2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

High-fidelity compression of extracellular recordings from motor cortex

Rachel Zhang[1], Gang Pan*[1] , Yueming Wang[1,2], Zhenfang Hu[1]
[1]Department of Computer Science, Zhejiang University, Hangzhou, China
[2]Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, China

***Abstract*—In invasive brain-machine interfaces (BMI), the recorded high-quality neural signals produce a large data volume. This calls for effective compression. In this paper, we focus on extracellular recording of motor cortex. First the characteristics of the signals are studied, one of which is that peaks of DCT coefficients at high frequency may correspond to spike firing patterns. Based on these characteristics, we propose a high-fidelity compression framework for these signals. The DCT coefficients of the signal are divided into two parts according to amplitude, rather than frequency. The Low-Amplitude-Component (LAC) is encoded by a phase called Symbol Encoding, which helps to reduce overall distortion. The High-Amplitude-Component (HAC), containing major information and spikes, is encoded by another phase called Hybrid Encoding. It combines the Huffman encoding and a novel Zero-Length-Encoding. Experiments show that the algorithm achieves a compression ratio of 18% without obvious distortion. Moreover, spikes are reserved more than 92%, outperforming existing work. Our algorithm enables low-cost storage devices to store long-time neural signals.**

I. Introduction

Biological brain is one of the most complex systems ever to be studied. It has a huge sensory system, transmitting mechanism and action executor that alter the state of mind or body continuously. Recent development on neural recording techniques has made it possible to collect neuron activities from brain, leading to the development of brain-machine interfaces (BMI) system [1], [2], [3].

BMI systems can be classified to invasive and non-invasive. Non-invasive methods such as scalp electroencephalogram (EEG) is easily accessible, but with low signal precision. On the contrary, the invasive methods use surgically implanted electrodes, recording extracellular neurons' signals with finest detail [4], [5], which is referred to *"action potential"*, or *"spikes"* in individual neurons. When excited, neurons create ion currents through their membranes, causing the cell to depolarize and trigger a spike.

This paper focuses on the motor cortex neuronal signals recordings. As an important part of cerebral cortex, motor cortex is in charge of planning, controlling and executing voluntary movement of body. In the researches of motor cortex function, an extracellular recording of a channel is split into different frequency bands. Lower frequencies (cut off at 100Hz) correspond to local field potential (LFP), while medium to high frequencies correspond to spikes. The LFP mainly originates from pre-synaptic activity, composed of more sustained currents reflecting the averaged dendric activity. And the spikes mainly represent activities of excitatory neurons. Both of them are significant to signal decoding. For motor cortex, spikes usually last for less than one millisecond. Therefore high resolution device multi-electrode array (MEA) is penetrated into tissues to capture signals from hundreds of interested neurons. The mammalian neuronal signal of motor cortex is usually recorded by MEA with 128 channels at 20-30 kHz to well preserve the detail of spikes. Consequently, with 16-bit A/D resolution and a maximum sampling rate of 30kHz, data stream of such raw format in overall 128 channels is recorded at 7.68MB/s. In other words, it produces 28.8GB raw data in an hour. This not only brings significant cost for data storage, but also challenges data transmission. Therefore, compression is desirable for neuronal recordings.

Although BMI systems are well established, compression for cerebral extracellular recording is not deeply investigated. Some relevant work such as the compression on Electromyography (EMG) and Electroencephalography (EEG)[9], [12], [13] take signal characteristics into consideration for effective compression. However, the invasive extracellular recording is quite different.

Existing compression algorithms for multi-channel extracellular recording is implemented from two threads. One is to compress signal of each channel individually using intra-channel properties; the other is to decrease the redundancy among channels using inter-channel correlation. From the first perspective, Weber et al. [14] compress somatosensory cortex (S1) neuronal responses of rat by a wavelet based coder, achieving the compression ratio low to 5%. However, this compensates for the loss of 25% of the spikes, which is not desirable for future analysis. For the same recorded data of rat's S1 response, Chen et al's result achieved Signal to Noise Ratio (SNR) at about 25db with compression ratio larger than 25% [7] by adaptively qualification, in which both compression ratio and signal quality is not guaranteed perfectly. To improve their work in terms of the second point of view, Chen et al. [8] take advantage of correlation between channels, achieving 5% compression ratio with SNR at 25db by a video compression method. However, all the work above loss much detail signal, making the high quality raw

* Corresponding author (e-mail: gpan@zju.edu.cn).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

signal acquired in vain.

This paper proposes a framework that compresses extracellular recording of motor cortex with high fidelity. First, three special characteristics of these signals are investigated: 1) the power centralization on low frequency; 2) peaks of discrete cosine transform (DCT) coefficients at high frequency may correspond to spike firing patterns; 3) the Inter-channel correlation is unstable. According to 3), we encode each channel of the signal independently. According to 2), a novel Amplitude Filter is proposed to divide the DCT coefficients into two parts by amplitude rather than frequency. The value of the Low-Amplitude-Component (LAC) is encoded by a phase called Symbol Encoding to reduce overall distortion. The High-Amplitude-Component (HAC), containing major information and spikes, is encoded by another phase called Hybrid Encoding, which consists of the Huffman encoding and a novel Zero-Length-Encoding. The main features of our framework are as follows.

