Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 6, June 2013, pg.11 – 18

SURVEY ARTICLE

A Survey on Brain–Machine Interface used in VLSI Field-Programmable Mixed-Signal Array

Umayal.S[1]

[1]Assistant Professor, Department of ECE, PSNA CET, India

[1] *Umayal.16@gmail.com*

Abstract— A very large scale integration field -programmable mixed- signal array specialized for neural signal processing and neural modeling has been designed. This has been fabricated as a core on a chip prototype intended for use in an implantable closed-loop prosthetic system aimed at rehabilitation of the learning of a discrete motor response. The chosen experimental context is cerebellar classical conditioning of the eye-blink response. The programmable system is based on the intimate mixing of switched capacitor analog techniques with low speed digital computation; power saving innovations within this framework is presented. The utility of the system is demonstrated by the implementation of a motor classical conditioning model applied to eye -blink conditioning in real time with associated neural signal processing. Paired conditioned and unconditioned stimuli were repeatedly presented to an anesthetized rat and recordings were taken simultaneously from two precerebellar nuclei. These paired stimuli were detected in real time from this multichannel data. This resulted in the acquisition of a trigger for a well- timed conditioned eye-blink response, and repetition of unpaired trials constructed from the same data led to the extinction of the conditioned response trigger, compatible with natural cerebellar learning in awake animals.

Key Terms: - Brain–machine interface; closed-loop; field-programmable; learning; neuroelectrophysiology; neuroprosthesis; prosthesis; very-large-scale integration (VLSI)

I. INTRODUCTION

WHERE brain functions are impaired through brain damage or through degeneration caused by aging, it may be possible to develop prostheses which could interact with the brain in order to replace this functionality. While existing neural prostheses either provide input to the nervous system (e.g., cochlear prostheses [1], deep-brain stimulators [2], etc.) or take output from it (e.g., motor cortical prostheses [3]), a largely unmet challenge is the creation of devices that take input from the brain and provide output to it, in order to replace or supplement the functionality of a circuit internal to the brain, although software-based prototypes are appearing [4], [5].

The aim of the European ReNaChip project [6] was to provide a proof of concept for such a closed-loop prosthetic system. The cerebellum was chosen as a target brain area because its well-de fined inputs and outputs facilitate physical interventions while its relatively simple internal structure have proved fertile grounds for neural modeling from Marr onwards [7]. Eye- blink conditioning was chosen as a well-studied target behavior against which success can be measured. It is intended that the replacement system should be biomimetic, i.e., its architecture and functionality should mimic the characteristics of the area which it replaces according to a neural model of the behavior of the area. While the system is not specifically intended for clinical application, there has been a focus on practical constraints such as miniaturization and power constraints for implantability. The project has involved electrode design, neurophysiology, modeling of cerebellar learning, signal processing methods, real- time system integration and chip design. This article focuses on chip design, particularly how a

field-programmable mixed-signal array is used to fulfill the computational requirements. First, in Section II, the target system is described, including: the eye-blink paradigm; electrode placements for recording and stimulation; signal processing methods for real-time extraction of stimulus related events from neural recordings; and the model of cerebellar function which allows online learning. Then, in Section III, the chip prototype is introduced and its features explained. The key experiment by which the performance of the developed circuitry is demonstrated is the real-time acquisition and extinction of a learned timed response based on *in vivo* recorded data, for which methods and results are presented in Sections IV and V, respectively.

II. **TARGET PROSTHETIC SYSTEM**

A. Eye-Blink Conditioning

Eye-blink conditioning is a form of classical conditioning that is commonly investigated with the delay paradigm [8]. An auitory stimulus (conditioned stimulus—CS) and air-puff to the eye (unconditioned stimulus—US) are applied according to the timing scheme in Fig. 1(a) (bottom), in which the CS onset pre-cedes the US onset by an inter-stimulus interval (ISI) of a few hundred ms and the two stimuli then co-terminate. A US alone causes the subject, whether human or rodent, to blink; this is called an unconditioned response (UR). After many repetitions of these paired stimuli, however, the subject learns to blink in response to the CS, prior to the US, at an appropriate time to anticipate the aversive stimulus; this is called a conditioned response (CR). It is known that the cerebellum is necessary for this learning to occur [9]. The target structure for replacement, therefore, is a microcircuit of the cerebellum.

