis [5]. However, dictionary learning method is computational expensive and not suitable for real-time longitudinal analysis.

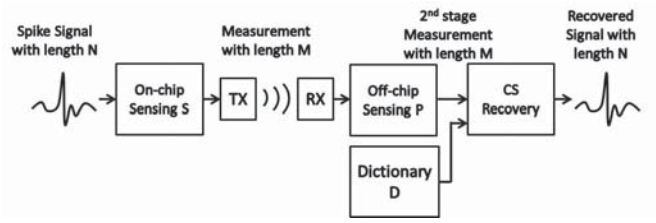


Fig. 1. The system diagram of the proposed two-stage CS approach.

To improve upon our prior work [5] and address the challenges, we developed a two-stage CS approach, as shown in Fig1. We adopted the random Bernoulli matrix $\mathbf{S}$ because of its simple implementation and add an off-chip sensing using Puffer Transformation $\mathbf{P}$. For sparsifying dictionary $\mathbf{D}$ used in CS recovery, we use measured data, instead of standard basis or trained dictionaries. Compared to other approaches, this approach can achieve three advantages simultaneously,

a) *Energy efficiency*: We can achieve a higher compression rate and better reconstruction performance than previous CS approaches (including our prior work [5]), which could significantly reduce the power consumption of transmission.

b) *Simple circuit*: The on-chip sensing matrix $\mathbf{S}$ can be realized with simple circuits while the performance can be further boosted by the off-chip sensing $\mathbf{P}$.

c) *Computational efficiency*: The matrix $\mathbf{P}$ and $\mathbf{D}$ can be updated incrementally with reduced complexity. This opens up the possibility for real-time processing using devices with limited computing power, such as smartphones.

978-1-4799-1471-5/13/$31.00 ©2013 IEEE

The rest of the paper is organized as follow. In Section II, we cover each component of our two-stage CS method, including the on-chip/off-chip sensing matricies and the sparsifying dictionary. Then we present an algorithm for incremental updates of these components. In Section III, we compare the proposed approach with the other CS based approaches and spike detection method using both simulation and real datasets. We end the paper with a conclusion in Section IV.

II. Two-Stage Compressed Sensing Method

A. Background on Compressed Sensing

Compressed sensing (CS) introduced a theoretical framework regarding the exact recovery of a K-sparse signal $\mathbf{x} \in \mathbb{R}^N$ from a measurement vector $\mathbf{y} \in \mathbb{R}^M$, where $K < M << N$. The K-sparse signal is defined as a signal that has exactly K non-zero coefficients, or can be well approximated by its largest K coefficients. Given that a sensing matrix $\mathbf{S}$ satisfying the Restricted Isometry Property (RIP) and $M \sim K\log(\frac{N}{K})$ [7], the K-sparse vector $\mathbf{x}$ can be recovered with high probability by solving the following l_1-norm minimization problem:

$$\underset{\mathbf{x}}{\operatorname{argmin}} ||\mathbf{x}||_1 \;\; s.t. \;\; ||\mathbf{y} - \mathbf{S}\mathbf{x}||_2 \leq \epsilon \tag{1}$$

where $\mathbf{S} \in \mathbb{R}^{M \times N}$ is the sensing matrix. Most often, the signal is not sparse in spatial domain but with respect to some basis $\mathbf{D}$. Then the optimization problem becomes,

$$\underset{\alpha}{\operatorname{argmin}} ||\alpha||_1 \;\; s.t. \;\; ||\mathbf{y} - \mathbf{S}\mathbf{D}\alpha||_2 \leq \epsilon \tag{2}$$

where $\mathbf{D} \in \mathbb{R}^{N \times L}$ is the dictionary that sparsify the signal and ϵ controls the approximation quality. After recovering the K-sparse signal $\alpha \in \mathbb{R}^L$, the estimated signal is

$$\mathbf{x}_{rec} = \mathbf{D}\alpha \tag{3}$$

Typically, we have an overcomplete dictionary ($N \leq L$). We will explain how to choose $\mathbf{S}$ and $\mathbf{D}$ in the following sections.