- A novel Amplitude Filter is designed specially to extracellular recording. It divides the DCT coefficients into two parts according to amplitude rather than frequency. This avoids the loss of spike information.
- A Symbol Encoding method is proposed to encode the values of Low-Amplitude-Component, rather than simply discarding them. It helps to reduce the overall distortion of signal.
- A Hybrid Encoding method consisting of the Huffman encoding and a novel Zero-Length-Encoding is devised to encode the High-Amplitude-Component and the frequency indices of Low-Amplitude-Component. The spike information is thus preserved with concise structure.

A number of instances have been examined on our proposed framework, achieving an average SNR at 36db and compression ratio of 18%. The fidelity of spikes is also kept higher than 92%, making reconstruction performance guaranteed.

II. Characteristics of Extracellular Recordings from Motor Cortex

To compress effectively while maintaining the quality of signal, the properties of recorded multi-channel extracellular recording are analyzed. Our dataset is described in SectionV. Three characteristics are summarized from intra-channel property to inter-channel correlation. All the characters are vital to the compression algorithm as well as experiment measuring.

1. Power centralizes on low frequency

To investigate the characteristic of the recorded temporal signal in spectral domain, discrete cosine transformation (DCT) has been adopted. As a variation of Fourier Transformation, DCT is preferable because it derives a set of real number coefficients, called DCT coefficients.

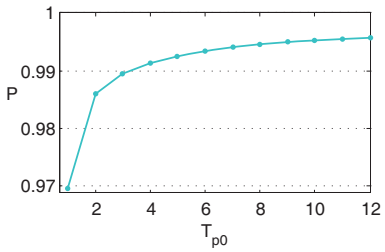

Fig. 1. Statistical Energy ratio of the first 12 dimension DCT coefficients. The horizontal axis T_{p0} denotes the number of components to be taken into consideration. The vertical axis is the energy proportion P of the first T_{p0} components.

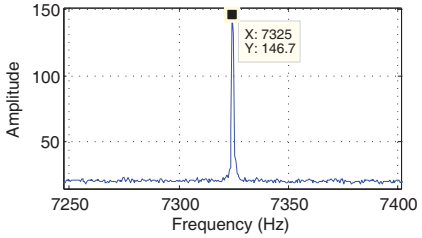

Fig. 2. DCT coefficients amplitude distribution of Electrophysiological Signal in High frequency

The transformed DCT coefficients vector of the i-th channel is denoted by $x_i = [x_i^1, x_i^2, ..., x_i^N]$ Let x_i^j be the j-th DCT component of x_i. The energy proportion of the low frequencies part is calculated on the whole dataset:

$$P = \frac{\sum_i \|x_i^1, x_i^2, ..., x_i^{T_{p0}}\|_2}{\sum_i \|x_i\|_2} \quad (1)$$

where the denominator is the total amount of energy over all channels and the numerator is the energy summation of the first T_{p0} DCT coefficients, i.e., energy of the low frequency part with the cutoff frequency at T_{p0}. The average value of P on the whole dataset with T_{p0} is shown in fig.1. It is clearly illustrated that few number of DCT components occupy the dominant energy. In other words, considerable power is centralized on low frequency domain.

2. Remarkable peaks locate at high frequency

Similar with other natural signal, the power of extracellular recording focus much on low frequency. However, such signal makes a difference at the medium and high frequency. As fig.2 shows a truncated spectrum at medium frequency. There is a peak at 7325Hz, which corresponds to a frequent neuronal firing pattern. Actually, experiment shows that some channels share the peak frequency positions while some are not. This can be comprehended from the sampling mechanism of Multielectrode array, by which the extracellular recording of a channel is composed by signal generated from 3 to 5 neurons with different spike firing patterns.

3. Unstable Inter-channel Correlation

The third character of motor cortex signal is the unstable correlation between channels. The correlation between channels is calculated for every sample. The average fluctuation extent (defined as the average divided standard deviation over time) among all samples is calculated to be 0.68. That is to say, the correlation coefficient varies seriously over time, so the correlation between channels is unstable in our obtained motor cortex neuronal responses, leading to difficulties in reducing redundancy among channels.

III. The Proposed Compression Method: An Overview

In this paper, we propose a high-fidelity compression framework for extracellular recording, taking into account the above mentioned characteristics. First of all, because the inter-channel correlation is unstable, we process each channel independently. The framework is shown in fig.3. It contains two consecutive modules: *"Preprocessing"* and *"Dual-Phase Encoding"*.[1]

For each channel, the original long signal is firstly segmented into small blocks of size S_b, then each block is processed by the following two modules:

a) ***Preprocessing****:* This module transforms signal into frequency domain by DCT. Since some peaks at high frequency may correspond to specific spike firing patterns, it is unreasonable to apply traditional low-pass filter for compression. So rather than divide the components by frequency and discard one part, we propose to divide them by amplitude and compress the two parts separately. Passed through an amplitude filter, the DCT coefficients above a threshold is stored in High-Amplitude-Component (HAC), which contains of salient LFP and action potential. The rest coefficients below the threshold are put to Low-Amplitude-Component (LAC).