The cerebellum has two inputs and one output, as shown in Fig. 1(c). Inputs related to all sensory stimuli come from the pontine nucleus (PN) while sensory inputs related to inherently aversive stimuli (US) also come from the inferior olive (IO). Both inputs arrive at the Purkinje cells (PU). Output from PU is inhibitory to the deep cerebellar nuclei (DN). A learned timed response manifests itself as activation of specific DN cells, from where signals go to premotor nuclei including the red nucleus and on to motor nuclei, such as the facial nucleus (FN) from where, in the case of this paradigm, an eye-blink is elicited.

The intended overall system is shown schematically in Fig. 1(a). Recording electrodes are inserted in PN, where a neural response to the CS can be detected, and in IO, where a response to the US can be detected. The signals from the recording electrodes are amplified and go through various stages of filtration (as detailed in the figure caption and Section IV-B), resulting in detections of CS and US events. These are input to a model of cerebellar function, whose output may be a timed response to a CS event. This output (the modelled CR) triggers a stimulator which elicits an eye-blink (behavioral CR) through an electrode implanted in FN. The system is therefore meant to bypass and emulate the neural circuitry that implements learning and effects the appropriately timed response. The following sections provide more detail on the aforementioned parts of this system.

B. Event Detection

The signals from the electrodes are treated as multiunit; i.e., the aim is to detect energy related to a population of spikes rather than to identify spikes from particular neurons; an increase in energy is observed in response to the stimuli, which is typically sustained in the case of PN [10] and phasic in the case of IO. The signal is amplified $(\text{gain} \approx 10000\times)$ and filtered in the frequency band associated with spikes (typically 300–3000 Hz), resulting in traces of magnitude $\approx$0.1 V rms. For the multichannel electrode in the PN, the signals are summed together according to a weighting calculated offline, based on the quality of event detection that can be obtained from each channel separately. Then signals are rectified and band -pass filtered to yield a measure monotonically related to the energy over a small window of time (the energy envelope), and a threshold is applied to yield onsets and offsets of detected events. The high cut-off frequency of the band-pass filter is a compromise be-tween the need to detect events immediately to act on them in real-time, and the need to aggregate more information over a longer period to make better detections. The low cutoff frequency is not critical but removes long-term drifts in the back-ground energy in traces, as can be observed in acute experiments with anesthetized animals. For PN, where detections may last a There are 500 components of the various types; this is there-fore a fine-grained design, (whereas most commercial designs have offered a small number of components [26], [27], [29]), and the intention is to operate with many small, low-quality components, using a combination of calibration and pooling of components to deliver accuracy where it is required. For details of the core architecture see Fig. 2(c).

Limitation of power consumption is a major concern for implantable hardware and a prominent reason for working with analog circuitry. In the following four sections, key aspects of this design are described that limit power consumption and otherwise make it fit for the domain of neural signal processing and neural modeling.

These aspects are: switched capacitor optimization (Section III-B); current control (Section III-C); leakage limitation (Section III-D); and the mixing of analog and digital signals (Section III-E). Then Section III-F, shows how rectification is performed, as an example computation which utilizes all components and which is part of the signal processing chain of Section II-B.