B. Design of Sensing Matrix $\mathbf{S}$

1) Prior Work: Random Bernoulli and Gaussian matrices satisfy RIP universally with small M [7]. However, given the dictionary or basis of the specific signal, an optimized sensing matrix could outperform random matrices. In [8], Sapiro et al. proposed to optimize $\mathbf{S}$ by reducing the average *mutual-coherence* and showed to achieve better performance than using random matrices. But the optimized $\mathbf{S}$ contains fraction values, which makes it hard for circuit implementations.

2) Our Approach: We propose a two-stage sensing scheme, including an on-chip sensing stage with $\mathbf{S}$ and an off-chip sensing stage with $\mathbf{P}$. The optimization problem now becomes,

$$\underset{\alpha}{\operatorname{argmin}} ||\alpha||_1 \; s.t. \; ||\mathbf{z} - \mathbf{P}\mathbf{S}\mathbf{D}\alpha||_2 \leq \epsilon \tag{4}$$

where $\mathbf{z} = \mathbf{P}\mathbf{y} \in \mathbb{R}^M$ is the measurement after second sensing stage and $\mathbf{P} \in \mathbb{R}^{M \times M}$ is the off-chip sensing matrix.

For on-chip sensing stage, we choose random Bernoulli matrix with values $\{1, -1\}$. Sensing operation using Bernoulli matrix can be implemented using M simple digital accumulators (as shown in Fig 2). The accumulators operate at the signal's Nyquist frequency [5]. Depending on the matrix entry, the accumulator either adds or subtracts the incoming signal value from the accumulated value. On the other hand, sensing matrices, both random Gaussian matrix and matrices with multiple significant bits, requires the implementation of multipliers. Thus the circuit complexity, area, and power consumption increases. From a on-chip storage point of view, only 1-bit is needed to store each entry of the Bernoulli sensing matrix, while it requires multiple bits to store one entry of a matrix with with more than one significant bits.

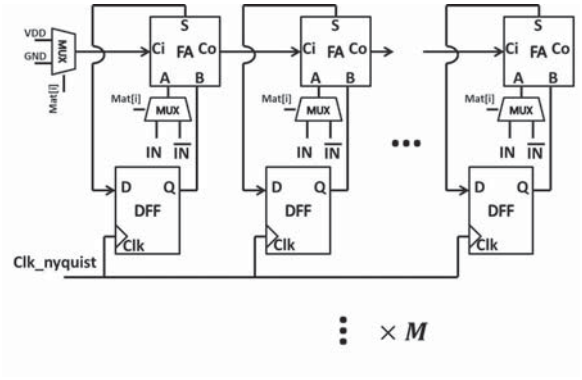


Fig. 2. The implementation of on-chip sensing matrix $\mathbf{S}$.

For the off-chip sensing stage, we adopt the concept of Puffer Transformation from the field of Statistics [9]. Given a design matrix $\mathbf{E}$, the corresponding Puffer Transformation $\mathbf{P}$ is the transformation that inflates its smallest nonsingular values. The Puffer Transformation is shown to improve the irrepresentable condition, which is related to *mutual-coherence* [9]. If we define the singular value decomposition (SVD) of $\mathbf{S}\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, the corresponding $\mathbf{P}$ will then be,

$$\mathbf{P} = \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{U}^T \tag{5}$$

For the case where the i-th singular value $\mathbf{\Sigma}_{i,i}$ is zero, we define $\mathbf{\Sigma}^{-1}_{i,i}$ to be zero as well.

C. Design of Dictionary $\mathbf{D}$

1) Prior Work: For spikes, time-frequency transformations such as wavelet transform are often chosen as $\mathbf{D}$. It has also been shown that a signal dependent $\mathbf{D}$ trained from previously acquired neural signals can yield better performance than these off-the-shelf choices [5]. An established dictionary learning method is K-SVD [10], which finds $\mathbf{D}$ by iterating between approximation of the training data and sparse decomposition of the training data with respect to the dictionary. For each iteration, one SVD is performed to update each dictionary atom, which makes it computationally expensive.