b) ***Dual-Phase Compression****:* In the first phase, LAC is compressed by Symbol Encoding. It is a lossy encoding scheme, which intends to represent each coefficient by one symbol. In the second phase, DCT coefficients contained by LAC are set to zeros in HAC. This achieved vector, which contains HAC, is then quantified and compressed by a hybrid encoding method blending the Huffman Encoding with a Zero-Length-Encoding. The Huffman Encoding deals with the high amplitude entries (nonzeros after quantization), while the Zero-Length-Encoding deals with the zeros. Compression of LAC and HAC separately would effectively preserves the spectrum positions of both LAC and HAC without storing additional information. In the end, the codes of the two phases are formatted to store.

[1]The reconstruction can be obtained by directly reversing the encoding steps.

IV. Dual Phase Encoding

The Dual-Phase Encoding module is composed of two compression phases for different parts of extracellular recordings.

A. Symbol Encoding for Low-Amplitude Component

Symbol Encoding is a lossy encoding scheme that encodes the value of each coefficient of LAC separately. The sign of coefficient is preserved and its magnitude is represented by the average magnitude taken over all blocks (of the original long signal) at the corresponding frequency. Thus, beforehand, a Quantization Table (QT) is established. Each row of it corresponds to a channel, storing the average magnitudes at all frequencies. Then, finally each coefficient of LAC, denoted by l_i, is actually encoded by one-bit symbol $symbol(l_i)$: 1 if it is positive, -1 otherwise. In precise,

$$symbol(l_i) = \begin{cases} -1, & -T_{LH} < l_i \leq 0 \\ 1, & 0 < l_i < T_{LH} \end{cases} \tag{2}$$

where T_{LH} is the threshold between LAC and HAC.

It is not necessary to record the positions of LAC for decompression. As the Low amplitude components are extracted for *Symbol Encoding*, their values are assigned to zeros in the DCT coefficients vector. Later processing would guarantee nonzero of other compressed components. As a consequence, LAC can be recovered by selecting out the zero entries.

B. Quantization for High-Amplitude Component

Since HAC includes LFP and salient spikes, this vital part is designed for better preservation. This component is first quantized to a small range for better compression, then a hybrid encoding method is taken for further compression. The first step referred to as quantization is presented in this section.

Let $QT \in R^{N_c \times S_b}$ denotes the quantization table, where N_c is the number of channels and S_b is the size of block. The c^{th} entry of quantized signal $\boldsymbol{H}_c^Q$ is computed by

$$H_c^Q = round(H_c./QT(c,:)), \tag{3}$$

where ./ is an entry-wise division, $round(X)$ is the operation that rounds the elements of X to the nearest integers. Note that the elements of H_c are greater or equal than the threshold T_{LH}, and the elements of $QT(c,:)$ are less than T_{LH}. Now the range of H_c^Q becomes greater or equal than 1, which facilities compression.

For high amplitude component to be quantized, its scale is determined by the individual signal as well as the threshold T_{LH}. However, such signal recorded from different individuals could obtain divergent results [15]. Whats more, different neural units may have different spectral distribution. Therefore quantization table varies on different channels for different samples. Therefore, the average amplitude of LAC are assigned to QT for each channel.

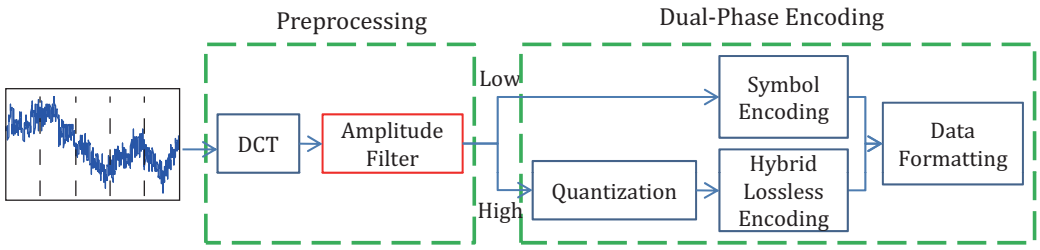

Fig. 3. Flow diagram of the overall compression algorithm.

This designing strategy for QT is proposed from several angles. First, diversity of individual recordings has been taken into consideration by yielding unique QT for each sample. Second, unshared quantizer for each channel takes account of the discrepant spectrum distribution among channels. Third, with the use of round function, the reverted value differs by no more than half of QT on corresponding position. Moreover, this makes all the quantized data less than T_{LH}. As a consequence, the quantized values are no less than one, sufficing the non-zero condition mentioned at the end of section IV-A.