C. Switched Capacitor Optimization

The choice of SC circuitry allows great flexibility but is not ideal for power consumption, since repetitive charging and discharging of clock nodes can pass significant current with respect to the charging and discharging of the voltage-mode signal nodes that they act on. Nevertheless there is much that can be done to limit power consumption. Firstly, CSCs are clocked by a single signal and each contains a state machine for locally generating a pair of non-overlapping pulses in response to a rising edge [Fig. 2(b)]. These halves the power used in charging and discharging clock nodes compared to transmitting the two non-overlapping clocks on separate wires. Secondly, clocks are not global but rather generated by PGNs and routed only to where they are needed. The CSCs take their clock signals from the programmable matrix, also allowing them to pass single packets of charge in response to irregular events generated elsewhere within the array; this has possible uses in neuromorphic modeling, a novelty which sets this design apart from other SC FPAAs, but which is not exploited in this article. The aforementioned state machine is insensitive to the slew rate of the clock, thus reducing the requirement for the strength of the driver of the clock signal, which needs to source and sink current only just fast enough to charge and discharge the clock node once per cycle. (The state machine is based on the slew -rate insensitive D-type flip-flop of [30]). This can reduce the effect of clock noise in the system, since clock nodes typically slew much more slowly than in digital systems, meaning that driven nodes onto which these signals are coupled may have much smaller transients as a result. Thirdly, PGNs can be enabled by routed digital signals, thus processes that are active with only a short duty cycle (there are many within the cerebellar model; see Section IV-D) may consume much less power than if they were continuously clocked.

D. Current Control

The signals involved in the initial stages of the chain of filters must pass signals of up to 3 kHz, implying a Nyquist rate of 6 kHz and a clock frequency for CSCs significantly higher (the core has been designed for frequencies up to $\approx$100 kHz). Later stages in the process have high cutoff frequencies on the order of just 1 Hz, and the cerebellar model of Section II-C needs a slowly ramping signal representing PU activation [trace 2 of Fig. 1(d)] which decreases over a period of order 1 s, for which clocked processes of order 10–100 Hz may be sufficient. There is therefore a range of greater than 3 orders of magnitude of different frequencies of operation and it should be possible to set the currents associated with these various processes appropriately so as not to waste power. The core is divided into 10 bands of components, each of which has associated bias currents which can be set to bias the AMPs, the CSCs' state machines, and the CLBs (Section III-E). It is intended that different circuitry operating at different speeds be placed within these bands, so that only those components with a high speed requirement are run at high power. The 24-bit programmable current generators of [32] have been reworked for SRAM programming. The currents are used both to bias components and to drive oscillators in the PGNs. The current of each generator can be individually altered over several orders of magnitude from a master current of $2\,\mu\Lambda$ down to < 1 pA, producing oscillator frequencies from $\approx$100 kHz down to $\ll$1 Hz. Taking the aforementioned slowly ramping PU-activation signal as an example, this was constructed as a SC integration [31], with a PGN driving a CSC, an AMP for active operation and a CLB (Section III-E) controlling the activation of the ramping. The PGN was biased at 330 pA, giving a frequency of $\approx$100 Hz, which (for chosen capacitor ratios) set the speed of the ramping. (Other less critical biases were set in a similar range: 3 nA for the AMP and 250 pA for the CLB and CSC).

E. Leakage Limitation gs

Since some signals, e.g., the level of PU activation, are in-tended to vary with a time constant of order 1 s or below, the leakage of charge through switches to such nodes becomes a cause for concern. Leakage is reduced in a mode suggested by [33]. The chip has two pairs of power rails, an inner and an outer pair. The outer pair, *vdd* and *gnd*, are separated by a standard 3.3 V, whereas the inner pair are offset by programmable voltages from the outer pair, e.g., to 3.1 V and 0.2 V respectively. All in-puts to the programmable interconnect are powered by the inner power rails and are thus constrained to remain between them, whereas the SRAM cells which control the T-gate switches are powered by the outer rails. This means that if a node required to carry a stable voltage is separated from other nodes carrying unknown voltages by a switched off T-gate, V is guaranteed to be a maximum of 0.2 V (for the NMOS), thus limiting the cur-rents through the transistors to the fA range. A suitable choice of the offset voltage at each power rail can reduce the currents through the transistors until they are comparable to the reverse diode leakage current, which ultimately limits the stability of a node. The use of inner and outer power rails to reduce leakage has been demonstrated in a different context in [34, Sec. 3B]. Measurements on this chip show that a typical net consisting of 30 routing wire segments and

only parasitic capacitance can achieve a leak as low as 35 mV/s, a 200-fold reduction com-pared to when no offset is used. Thus this technique can re-duce leakage by orders of magnitude and allow voltages stored on capacitors to remain almost stable over time scales relevant for neural modeling. For this, a proportion of the voltage range available for analog computation has been sacrificed. The transistor - level design of the AMP component is given in Fig. 3 as an example of how the dual power rails are utilized. It is a single-ended output amplifier based on a standard rail- to- rail topology but is altered so that its output stage is limited to the inner power rails whereas its input stage operates between the outer power rails, optimizing linearity over the input range.

