A Low-Power Integrated Circuit for Analog Spike Detection and Sorting in Neural Prosthesis Systems

A. Bonfanti¹, T. Borghi¹, R. Gusmeroli¹, G. Zambra¹, A. Oliyink², L. Fadiga², A. S. Spinelli¹, and G. Baranauskas³

¹Dipartimento di Elettronica e Informazione, Politecnico di Milano – IU.NET, Milano, Italy,
²Dipartimento di Scienze Biomediche e Terapie Avanzate, Università degli Studi di Ferrara, Ferrara, Italy,
³Department of Robotics, Brain and Cognitive Sciences, IIT, Genova, Italy

Abstract—Since the proof of viability of prosthetic devices directly controlled by neurons, there is a huge increase in the interest on integrated multichannel recording systems to register neural signals with implanted chronic electrodes. One of the bottlenecks in such compact systems is the limited transmission data rate of the wireless link, requiring some sort of data compression. We propose an analog low power integrated system for action potential (AP) detection and sorting capable of reducing the overall throughput of more than 100 times. In this system, AP detection is performed by a double threshold method that reduces the probability of false detections while AP sorting is based on the measurement of peak and trough amplitudes and peak width. The circuit has been implemented in 0.35 – μm CMOS technology with power consumption of 70 µW per channel including the pre-amplifier. The system was tested with real recorded traces: compared to standard AP sorting techniques, the proposed simple AP sorter was able to correctly assign to single units over 90% of detected APs.

I. INTRODUCTION

In recent years, a number of laboratories have shown that signals attained with multichannel extracellular recordings are sufficient to control the movements of a prosthetic device in real time [1]. However, we are still far away from any device to be clinically tested because of enormous technological and scientific challenges [2]. According to current thinking, a future neural prosthetic device should contain a set of neural amplifiers connected to a signal-processing unit and a wireless link to transmit the amplified signals down to an actuator (e.g., a robotic arm, a remote controller, a mouse, and so on). Unfortunately, stat-of-the-art wireless systems cannot afford the high data rate required for this application with limited power budget (800 µW/mm²): for example, 100 channels sampled at 30 kHz with 8 bits of resolution would require a 24 Mbps data rate. Data compression is thus a primary need to enable transmissions from neural implants. The most efficient way to compress neural signals is to transmit a single bit for each action potential. Investigators in [3] used simple threshold detectors to find APs in the signal and then transmitted zeros when no AP were detected and ones for each threshold crossing event. However, the efficiency is traded off with the accuracy, since each recording site does usually collect the activity from several close neurons and the aforementioned compression strategy cannot give information on which neuron fired at a given time. Such a drawback limits the use of simple compression techniques, especially in motor prostheses applications where it is important to isolate the activity of each neuron to predict the intended movement [1].

This paper presents a low-power analog compression algorithm that deals with the problem of neuron discrimination and is able to reduce the bandwidth by at least a factor of 100. Neural signal compression is based on the extraction of three parameters for a single spike, namely the peak and trough amplitudes and the width of the peak. Although the use of peak detectors to extract these features has been previously described [4], in this work we introduce significant innovations: a) the implementation of a new spike-detection algorithm [5] that decreases several-fold the rate of false detections; b) the optimization of the algorithm that now enables the peak amplitude detections over a wide range and c) the use of an additional feature, namely the peak width. Moreover, the amplification of the neural signal is performed with an amplification topology that provides the best trade-off available in the literature between noise and power consumption. We tested our system employing artificial and real signals and we conclude that, in the majority of cases, with few independent units present in a trace, the extracted three parameters can be effectively used to achieve single-units separation.

II. SYSTEM ARCHITECTURE AND IMPLEMENTATION

The system architecture is shown in Fig. 1, and comprises a low-noise amplifier (LNA), a spike detector and a spike sorter. The spike detector splits the amplified signal in two paths and exploits a double-threshold strategy, which will be discussed in detail in the following. We now highlight the key performance of each block.

A. Low-Power and Low-Noise Neural Amplifier

Pre-amplification is obtained through a double-stage active filter, shown in Fig. 2. The first stage is an ac-coupled inverting high-pass filter, employing two MOS-bipolar pseudoresistors as feedback elements [6]. Mid-band gain is set to $G_1 = -C_1/C_2 = -66$, with $C_1 = 10$ pF and $C_2 = 150$ fF. The higher cut-off frequency is determined by the gain-bandwidth product of the first operational amplifier ($GBW/P_1$) and is set to about $GBW/P_1 = 20$ kHz, while the high-pass
pole frequency is set to about 10 Hz (neural signal bandwidth ranges from about 100 Hz to 10 kHz). The second stage was designed to reject the first stage offset while providing further signal amplification and filtering. This stage employs the same pseudoresistor elements in the feedback path, with a gain $G_2 = -50$ (achieved with $C_3 = 7.5 \, \text{pF}$ and $C_4 = 150 \, \text{fF}$). The operational amplifier in the first stage was designed in order to have an input-referred noise in the pre-amplifier band lower than $5 \, \mu V_{\text{rms}}$. Furthermore, since this circuit is thought to be used in an implantable system with a large number of electrodes, the current consumption was accurately minimized (to approximately 4 $\mu A$) designing a telescopic operational amplifier, which exploits the best trade-off between power consumption and noise. Details of the implemented operational amplifier can be found in [7].