C. Hybrid Lossless Encoding for Quantized Data

As an optimal symbol-by-symbol coding method, the Huffman Encoding yields the optimal length- variable code, which can be efficiently used in our quantized data. However, after partitioning by the amplitude filter, a lot of DCT coefficients at high frequency turn to be zeros. Fig.4 shows the coefficients distribution after quantization. What is more, the zeros at high frequency part often appear consecutively, forming series of zeros. Therefore, at high frequency part, rather than encoding each zero independently, recording the number of consecutive zeros can compress more effectively. This strategy is referred to as Zero-Length-Encoding. Denote the boundary between the low frequency part and the high frequency part by B. Consequently, all the nonzero coefficients as well as the zeros before B are encoded by the Huffman Encoding, while the zeros after B are encoded by the Zero-Length-Encoding. To exploit the two encoding methods effectively, the boundary B should be well-designed, which depends on the distribution of zeros.

In the following, we first give a brief introduction of the Huffman Encoding, and then introduce the Zero-Length-Encoding; finally we investigate how to set the boundary B.

1) Huffman Coding: Entropy Encoding is a lossless compression technique that typically creates a unique prefix-free code to each symbol of a set. As the most common method of Entropy Encoding, the Huffman Encoding [10] is used in our lossless encoding method, aiming at establishing an optimal tree that minimizes the weighted sum of heights (i.e. the total length of code). To transform original values into binary sequence, the Huffman Encoding derives length-variable code based on the estimated occurrence frequency of each symbol.

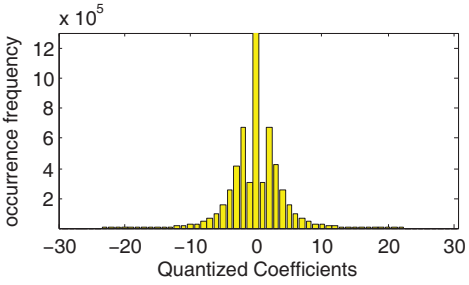

Fig. 4. Quantified Coefficient Frequency distribution, horizontal axis denotes the coefficients value ranging from -31 to 31. The vertical axis is the occurrence frequency of each value in x.

All the entries before B as well as the non-zero components after B are coded by the Huffman Encoding.

2) Zero-Length-Encoding: The Zero-Length-Encoding is used to encode the series of consecutive zeros at high frequency part. The number of zeros k_z of a series is represented by the octal number system, and the Huffman encoding of zero is used to separate adjacent these octal numbers. For example, if $k_z = (8A + B) \times 8 + C$, where A, B, C denote the 3^{rd}, 2^{nd}, 1^{st} order of octal number respectively, the form of code is shown below, where ***HCT***(0) is the Huffman code of zero.

$$\underbrace{\_\ \_\ \_}_{A}\ HCT(0)\ \underbrace{\_\ \_\ \_}_{B}\ HCT(0)\ \underbrace{\_\ \_\ \_}_{C}$$

Let $g(k)$ be the number of orders to represent zero count, we have

$$g(k_z) = \lceil \log_8(k_z + 1) \rceil. \tag{4}$$

3) Boundary between Hybrid Lossless Encoding: To make the best use of the above mentioned two encoding methods, the determination of boundary between them appears especially important.

Intuitively, as the number of zeros in the quantized signal increases, the code given by Zero-Length-Encoding would

be shorter than the code given by Huffman Encoding. To validate this idea, the code length given by the two methods is calculated.

Let ***HCT*** be the Huffman Code Table derived before, ***HCT***(x) means the Huffman Encoding of x , l_0 denotes the length of ***HCT***(0), k_z be the average number of contiguous zeros before a nonzero coefficient. The code length of quantized coefficients by Huffman encoding(l_1) and Zero-Length-Encoding(l_2) is:

$$\begin{cases} l_1 = \sum_{i=1}^{I}[HCT(x_i)] + l_0 \cdot k_z I \\ l_2 = \sum_{i=1}^{I}[HCT(x_i) + (3+l_0)g(k_z) - l_0] \end{cases} \quad (5)$$
$$, \ x_i \in H_c^Q, x_i \neq 0$$

where I is the size of the set including all the nonzero coefficients in H_c^Q. In l_1, $l_0 \cdot k_z I$ denotes the length of all the zeros. In l_2, $(3+l_0)$ is the number of bits required for each additional order. Use $g(k_z)$ to be the number of orders for k_z as defined in eq.(4), then $(3+l_0)g(k_z)-l_0$ is the number of bits denoting the average number of zero coefficient before a nonzero one.

Taking the difference between them, we have

$$l_1 - l_2 = [l_0 \cdot k_z - (3+l_0)g(k_z) + l_0]\, I = f(k_z) \cdot I \quad (6)$$

For constant I, we only consider function $f(k_z)$. Put eq.(4) into (6), we have

$$f(k_z) = \begin{cases} -3, \ k_z = 0 \\ l_0 \cdot k_z - (3+l_0)\lceil \log_8(k_z+1) \rceil + l_0, \ else \end{cases} \quad (7)$$

It can be easily derived that $f(k_z)$ can only be negative at the very beginning of k_z for l_0 no less than 1. With the increasing of k_z, $f(k_z)$ has a tendency of raise, making $f(k_z)$ has only one intersection point with zero. At this intersection, we have $f(k_z) = k_z * l_0 - 3$, i.e.,

$$k_z = 3/l_0 \quad (8)$$

Consequently, the boundary should be set at the point where the average number of continuous zeros is equal to $3/l_0$.