F. Intimate Mixing of Analog and Digital Signals with Asymmetric Logic

Digital logic is used to supplement analog computations where required. For example, in the model described in Section II -C, the direction of synaptic plasticity depends on the timed convergence of direct and modulatory inputs on synapses from CS and US signals respectively; such a decision can be implemented with a logical AND gate. Digital circuitry also allows the building of stable binary-valued memories of arbitrary precision, e.g., to store the weight value in the model. The CLB component allows these possibilities. In search of a simple flexible design, the CLBs have been placed in the same matrix of programmable interconnect as the other components [Fig. 2(c)], such that any component can act as an input to any other, e.g., an AMP implementing a threshold can act as an input to a CLB.

A standard approach to power reduction in digital logic is to increase the slew rate of signals so as to reduce "crowbar cur-rent." This is the current which flows through a logic gate, e.g., an inverter, when its input is not saturated at one of the power rails. In a system where analog signals may be used as digital inputs, slew rates may be arbitrarily slow, and thus a different solution is required. The CLB design [Fig. 2(b)] has been de-scribed in [30]. To summarize, this uses starved logic gates to limit crowbar current. As AMPs and the state machines of CSCs can be biased to define their speed of operation, the maximum currents that flow through the digital gates of the CLBs are like-wise programmable, also defining their intended speed of operation. The logic gates are starved asymmetrically, and this asymmetry allows useful circuits such as a D-type flip -flop which is insensitive to the slew rate of its clock, and a CLB configuration which checks the digital saturation of an input, as used in this experiment; see Section III-F.

Outputs of the CLBs are all current-starved in one direction, such that digital signals are allowed in the programmable matrix which transition upwards quickly but downwards more slowly (according to how they are biased). More generally, signals in the matrix are driven by currents which can vary over many orders of magnitude or which are driven only by switched capacitors and therefore undriven between pulses. This introduces several possibilities for signals with large and/or fast swings to couple capacitively to other signals which may be sensitive to noise. Capacitive coupling mainly occurs in the routing matrix and is especially problematic when two signals run alongside each other on parallel wires for long distances. Section IV-B gives an example in which a filter design was selected specifically to avoid such a problem. It is also possible for sensitive signals to be protected by the routing algorithm, for example by being flanked by grounded wires, though with an additional re-source cost.

G. Full-Wave Rectification

Rectification, as required in the chain of signal processing leading to event detection, is given as an example of how the components described above can be used together to perform computation. Fig. 4(a) shows a rectifier circuit, which uses the same principle as [35]. It is based on the active low pass filter circuit shown in Fig. 4(a) (inset), which is mapped into the components previously described. CSC1-2 act as R1 -2, respectively, and CSC3 acts as C1 (for clarity, the diagram shows only the ports of components which are used). *InputOffset* is a voltage bias at the level around which the input signal *In* is centered. PGN provides the regular pulse stream which drives CSC1-2. Using the same clock for both components simplifies the setting of the gain and cutoff frequency of the filter to a matter of adjusting the ratios of capacitance in CSC1-3. Each CSC shown here may be composed of more than one physical component wired in parallel in order to achieve the desired capacitance. AMP1 determines whether *In* is above *InputOffset*. AMP2 ap-plies further positive feedback to sharpen the previous decision. CSC1-2 act in lossless mode [31], with their ground set to a voltage bias *OutputOffset*. This bias is set in a calibration phase to a level which compensates for any systematic offsets due to mismatch, to deliver an output centered around the desired voltage (as will be described in Section AVID). CSC2 acts as a transresistance, whereas the output of AMP2 is used as the "polarity" of CSC1 (a specialization of the CSC component, which is controlled by the input labelled "P," such that $\phi x/y$ are $\phi 1/2$ or vice versa), so that CSC1 either acts directly as a transresistance when *In* is below *InputOffset*, giving negative or inverting gain, or otherwise acts as a negative transresistance, giving positive gain, effectively rectifying the input. The CLB is programmed with the XNOR function to act as a logic level detector as in [30] on the polarity, disabling the