B. Spike-detection Circuit

The amplified neural signal is sent to a spike-detection circuit. The spike detector is based on the two-threshold detection method, using two different comparators, one for the peak and one for the trough. The main difference with the traditional architecture is the low-pass filter implemented by a $G_M - C$ cell in the trough detection channel (see Fig. 1). The proposed algorithm, called two-threshold with filtering [5], improves the spike detection up to two-times three with respect to the traditional two-threshold method. A spike is detected only if a trough is found within a time window after the peak arrival. When the amplified signal overcomes the positive threshold $V_M$, the comparator output triggers the time window generator providing a positive impulse of selectable duration. Similarly, the signal filtered by the low-pass cell is compared to the negative threshold $V_L$ by a second comparator. The two channel outputs are then sent to a D-type flip flop, acting as AND gate. Its output is high (i.e. a spike is detected) only when the negative threshold is passed and the TIME WINDOW signal is up. The SPIKE DET signal extends from the negative threshold crossing to the end of time window. In the measurements presented in Section IV, the time window duration was set to 2 ms, while the low-pass cut-off frequency of the $G_M - C$ filter was set to 5 kHz.

C. Spike-sorter Circuit

The amplified signal is also processed by a spike-sorter circuit (see Fig. 1) that measures the peak and the trough amplitudes of the spike and the width of the peak. The peak detector circuit is shown in Fig 3(a). If the ENABLE signal is low, the circuit acts as a traditional voltage follower, while when ENABLE is high the source follower $M_6$ is biased with the small leakage current of transistor $M_7$ and it becomes only able to follow the rising edge of the signals. In the implemented circuit, the ENABLE signal comes from the time-window generator output: when the amplified signal crosses the positive threshold, the time-window signal is set high, forcing the detector to track the peak. This control signal allows detecting peaks with very different amplitudes and very close to each other, that was not possible with the implementation in [4] because of the slow discharge after a peak detection. Two aspects were particularly taken into consideration in the design of the peak detector. First, the $M_1 - M_4$ amplifier gain was maximized (to about 50 dB) in order to reduce the systematic offset of the detector (which causes a systematic error in the peak amplitude evaluation). Second, the gate-source capacitance of the follower transistor $M_6$ was minimized in order to keep low the voltage drop caused by the input voltage that drops below the peak voltage. The trough detector is the symmetrical PMOS implementation of the peak one. Furthermore, the enable signal for the trough detector is the spike-detector output itself. The width detector circuit, shown in Fig. 3(b), is a time-to-amplitude converter, based on the charging of a capacitor with a constant current, with two control signals, ENABLE and RESET. These two signals are the positive spike comparator output (POS SPIKE in Fig. 1) and the time-window signal (TIME WINDOW), respectively. The crossing of the positive threshold activates the current generator, $M_5$, and switches off $M_6$. The capacitor $C = 10 \, \text{pF}$ is charged by a 30 nA constant current until the end of the positive spike signal. Then, the capacitor voltage is kept constant to a value proportional to the width of the spike peak, until the time window signal goes down discharging the capacitor. The gain of the width detector is about 3 V/ms.

III. CIRCUIT CHARACTERIZATION

A test structure with pre-amplifier, spike detection circuit and analog sorter was implemented in 0.35-$\mu$m AMS CMOS.
process occupying a total area of 0.24 mm² (see Fig. 4). The LNA was characterized using an HP35665A Dynamic Signal Analyzer. The measured transfer function is depicted in Fig. 5, showing a mid-band gain of 71.5 dB and a 29 Hz–22 kHz band. The same figure shows the transfer function of the LNA plus $G_M - C$ filter implemented in the trough path of the spike detector. The upper cut-off frequency is set to about 5 kHz. The input-referred noise in the LNA band is about 4 $\mu V_{rms}$, while the measured current consumption is 4.7 $\mu A$ for the pre-amplifier first stage and 0.5 $\mu A$ for the operational amplifier in the second stage. The overall Noise Equivalent Factor (NEF) of the LNA is 2.45, which is the best result reported to date [7]. Also, the spike detection circuit was designed in order to keep the power consumption very limited: each comparator draws 2 $\mu A$ while the $G_M - C$ filter has a current consumption of about 0.1 $\mu A$. The window generator dissipates only when the positive threshold is crossed: for a firing rate of 100 Hz the average current consumption is 10 $\mu A$. The peak and the trough detector draw 2 $\mu A$ each, while the width detector has a negligible consumption. The power dissipation of the whole system (output buffers excluded) is about 70 $\mu W$ considering a 3-V power supply. In order to verify the correct detection of the three features (peak, trough and width amplitudes), simulated spikes signals were first fed into the integrated circuit using a TTI TGA12104 arbitrary waveform generator. The amplified signal was recorded as well as peak, trough and width amplitudes. A snapshot of the measured signals is shown in Fig. 6. The three features measured by the spike sorter were compared with the same features directly extracted from the acquired LNA output signal using MATLAB. For amplified spikes with 400 mV of peak amplitude the error was never larger than 30 mV. This corresponds to an error of 8 $\mu V$ for a 106 $\mu V$ peak amplitude input spike. In the same way we obtained a maximum error of 50 $\mu S$ for spike width of 500 $\mu S$.