The overall compression Algorithm is shown in algorithm(1). For clearer demonstration, this algorithm considers compression applied on one channel, compression between different channels are independent according to the third character exploited in section II. After dividing the signal into blocks, we firstly calculate the Quantization Table ($\boldsymbol{QT}$) for a sample(line 3-8), where $\boldsymbol{F}_{(i)}$ is the DCT coefficients of $\boldsymbol{X}_{(i)}$ and $\boldsymbol{low}_{(i)}$ is the indices of Low Amplitude Components $\boldsymbol{Lc}$. Then each block sample is compressed with the Dual-Phase Encoding(Line 9-18). $\boldsymbol{S}$ is assigned as the signs of $\boldsymbol{Lc}$ in *Symbol Encoding* (line 11). Let $\boldsymbol{Hc}$ be the high amplitude component, it is first quantized(line 13) then encoded by Hybrid Lossless Encoding. The output of our algorithm includes the compression code $\boldsymbol{Y}$ and the length of *Symbol Encoding*, $\boldsymbol{Z}$, serving as the separator between Symbol Encoding and Hybrid Encoding for decompression.

Algorithm 1: Overall Compression Algorithm

Input: $\boldsymbol{X}$, the signal; S_b, the block size; T_{LH}, the threshold between HAC and LAC; ***B***, the boundary within Hybrid Encoding

Output: $\boldsymbol{Y}$, formatted compression result; $\boldsymbol{Z}$, lengths of *Symbol Encoding* codes for all blocks

1 Divide $\boldsymbol{X}$ into blocks of size S_b, $\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, ..., \boldsymbol{X}_{(N)}$;
2 **for** $i = 1, ..., N$ **do**
3 $\quad \boldsymbol{F}_{(i)} \leftarrow DCT(\boldsymbol{X}_{(i)})$;
4 $\quad \boldsymbol{low}_{(i)} \leftarrow$ find indices $(\boldsymbol{F}_{(i)} < T_{LH})$;
5 $\quad \boldsymbol{Lc}_{(i)} \leftarrow F(\boldsymbol{low}_{(i)})$; %*LAC*
6 **end**
7 $\boldsymbol{QT} \leftarrow$ average over $|\boldsymbol{Lc}_{(i)}|$, $i = 1, ..., N$;
8 $\boldsymbol{Y} \leftarrow [\,]$;
9 **for** $i = 1, ..., N$ **do**
10 $\quad \boldsymbol{Hc}_{(i)} \leftarrow \boldsymbol{F}_{(i)}$; $\boldsymbol{Hc}_{(i)}(\boldsymbol{low}_{(i)}) \leftarrow 0$;
11 $\quad \boldsymbol{S} \leftarrow sgn(\boldsymbol{Lc}_{(i)})$; $\boldsymbol{Y} \leftarrow [\boldsymbol{Y}\ \boldsymbol{S}]$; *%Symbol Encoding*
12 $\quad \boldsymbol{Z}_{(i)} \leftarrow length(\boldsymbol{S})$;
13 $\quad \boldsymbol{H}_{\boldsymbol{c}}^{\boldsymbol{Q}} \leftarrow round(\boldsymbol{Hc}_{(i)}./\boldsymbol{QT})$;
14 $\quad \boldsymbol{H} \leftarrow Huffman(\boldsymbol{H}_{\boldsymbol{c}}^{\boldsymbol{Q}}(1:B))$; $\boldsymbol{Y} \leftarrow [\boldsymbol{Y}\ \boldsymbol{H}]$;
15 $\quad$ **forall the** $x \in \boldsymbol{H}_{\boldsymbol{c}}^{\boldsymbol{Q}}((B+1):end)$ **do**
16 $\quad\quad$ **if** $(x \neq 0)$, $\boldsymbol{Y} \leftarrow [\boldsymbol{Y}\ Huffman(x)]$;
17 $\quad\quad$ **else** $\boldsymbol{Y} \leftarrow [\boldsymbol{Y}\ ZeroLength(x)]$;
18 $\quad$ **end**
19 **end**

V. EXPERIMENTAL RESULTS

A. Dataset

The dataset of our experiments comes from two male rhesus monkeys (Macaca mulatta) by the BMI system at Zhejiang University Qiushi Academy [6]. In this system, each monkey was trained to perform a four-direction centered-out task by turning a joystick according to some prompt. After mastering this task, the monkey was implanted with a multi-electrode array in the primary motor cortex (M1) of its cerebral hemisphere contralateral to track the neural signal as the hand moves. Each experiment takes approximately 60 minutes.

The sample records 106 neurons' signal simultaneously from 96-electrode array, with 16-bit accuracy at a sampling rate of 30 kHz. It produces a data stream with 5.76MB/s. To verify our compression algorithm, we randomly selected 12 records; each has a length about 300s.

B. Criteria Used

We use three criteria to evaluate our compression method: Signal to Noise Ratio, Spike ratio and Compression ratio.

