Journal of Neuroscience Methods 230 (2014) 51–64

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Computational Neuroscience

Minimum requirements for accurate and efficient real-time on-chip spike sorting

Joaquin Navajas[a,*], Deren Y. Barsakcioglu[b], Amir Eftekhar[b], Andrew Jackson[c], Timothy G. Constandinou[b], Rodrigo Quian Quiroga[a]

[a] Centre for Systems Neuroscience, University of Leicester, 9 Salisbury Road, LE1 7QR, United Kingdom
[b] Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, United Kingdom
[c] Institute of Neuroscience, Newcastle University, Newcastle-upon-Tyne NE2 4HH, United Kingdom

HIGHLIGHTS

- We test the feasibility of a two-stage method for real-time on-chip spike sorting.
- To evaluate this procedure, we use realistic simulations of extracellular recordings.
- We study the minimum signal requirements that allow accurate and efficient spike sorting.
- We describe sets of specifications that optimise performance and minimise complexity.

ARTICLE INFO

Article history:
Received 3 January 2014
Received in revised form 11 April 2014
Accepted 14 April 2014

Keywords:
Spike sorting
Extracellular recordings
Real-time
On-line
On-chip
BMIs
Template matching

ABSTRACT

Background: Extracellular recordings are performed by inserting electrodes in the brain, relaying the signals to external power-demanding devices, where spikes are detected and sorted in order to identify the firing activity of different putative neurons. A main caveat of these recordings is the necessity of wires passing through the scalp and skin in order to connect intracortical electrodes to external amplifiers. The aim of this paper is to evaluate the feasibility of an implantable platform (i.e. a chip) with the capability to wirelessly transmit the neural signals and perform real-time on-site spike sorting.
New method: We computationally modelled a two-stage implementation for online, robust, and efficient spike sorting. In the first stage, spikes are detected on-chip and streamed to an external computer where mean templates are created and sent back to the chip. In the second stage, spikes are sorted in real-time through template matching.
Results: We evaluated this procedure using realistic simulations of extracellular recordings and describe a set of specifications that optimise performance while keeping to a minimum the signal requirements and the complexity of the calculations.
Comparison with existing methods: A key bottleneck for the development of long-term BMIs is to find an inexpensive method for real-time spike sorting. Here, we simulated a solution to this problem that uses both offline and online processing of the data.
Conclusions: Hardware implementations of this method therefore enable low-power long-term wireless transmission of multiple site extracellular recordings, with application to wireless BMIs or closed-loop stimulation designs.

1. Introduction

For more than half a century neuroscientists have recorded extracellular action potentials, namely spikes, generated by neurons, in order to understand how information is represented and transmitted through the nervous system (Hubel, 1957). Until recently, these experiments involved sampling small numbers of neurons over short sessions of a few hours, but with advances in chronic electrode arrays it has become possible to record spikes from large numbers of neurons over several months (Harris et al., 2003; Buzsaki, 2004; Quian Quiroga and Panzeri, 2009). These techniques inspired translational efforts to develop brain–machine interfaces (BMIs) that communicate directly with the nervous system for therapeutic benefit (Donoghue, 2002; Carmena et al., 2003;

* Corresponding author. Tel.: +44 01162525540.
E-mail address: jmna1@le.ac.uk (J. Navajas).

http://dx.doi.org/10.1016/j.jneumeth.2014.04.018

Chapin, 2004; Musallam et al., 2004; Velliste et al., 2008; Nicolelis and Lebedev, 2009). For example, neural signals from the motor cortex of paralysed patients have been recently used to operate assistive devices such as computers and robotic prostheses (Hochberg et al., 2006).

The first step for invasive BMI systems (i.e. those involving the use of intracranial electrodes), as for any extracellular recording, is to detect the spikes and then identify which spike corresponds to which neuron in a process called ‘spike sorting’ (Letelier and Weber, 2000; Pouzat et al., 2002; Quian Quiroga et al., 2004; Zhang et al., 2004; Rutishauser et al., 2006; Vargas-Irwin and Donoghue, 2007). For this, spikes are classified according to their shapes, assuming that, in principle, each neuron fires spikes with a stereotyped waveform (Buzsaki et al., 2012; Quian Quiroga, 2012). Data is then relayed to an external computer where the motor commands are decoded based on the activity of several individual neurons (Quian Quiroga and Panzeri, 2009). A main caveat of current invasive BMI systems, and of any setup for extracellular recordings, is the need of wires passing through the scalp and skin in order to connect intracortical electrodes to external amplifiers. Ideally, to avoid risks of infections, among other things, one would like to implant a low-power device with a wireless link transmitting the signal to an external receiver (Aghagolzadeh and Oweiss, 2011; Harrison, 2008; Rapoport et al., 2012). A first strategy to do this is to stream all the recorded data wirelessly for processing in an external computer. However, to perform spike detection and sorting, neural signals are typically sampled at very high rates – around 25 kHz per channel or more – thus producing an immense amount of data to be transmitted (approx. 500 kbit/s per channel). Bandwidth limitations severely restrict the number of channels that can be transmitted, as well as power requirements and consequently battery lifetime. Another alternative is to perform spike sorting directly in the implanted device (Mavoori et al., 2005; Zhang et al., 2012; Zumsteg et al., 2005; Zviagintsev et al., 2005). By doing so, the amount of data to be transmitted is reduced to just binary events signalling the firing of spikes (typically occurring at the order of Hz, compared to the kHz time resolution needed to transmit the raw data). However, given the computational complexity of spike sorting, these methods compromise either performance or hardware practicability.