pulse generator when *In* is close to *InputOffset* to prevent an intermediate po-larity input to CSC1 causing an improper switch sequence. An example output from the chip is shown in Fig. 4(b) (the PGN operated at $\approx$50 kHz and the filter was programmed and calibrated for a gain of $\approx 22\times$); additional phase shift can be seen, as well as clipping at the bottom towards the threshold due to the clock disablement; however, performance is more than adequate for its subsequent use in energy detection.

III. METHODS

A. Electrophysiology

The data was selected from a batch of six electrophysiology sessions. In each session an anesthetized rat had a three -twisted platinum wire (California fine wire) electrode inserted into the PN to detect the CS and a 5 MΩ tungsten needle electrode (A-M Systems, USA) or a stainless steel entomological pin #000, insulated except for 0.15 mm tip, into the IO to detect the US. These electrodes were connected to a standard amplification system (MCP-plus, Alpha-Omega, Israel) which applied $10000\times$ gain and Butterworth filters: two-pole high-pass at 300Hz; four- pole low- pass at 3000 Hz. The four signals were then digitized at 14286 Hz per channel with a standard sampling system (Power1401, CED, U.K.). The CS was a white-noise stimulus of 67–70 dB for 470 ms delivered through a hollow ear -bar of a stereotaxic head holder to the right ear. The US was an air-puff of 1.5 bars at source for 100 ms delivered through a nozzle about 2 cm from the right eye. The ISI was 370 ms, such that CS and US co-terminated. 60 paired CS-US trials were delivered, with an inter-trial interval (not including CS duration) of 8 s. The rat was sacrificed and electrode locations were confirmed with histology. All procedures were approved by the Tel Aviv University Animal Care and Use Committee (P-05-004).

B. Simulation of Event Detection and Parameter Setting for Model

The signal processing was conceived as a chain of first -order filters, where the first in the chain was rectifying as in Section III- F, with cutoffs for IO at 3000 Hz LP (rectifying); 30 Hz LP; 6.4 Hz LP; 1 Hz HP. The 30 Hz step was added in order to avoid extreme capacitor ratios in the step down to 6.4 Hz. For PN, the final three cutoff frequencies were instead: 10 Hz LP; 1.6 Hz LP; and 0.2 Hz HP. It has been noted in Section II-B that the precise filter frequencies are not critical but are based on heuristics. For PN, an additional 3000 Hz LP filter was included at the beginning of the chain which had one input for each channel and performed weighted summation. Gain was introduced at each filter stage. In the first one or two stages for IO and PN respectively, sufficient gain was introduced to bring the signal to 500 mV rms. Then gain was 4 and 3 for the two low-pass stages (these values were selected to keep the signal utilizing the available voltage range). An active high-pass filter has only parasitic capacitance on its virtual ground and this node can therefore suffer from capacitively coupled clock noise in the programmable interconnect. Thus a passive high - pass filter was used for the final stage (the gain was therefore unity). These signal processing chains were applied in software to each digitized trace separately using IIR filters. Following the final stage, a threshold was applied, where an iterative search yielded the threshold which to the nearest 1 mV maximized bespoke quality measures. In the case of IO, the quality measure was based on the background frequency of US detections being as close as possible to 1 Hz (a level which empirically works well over diverse recordings). In the case of PN, the quality measure rewarded CS detections which started in a time window up to 100 ms after the CS onset and lasted for the correct duration, and punished deviations from this ideal. (Details of the quality measures and further insights on these methods will be published separately). For the PN, which has multiple recording points, the quality measure was used to provide weights for the summation of the channels in the first filter stage, such that channels which individually provided better information about the stimulus contributed more.