IV. Test with Signals

To test the efficiency of the spike sorter, we used a variety of traces obtained during in-vivo recordings in rats and monkeys employing tungsten electrodes (impedance in the range of 0.5 $- 2$ MΩ). The signals were downloaded to the waveform generator and then fed into the IC through an attenuation filter to simulate the real amplitude of the neural signal and the input resistance of typical electrodes. For AP detection and sorting, the peak threshold was set to about 4 times the rms background noise while the trough detector threshold was 0.6 times the

Fig. 3. Simplified schematic of the peak detector (a) and width detector (b).

(a) Peak detector

(b) Width detector

Fig. 4. Die photo of the proposed circuit.

Fig. 5. Measured transfer functions for the LNA amplifier (22-kHz bandwidth) and for the LNA + $G_M - C$ filter (5-kHz bandwidth).

Fig. 6. Snapshot of the measured signals of the analog sorter: amplified neural signal (gold line), peak detector output (red line), trough detector output (blue line) and width detector output (green line).
peak threshold [5]. All these tests yielded similar results, so that here we only describe a typical case of monkey recording trace (see Fig. 7). We compared the chip sorted waveforms with the results obtained employing the Principal Components (PCs) analysis of AP waveforms and identifying clusters of events in the space of the three largest PCs [8]. In this case, the AP waveforms were obtained by detecting the events of threshold crossing and collecting about 0.3 ms of trace before and 1.3 ms after the event. Since we used real data, there is no assurance that spike shapes are sorted correctly. Therefore, to assess the quality of unit separation, we used spike occurrence auto- and cross-correlograms [9]. The method is based on the fact that each neuron has a certain refractory period of 2–3 ms after firing a spike during which no action potential can be generated. Thus, if all spike shapes are derived from a single neuron, there should be no action potentials occurring at intervals less than the refractory period of the neuron. This absence of spikes at brief intervals will correspond to no spikes around the midpoint (0 ms interval) in the auto-correlogram (Fig. 7). In the case of a pure noise signal, the probability to detect events around the midpoint (i.e.: events separated by small time intervals) would be equal to the probability of detecting events separated by large time intervals. Thus, by comparing the frequency of events at very short intervals with the events corresponding to very long intervals it is possible to estimate how good is the achieved separation of units. In the example shown in Fig. 7, both methods were able to detect two units and the remaining events turned out to be noise. While in PCA analysis both units has clean refractory periods with no events for intervals of < 3 ms, the chip separated units has a small background noise (Fig. 7). However, compared to the large interspike intervals, the frequency of these short interspike interval events was very low (< 10 fold) suggesting that less than 10% of identified units were misclassified. Similarly, the number of events detected by our spike sorter was > 90% of the ones detected by PCA analysis confirming that our simple sorter is able to achieve 90% efficiency of the PCA analysis method.

V. DISCUSSION AND CONCLUSIONS

The described analog integrated circuit for on-line AP waveform separation was designed with intention to be a part of a compact integrated multichannel system where low power consumption, small size and data compression are key features. The bandwidth reduction by means of the three selected features does obviously depend on the neurons firing rate: in a "worst case" of 100 Hz of firing rate per neuron we would have a 100 fold data reduction if compared to a channel continuously sampled at 30 kHz. The three features selected to perform the discrimination of different spikes were proved to work with real signals and were chosen in order to be implemented with analog circuits. Our implementation overcomes some of the problems of similar systems [4]. Moreover the proposed architecture limits the features extraction only in correspondence of detected spikes (see Fig. 6) and thus enables the tracking of peaks with very different amplitudes for spikes fired close to each other. Finally the circuit has an extremely low-power consumption and the overall power density (< 300 µW/mm²) makes this circuit suitable to be used in multichannel implantable systems.

VI. ACKNOWLEDGMENT

This work was partially supported by IIT (Italian Institute of Technology).

REFERENCES