2) Our Approach: Physiological recordings and Neuroscience theories have suggested that the shape of the spikes are quite reproducible for each neuron over time. Neuroscientists use this property to identify multi-neuron activities with spike

sorting techniques. Inspired by this *self-expressiveness* property, we propose to use acquired spike data as the dictionary $\mathbf{D}$ without the need of training. We assumes each spike signal can be sparsely represented as a linear combination of other spike signals. Using data as the dictionary obviously avoid the intense computation needed by dictionary learning. The $\mathbf{D}$ can be either initial acquisition at Nyquist rate for a certain period or the recovered spikes. In the case of sparsely firing neurons (no prior knowledge in the dictionary), we can always trigger the full Nyquist acquisition depending on the feedbacks from Bayesian sparse recovery [11].

D. Incremental Update of $\mathbf{P}$ and $\mathbf{D}$

In some scenarios, the Neuroscientists need to perform longitudinal analysis of spike shapes. In other cases, the sparsely firing neurons need special attentions. Both circumstances require an update of the dictionary $\mathbf{D}$ and the corresponding off-chip sensing matrix $\mathbf{P}$. The dictionary update in our approach is just $[\mathbf{D} \quad \mathbf{D_{new}}]$, where $\mathbf{D_{new}} \in \mathbb{R}^{N \times L_{new}}$ is the new observations. Because the off-chip sensing matrix $\mathbf{P}$ is related to the SVD of $\mathbf{SD}$, we can use the incremental PCA scheme [12]. The incremental update of sensing matrix $\mathbf{P}$ is shown in Algorithm 1, where $orth(\cdot)$ performs orthogonalization via QR and $svd(\cdot)$ performs SVD. The proposed algorithm has a computation complexity of $O(NL_{new}^2)$, versus $O(N(L+L_{new})^2)$ for recomputing the SVD of the whole dictionary. Moreover, the total storage required reduces to $O(N(L+L_{new}))$, down from $O(N(L+L_{new})^2)$ with the naive approach. We can also add a forgetting factor to remove the old data from the dictionary.

Algorithm 1: Incremental Update of Sensing Matrix $\mathbf{P}$

Input: $\mathbf{S}$, $\mathbf{D_{new}}$ and $\mathbf{U}$, $\mathbf{\Sigma}$ from the SVD of $\mathbf{SD}$
Output: The updated sensing matrix $\mathbf{P_{new}}$

$\tilde{\mathbf{D}}_{\mathbf{new}} \leftarrow orth(\mathbf{SD_{new}} - \mathbf{UU}^T\mathbf{SD_{new}})$
$\mathbf{R} \leftarrow \begin{bmatrix} \mathbf{\Sigma} & \mathbf{U}^T\mathbf{SD_{new}} \\ \mathbf{0} & \tilde{\mathbf{D}}_{\mathbf{new}}(\mathbf{SD_{new}} - \mathbf{UU}^T\mathbf{SD_{new}}) \end{bmatrix}$
$\tilde{\mathbf{U}}, \tilde{\mathbf{\Sigma}} \leftarrow svd(\mathbf{R})$
$\mathbf{U_{new}} = [\mathbf{U}\ \tilde{\mathbf{D}}_{\mathbf{new}}]\tilde{\mathbf{U}}$
$\mathbf{\Sigma_{new}} = \tilde{\mathbf{\Sigma}}$
return $\mathbf{P_{new}} = \mathbf{U_{new}}\mathbf{\Sigma_{new}^{-1}}\mathbf{U_{new}^T}$

Given $\mathbf{S}$, $\mathbf{P}$ and $\mathbf{D}$, we use Bayesian Compressive Sensing [11] to solve Eq(4) and recover the signal $\mathbf{x}_{rec}$ by Eq(3).

III. Experimental Results

In this section, we first compare the recovery performance of our dictionary with the standard basis and signal dependent dictionaries. Then the proposed two-stage sensing approach is compared with other sensing matrices. The experiments are performed on the Leicester neural signal database [5], which contains 20 simulation datasets. Each dataset contains spikes from three different neurons under different noise levels. We also test the proposed approach on the publicly available dataset hc-1 [13], which is the recording from nearby neurons in the hippocampus of an anesthetized rat. The hc-1 dataset contains four different types of spikes. We take N discrete samples (128 for Leicester datasets and 40 for hc-1 dataset) around each spikes to form the signal frames and retain only the signal segments containing one spike. The experiments are performed five times on each dataset and the average performance is found in terms of Signal to Noise and Distortion Ratio (*SNDR*) with the unit in dB [1], which is defined as,

$$SNDR = 20\log\frac{||\mathbf{x}||_2}{||\mathbf{x}-\mathbf{x}_{rec}||_2} \tag{6}$$

Since the comparison result is consistent across different datasets with different M, we only report part of it.