1) Signal to Noise Ratio: In communication theory, Signal to Noise Ratio (SNR) is used to judge the fidelity of compressed data by comparing original signal with the reconstruction error. Let S_o and S_r be the original signal and recovered signal in a channel respectively, SNR is defined by the power of S_o divided by the power of background noise:

$$SNR(S_o, S_r) = 10 \cdot \log_{10} \frac{\|S_o\|_2^2}{\|S_o - S_r\|_2^2} \tag{9}$$

2) Spike ratio: High frequency components play a minor role in SNR. To examine the preservation of spikes, the Spike Ratio is taken into consideration. It measures the ratio of spikes preserved after reconstruction. In our validation, the well-known amplitude threshold technique [16] is used for spike detection, where the threshold (Thr) is set as

$$Thr = \alpha \cdot \sigma_n, \ \sigma_n = median\left(\frac{|x|}{0.6745}\right) \tag{10}$$

where α is a constant factor, σ_n is an estimate of the standard deviation of the background noise. A point is regarded as the beginning of a spike with amplitude higher than Thr. Notice that the spike ratio not merely counts the spikes, but counts the spikes correctly matched.

3) Compression ratio: In addition to measuring the signal fidelity by SNR and the spike preservation, compression ratio (CR) is also taken into consideration, from a data reduction perspective. The compression ratio is defined as the size ratio of compressed file to the original one.

C. Parameter setting

In this section, we investigate the choice of three parameters in our model: T_{LH}, the threshold of Amplitude Filter; ω, the scale of Quantization Table and S_b, the size of a block in preprocessing.

1) The threshold between LAC and HAC: T_{LH}, the threshold between LAC and HAC determines the boundary for *Symbol Encoding* and *Quantization*. As T_{LH} increases, more DCT coefficients are processed by Symbol Encoding, bringing more loss while decreasing compression ratio. Therefore, the determination of T_{LH} can be viewed as a tradeoff between distortion and compression ratio.

2) Quantization Table Scale: Quantization Table (QT) is used by Symbol Encoding in our proposed method, which is designed to be the average magnitude of Low Amplitude Component. Therefore, QT increases if T_{LH} gets higher; and in this case the distortion increases. However, it is interesting to know whether the compression result will be improved by scaling QT, i.e. fine-tuning QT by multiplying different factors ω from 0.5 to 2.5 with equal difference of 0.5.

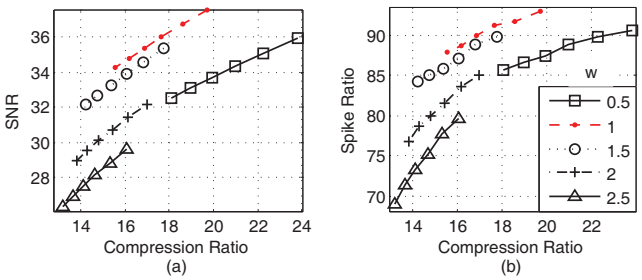

Fig. 6. Comparison of compression performance at different ω. Each curve represents compression performance under a QT scale.

Results under different parameters settings are shown in fig. 5. Each evaluation is the average result on the whole dataset by a systematic variation of T_{LH} and Quantization Table multiplier ω. Accordingly, both SNR and Spike Ratio go down with the increase of T_{LH}, and ω=1 is always optimal in the view of SNR and Spike ratio, and its compression ratio ranks second.

To determine the selection of ω, signal fidelity is compared under fixed compression ratio as shown in fig. 6. For the same compression ratio (at horizontal axis), we see that ω = 1 consistently outperforms the others. Such result validates our setting of Quantization Table.

The choice of T_{LH} depends on our requirement on compression ratio and compression fidelity. For a desired SNR higher than 30db and Spike ratio no less than 90%, T_{LH} = 24 is selected, resulting in an average compression ratio of 17.75% (SNR 36.24dB and Spike ratio 92%).

3) Block size: The size of a block in preprocessing reveals the precision in time domain. All the above experiments are taken on a fixed block size of $S_b = 1600$, but it is still a question how S_b changes compression result. The evaluation is shown in fig. 7, with S_b tested between 1500 to 28500, fixing $T_{LH} = 24$ and ω=1.

This figure illustrates that with the increasing of S_b, the signal fidelity firstly increases and then decreases. This phenomenon can be interpreted in the following way. First, larger S_b brings more refined DCT coefficients, therefore, decreasing the error in decompression. However, more coefficients are involved into the low amplitude component, which means more loss according to the previous analysis. Therefore, the increase and decrease of fidelity reflect the tradeoff between the two factors. According to the given experimental result, an optimal size of block is determined at S_b = 7500, achieving an average compression ratio of 17.7%, SNR of 36.6dB and Spike ratio of 91.9%.