The goal of this paper is to computationally test the feasibility of a hybrid, two-stage strategy for a real-time, hardware-efficient and robust spike sorting implementation. In the first ‘template building’ stage, the implanted device automatically detects and wirelessly streams spikes to an external computer performing the heavy processing steps of spike sorting (i.e. template definition by feature extraction and clustering). The mean spike templates for each channel are then sent back to the implanted device and, in the second ‘template matching’ stage, spikes detected in real time are compared with these mean templates and assigned to the one with the least distance. Given the low complexity and robustness of template matching, this approach enables low-power processing while preserving the quality of the signal.

To the best of our knowledge, there is only one previous study that proposed the idea of doing real-time spike sorting using downloaded templates from a training session (Rizk et al., 2009). In that study, Rizk and collaborators presented a 96-channel implantable platform with the capability to wirelessly stream detected spikes and perform real-time template matching in an external device. However, one caveat of the implementation proposed in that study – as argued by the authors – is that the power consumption of the chip is far from ideal, and therefore the device does not have enough resources to perform template matching inside the body. Here, we study the extent to which we can reduce the specifications of a hypothetical chip similar to the one described in Rizk et al. (2009) without significantly impoverishing performance. Using a set of synthetic data mimicking real extracellular recordings (Martinez et al., 2009; Camunas-Mesa and Quian Quiroga, 2013), we systematically investigated the minimum data and processing requirements – in terms of filtering, sampling frequency, signal resolution, etc. – that still allows reliable spike detection and sorting, but reducing hardware resource (energy and complexity) requirements. We report that typical specifications of extracellular recordings (e.g. 28 kHz of sampling rate and 16-bit resolution) can be largely reduced without considerably dropping performance. Finally, we validate the choice of these minimum requirements using real data.

2. Methods

2.1. Synthetic extracellular recordings

Simulated extracellular recordings were generated by modelling the contribution of the local field potentials (LFPs), background noise, multi-unit activity and single-unit activity (Fig. 1).

To simulate the LFPs and the background noise, we used surrogates of one real extracellular recording from the human medial temporal lobe (MTL). The subject was a patient with pharmacologically intractable epilepsy who was implanted with intracranial electrodes for clinical reasons (Fried et al., 1997). To record single-neuron activity, the intracranial probe had a total of 9 micro-wires at its end, with 8 active recording channels and 1 reference. The differential signal from the micro-wires was amplified, sampled at 28 kHz and 16-bit resolution with a signal input range of $\pm$1 mV. The size of the electrodes used in this recording was 40 μm, and the signal bandwidth was 0.1–9000 Hz. The recording was performed with a Neuralynx “Cheetah-32” system, which works in conjunction with “Lynx-8” amplifiers. These amplifiers have an input noise of 15 μV_{p-p} and an output noise of 10 mV_{p-p}. A full list of specifications about the “Lynx-8” amplifier can be found in http://neuralynx.com/manuals/Lynx-8_Manual.pdf.

The channel used to create the surrogates did not exhibit single-unit nor multi-unit activity, but had the same power spectrum and amplitude characteristics of neighbouring channels that had both types of activity. Surrogates were constructed applying the Fourier transform, shuffling the phases, and then applying the inverse transform. Here, we used the implementation proposed by Schreiber and Schmitz (2000), which uses a constrained randomisation, thus preserving not only the power spectrum but also the amplitude distribution. Each simulation was built using a different surrogate.

Synthetic multi-unit and single-unit activity were added to the LFP and background noise with the implementation described in Martinez et al. (2009). The simulations were created using a database with 594 different average spike shapes, taken from real recordings from the monkey neocortex and basal ganglia. Multi-unit activity – i.e. spikes that can be detected but cannot be clustered into different single units due to their relatively small amplitude – was created by mixing the activity of the whole database of 594 spikes shapes, leading to multi-unit amplitude uniformly distributed between 20 μV and 40 μV. The multi-unit firing rate was set at 20 Hz. Single-unit activity was generated by adding spike shapes with varying amplitudes (as specified below) to the background noise and LFP. The spike shapes added to the signal were 92 samples long and had a smooth decay to zero in order to avoid introducing edge artefacts. The spike train of each single-unit followed a Poisson process, with a mean firing rate of 5 Hz. Spikes that fell within a 2 ms window following another spike were removed to introduce a refractory period and delete overlapping spikes.