TABLE I
Comparison of recovery performance (in *SNDR*) between different dictionary methods. M is the number of measurements and N is the signal length.

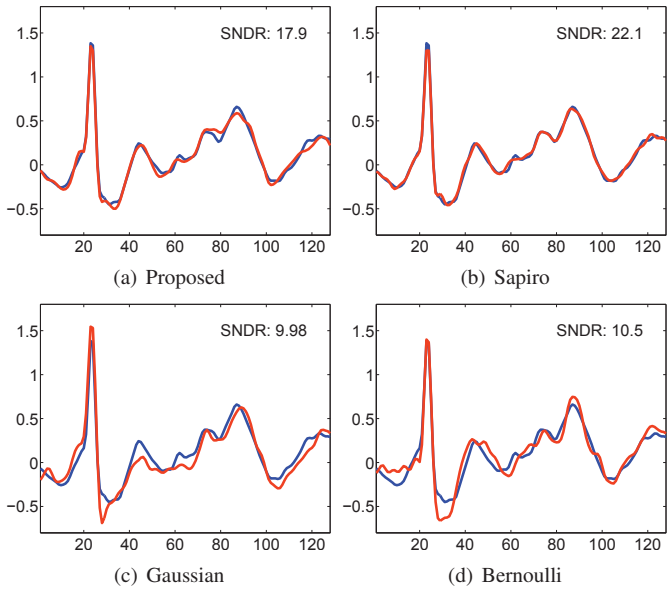
	Compression Ratio $\frac{M}{N}$ = 0.2			
	This Work	K-SVD	Wavelet	Spike Detection
Easy1-noise02	9.9	9.7	-0.1	3.4
Easy2-noise02	9.9	9.4	0.0	4.1
Difficult1-noise02	9.4	8.8	-0.7	3.4
Difficult2-noise02	10.0	9.7	-0.1	3.7
hc-1	5.5	5.8	-1.8	5.2
	Compression Ratio $\frac{M}{N}$ = 0.3			
Easy1-noise02	13.2	13.3	3.9	4.3
Easy2-noise02	13.3	13.5	4.8	4.8
Difficult1-noise02	12.6	12.7	2.5	4.2
Difficult2-noise02	12.9	13.4	3.2	4.4
hc-1	7.6	8.2	-0.5	6.5

A. The Recovery Performance of Our Dictionary

Under different compression ratios $\frac{M}{N}$ of Nyquist rate, we compare four different choices of dictionary $\mathbf{D}$ in CS framework, including the measured data dictionary, trained dictionary [5], and db-8 wavelet dictionary [1], [4]. The spike detection method [2] is also included for comparison. For spike detection method, we use threshold crossing to find the segment containing spikes with length M. In CS framework, we use same random Bernoulli matrix for $\mathbf{S}$. For methods that involve dictionary training, we randomly split the data into two halves with one part for training and one part for testing. The parameters for dictionary learning is the same as in [5]. The result is shown in Table 1. We can see that current approach using data as the dictionary works comparably to using the trained dictionary and far better than other approaches. Note that the dictionary for our approach is much simpler to update than the dictionary training approach, which makes it more suitable for real-time monitoring applications.

B. The Recovery Performance of Two-Stage Sensing Approach

Under different compression ratios $\frac{M}{N}$, we also compare four different choices of sensing matrices $\mathbf{S}$ in CS framework,



(a) Proposed (b) Sapiro (c) Gaussian (d) Bernoulli

Fig. 3. An example of the recovered signal using different sensing approaches.

including the two-stage sensing, Sapiro's approach [8], random Bernoulli matrix [5] and random Gaussian matrix. We use data directly for dictionary $\mathbf{D}$. The other experiment setup is same as previous section. The average *SNDR* is shown in Table 2. The proposed two-stage sensing approach works a little worse than the Sapiro's approach. However, it has the benefit of easy circuitry while Sapiro's approach requires a much more complicated implementation. Moreover, the proposed two-stage CS approach works consistently about more than 2dB better than using random matrices. To be more intuitive, we also show an example of the recovered signal using different sensing approaches under a compression ratio of 0.2 in Fig 3. The blue curve is the groundtruth and the red curve is the recovered signal under different approaches. Both this work and Sapiro's approach can recover the signal at coarse and detail level. But the recoveries using random matrices are bad at details, which could jeopardize the subsequent analysis, such as spike sorting.