D. Effect of Symbol Encoding

Symbol Encoding focuses on those signals with lower amplitude. We need to exploit whether it brings contribution

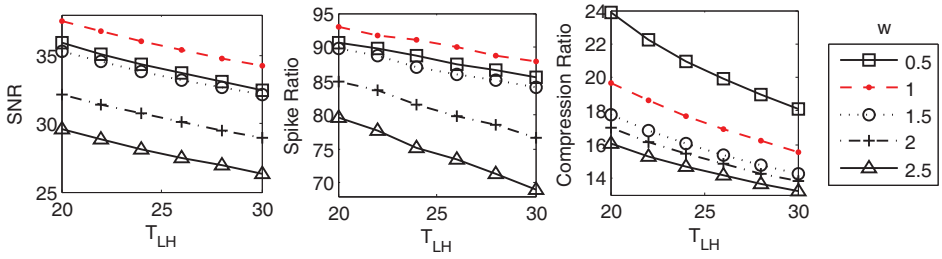

Fig. 5. SNR, Spike Ratio and Compression Ratio for different T_{LH} and QT scales.

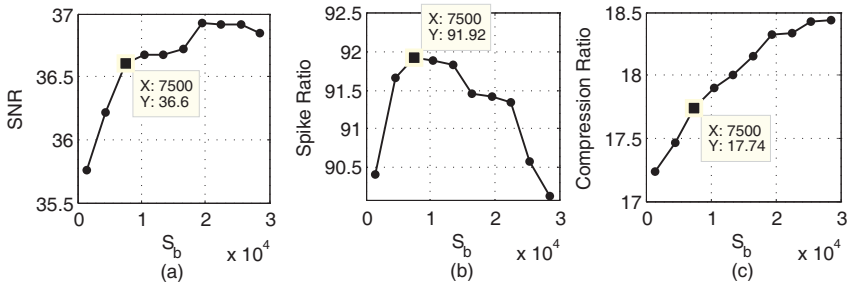

Fig. 7. Compression result of different block size S_b. The horizontal axis is the size of a block.

to signal preservation. Given the parameters selected in the last section, the improvement brought by symbol encoding is shown in Table I.

TABLE I. SNR, SPIKE RATIO AND COMPRESSION RATIO WITH AND WITHOUT SYMBOL ENCODING (LOW-AMPLITUDE)

Symbol Encoding	SNR(db)	SR	CR
without	31.4	83.7%	13.1%
with	36.6	91.9%	17.7%

It is clear that taken the Low-Amplitude part into consideration has to save more information, i.e., leads to higher compression ratio, but the signal fidelity has been greatly improved. This is achieved by the inherent distribution of coefficients in the high frequency part. Therefore, symbol encoding is adopted as an effective way to preserve high frequency components.

E. Comparison with other approaches

Section I has mentioned some relevant work on neural signal compression. However, due to the lack of public data sets and other compression standards on extracellular recording from motor cortex, the general data compression methods and the state-of-the-art audio compression algorithms are considered to be compared with our compression algorithm. By investigating the compression result of both lossless and lossy audio compression algorithm, we find that our compression enables a balance between compression ratio and fidelity.

1) Lossless compression: Lossless compression methods produce exact reconstruction of the original file and storage will be reduced by a more compact coding format. For audio compression, Codecs like FLAC use linear prediction to estimate the spectrum of the signal, achieving a compression ratio of 50%-60% for general waveforms [11]. However, unlike audio signal, neural signal is more complicate and difficult to predict. Likewise, data file compression format such as Zip, 7-Zip and RAR also cannot achieve a relative low compression ratio. Table II shows the compression ratio of different lossless compression techniques. The best compression method for the given neural data is APE (Monkey's Audio), achieving lowest compression ratio 56.88%.

2) Lossy Compression: Different from lossless compression, lossy audio compression takes advantage of human acoustic perception that is only sensitive to specific frequency band and amplitude, and only quantifies and encodes the perceptible parts. As a state-of-the-art audio encoding algorithm, Advanced Audio Coding (AAC) is used on neural signal and compared with our coding method. AAC is a part of MPEG-2 standard and provides better signal quality than MP3 with 30% reduction of file size. Fig. 8 shows the comparison between the two methods.

For audio compression, we use high bitrate ranging from 300kbps to 600kbps, intending to achieve good reconstruction performance. However, the result is not ideal for extracellular recording. Fig.8 illustrates that our method is higher than AAC in both SNR (exceeds by 46.4%) and Spike ratio (exceeds by 80.4%), under the same compression ratio. It implies that the characteristics of neural signal are quite different from those of audio signal, so that conventional methods fail to perform well on the neural signal.

TABLE II. PERFORMANCE OF LOSSLESS COMPRESSION FOR COMPARISON

	Lossless Compression Format					Ours
Configuration	Audio Codec			Archive File Format		SNR=36db Spike Ratio=92%
	Lossless WMA	FLAC	APE	Zip	RAR	
Compression Ratio	70.89%	54.27%	53.08%	70.04%	60.91%	17.74%

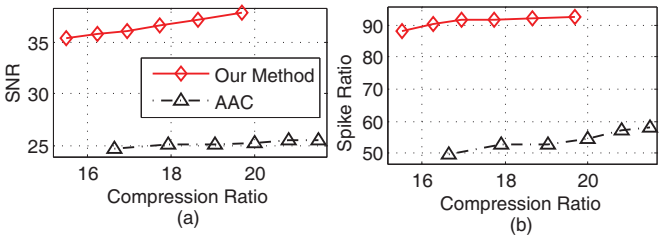

Fig. 8. Compression result comparing between AAC encoding and our compression method.