IV. Conclusion

We presented a two-stage compressed sensing approach for implantable neuron recordings. Using data as the dictionary, our method is designed to allow simple circuit design and real-time monitoring. Experiments on simulation and real datasets have shown that the proposed approach outperform other approaches when compression rate, recovery quality and computational efficiency are all important. Nevertheless, our approach can also be applied to other biological devices.

Acknowledgment

The authors are partially supported by NSF under Grant 1057644, CCF-1117545 and Grant DMS-1222567, ARO under Grant 60219-MA, ONR under Grant N00014-12-1-0765 and Grant N00014-10-1-0223.

TABLE II

Comparison of recovery performance (in *SNDR*) between different sensing methods.

	Compression Ratio $\frac{M}{N}$ = 0.2			
	This Work	Sapiro	Gaussian	Bernoulli
Easy1-noise02	12.9	18.5	10.6	10.8
Easy2-noise02	13.9	18.3	10.9	10.2
Difficult1-noise02	13.2	17.7	9.6	9.0
Difficult2-noise02	13.8	18.7	11.0	10.1
hc-1	7.8	9.4	6.9	6.1
	Compression Ratio $\frac{M}{N}$ = 0.3			
Easy1-noise02	20.4	24.6	13.7	13.8
Easy2-noise02	20.6	24.3	14.2	13.8
Difficult1-noise02	19.6	24.0	13.0	13.5
Difficult2-noise02	20.2	24.2	12.7	12.9
hc-1	9.6	12.0	9.2	8.3

References

[1] F. Chen, A. P. Chandrakasan, and V. M. Stojanovic, "Design and analysis of a hardware-efficient compressed sensing architecture for data compression in wireless sensors," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 3, pp. 744–756, 2012.

[2] B. Gosselin, A. E. Ayoub, J.-F. Roy, M. Sawan, F. Lepore, A. Chaudhuri, and D. Guitton, "A mixed-signal multichip neural recording interface with bandwidth reduction," *Biomedical Circuits and Systems, IEEE Transactions on*, vol. 3, no. 3, pp. 129–141, 2009.

[3] K. G. Oweiss, A. Mason, Y. Suhail, A. M. Kamboh, and K. E. Thomson, "A scalable wavelet transform vlsi architecture for real-time signal processing in high-density intra-cortical implants," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 54, no. 6, pp. 1266–1278, 2007.

[4] Z. Charbiwala, V. Karkare, S. Gibson, D. Markovic, and M. B. Srivastava, "Compressive sensing of neural action potentials using a learned union of supports," in *Body Sensor Networks (BSN), 2011 International Conference on*. IEEE, 2011, pp. 53–58.

[5] J. Zhang, Y. Suo, S. Mitra, C. Peter, T. D. Tran, F. Yazicioglu, and R. Etienne-Cummings, "A signal dependent sparse representation and a recovery algorithm of neural signals," in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*. IEEE, 2013.

[6] C. Bulach, U. Bihr, and M. Ortmanns, "Evaluation study of compressed sensing for neural spike recordings," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 3507–3510.

[7] D. L. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[8] J. M. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *Image Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1395–1408, 2009.

[9] J. Jia and K. Rohe, "Preconditioning to comply with the irrepresentable condition," *arXiv preprint arXiv:1208.5584*, 2012.

[10] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[11] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *Signal Processing, IEEE Transactions on*, vol. 56, no. 6, pp. 2346–2356, 2008.

[12] A. Levey and M. Lindenbaum, "Sequential karhunen-loeve basis extraction and its application to images," *Image Processing, IEEE Transactions on*, vol. 9, no. 8, pp. 1371–1374, 2000.

[13] D. A. Henze, Z. Borhegyi, J. Csicsvari, A. Mamiya, K. D. Harris, and G. Buzsáki, "Intracellular features predicted by extracellular recordings in the hippocampus in vivo," *Journal of neurophysiology*, vol. 84, no. 1, pp. 390–400, 2000.