F. Computational cost

Previous sections have demonstrated the ability of our method on balancing compression ratio and reconstruction error. At last, the compression efficiency is taken into consideration. The following results come from the experiments implemented using MATLAB. The initialization process to get Quantization table, Huffman Code Table and boundary in Lossless encoding needs 2.86Mb/s, the compression consumes 0.13Mb/s and the decompression speed is 0.14Mb/s.

VI. CONCLUSION

In this paper, a lossy compression algorithm for motor cortex extracellular recordings is given. We first exploit the intra and inter channel relationship in motor cortex signal. Then a dual-phase encoding is proposed based on spectrum analysis. The prototype is successfully tested on the sampled Rhesus's motor cortex signals, achieving a compression ratio of 17.7% with SNR values 36.6dB and preserving 92% spikes. The result is remarkable compared with other biomedical signal compression methods, which achieves a SNR of 15-26dB and compression ratio of 1%-20% without consideration of the significant spike signal [7], [8], [13].

Our proposed method is validated on motor cortex extracellular recordings, but it may also be applied to other extracellular recordings with further exploitation.

ACKNOWLEDGMENT

This work was supported in part by the National 973 Program.

REFERENCES

[1] Lebedev, Mikhail A., and Miguel AL Nicolelis. "BrainCmachine interfaces: past, present and future." *TRENDS in Neurosciences* 29.9 (2006): 536-546.

[2] Zhaohui Wu, Gang Pan, Nenggan Zheng. "Cyborg Intelligence." *IEEE Intelligent Systems* 28(5):31-33, Sep/Oct 2013.

[3] Yipeng Yu, Dan He, Weidong Hua, Shijian Li, Yu Qi, Yueming Wang, Gang Pan. "FlyingBuddy2: A Brain-controlled Assistant for the Handicapped." *The 14th ACM International Conference on Ubiquitous Computing (Ubicomp'12) Poster & Video, Pittsburgh, PA, USA, September* 5-8, 2012.

[4] Rousche, Patrick J., and Richard A. Normann, "Chronic recording capability of the Utah Intracortical Electrode Array in cat sensory cortex." *Journal of neuroscience methods*, vol. 82, no. 1, pp. 1-15, 1998.

[5] Kipke, Daryl R., Rio J. Vetter, Justin C. Williams, and Jamille F. Hetke, "Silicon-substrate intracortical microelectrode arrays for long-term recording of neuronal spike activity in cerebral cortex." *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 151-155, 2003.

[6] Zhang, QiaoSheng, ShaoMin Zhang, YaoYao Hao, HuaiJian Zhang, JunMing Zhu, Ting Zhao, JianMin Zhang, YiWen Wang, XiaoXiang Zheng, and WeiDong Chen, "Development of an invasive brain-machine interface with a monkey model." *Chinese Science Bulletin*, vol. 57, no. 16, pp. 2036-2045, 2012.

[7] Chen Han Chung, Yu-Chieh Kao, Liang-Gee Chen, Fu-Shan Jaw, "Intelligent Content-Aware Model-Free Low Power Evoked Neural Signal Compression." *Advances in Multimedia Information Processing*, pp. 898-901, 2008.

[8] Chen Han Chung, Yu-Chieh Kao, Liang-Gee Chen, Fu-Shan Jaw, "Multichannel Evoked Neural Signal Compression Using Advanced Video Compression Algorithm." *Neural Engineering*, pp. 697-701, 2009.

[9] Antoniol, Giuliano, and Paolo Tonella, "EEG data compression techniques." *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 2, pp. 105-114, 1997.

[10] Huffman, D.A., "A method for the construction of minimum redundancy codes." *Proceedings of IRE*, vol. 40, no. 9, pp. 1098-1101, 1952.

[11] Coalson, Josh. "FLAC-Free lossless audio codec". 2009.

[12] Monica Fira and Liviu Goras, "Biomedical Signal Compression based on Basis Pursuit." *Proceedings of the 2009 International Conference on Hybrid Information Technology*, vol. 14, pp. 53-64, Jan 2010

[13] S. Aviyente, "Compressed sensing framework for EEG Compression." *IEEE/SP 14th Workshop on Statistical Signal Processing*, pp. 181-184, 2007.

[14] Birgitta Weber, Thomas Malina, Kerstin M. L. Menne, Volker Metzler, Andre Folkers, Ulrich G. Hofmann, "Handling large files of multisite microelectrode recordings for the European VSAMUEL consortium." *Neurocomputing*, pp. 1725-1734, 2001.

[15] Prentice, Jason S., Jan Homann, Kristina D. Simmons, Gašper Tkačik, Vijay Balasubramanian, and Philip C. Nelson, "Handling large files of multisite microelectrode recordings for the European VSAMUEL consortium." *PloS one*, vol. 6, no. 7, 2011.

[16] Quiroga RQ, Nadasdy Z, Ben-Shaul Y, "Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering." *Neural Computation*, vol. 16, no. 8, pp. 1661 - 1687, 2004.