The model was parametrized with: a threshold for the production of a CR when PU activation reached a proportion of 0.2 of its full baseline value (an arbitrary choice); a rate of (linear) reduction of PU activation such that it passed from maximum to minimum in 1 s; a delay from the CR onset to inhibition of IO of 80 ms (higher than the 20–30 ms observable in biology [17] in order to accentuate the observable effect in this experiment); and LTP and LTD rates which were set so that an acquisition of a well-timed response would ideally be achieved after 60 paired CS-US trials (a physiologically realistic number of trials would be $\approx$500 for rat but corresponding to rabbit and much fewer in humans) and extinguished after the same number (fluctuations in detection performance would cause deviations from these ideals, however).

The model was simulated based on the detections from the previous stage, to confirm that acquisition and extinction of the learning of a well-timed learned response was possible in principle based on applying these methods to the available data. To do so, the traces recorded in the 60-trial experiment were repeated twice, allowing there to be a phase of acquisition in which the weight value should decrease, followed by a phase of stability in which the weight value should be maintained in the same region by the negative feedback (in a control systems sense) effected by the (feedforward) inhibition from DN to IO. Thereafter, the traces were

repeated twice more but with the IO recording shifted forward in time by ITI/2, such that the increase in US-related events did not occur during the CS thus simulating unpaired trials; this allowed another 120 trials in which conditions for extinction were simulated and the weight value should increase to its maximum value and stay close to it thereafter.

Of recordings from the six electrophysiology sessions, some had S/N ratios from one or both of the nuclei too low for the described learning to recognizably occur (this will be quantified in a separate publication); the best simultaneous recording from both nuclei was selected for the experiment reported here. Having established that the learning was possible in principle, the same inputs were sent to the chip, yielding the results in Section V.

C. Chip Test Environment

The chip was placed on a bespoke PCB providing connections to DACs and ADCs and an integration board (XEM3010, Opal Kelly, USA) hosting an FPGA (Xilinx Spartan 3). The FPGA was used to programme the chip, manage the ADCs and DACs and stream data between the chip and a PC. The chip was designed to be packaged with a minimal pin-out of 56 pins in an 8×8 mm QFP package for implantation; however for testing it has a full pin-out of 144 pins. Of these, 58 are general purpose I/O ports to the FPMA core. Bespoke software for programming of the chip (placing, routing and calibrating) and monitoring of its operation was developed using Matlab (Math-works, USA). Programming SRAM, for example, involves generation of data words encoding the switch matrix settings generated by a routing algorithm. These words are transmitted via USB to the FPGA, which then affects a serial programming protocol. Programming each of the 337 rows took 2 ms.

D. Chip Programming: Place, Route and Calibrate

Various types of sub-circuit were defined, e.g., an active low-pass filter type, with rectifying as a sub-type, as demonstrated in Section III-F. The event detection chain and cerebellar model were decomposed into sub-circuits and described using a be-spoke description in Matlab code. Other sub-circuits included a delay (for example for timing the delay between CR onset and IO inhibition), a linear ramp (for example for describing the behavior of PU activation following a CS onset), a hysteretic threshold, etc. The delay and linear ramp are two examples of circuits which are event triggered and activate a PGN to drive their process only when required, so as not to waste power on unused clock cycles. Placement of components to form the necessary sub-circuits was performed deterministically based on heuristics from the user; in constructing filters, for example, trade-offs between clock rates and capacitance ratios were calculated from coded heuristics, as well as their relative placement to minimize necessary routing. Routing was then performed using a bespoke algorithm and the chip was programmed. The design used in this experiment employed 43% of CLBs, 89% of CSCs, 21% of AMPs, 38% of PGNs, and 39% of routing wires.

Each stage of processing introduces deviations from ideal performance due to mismatch, for example in amplifier offsets. To compensate for this, calibration routines were devised for each sub-circuit. For example, for active first-order filters, calibration consisted of streaming in a short section of recorded data, recording the filter output, comparing the output to that of the same filter in software and adjusting capacitor ratios and voltage biases to adjust gain and offset respectively. When initially laid out, an excess of programmable capacitance was made available beyond what was needed in the ideal case, to allow the capacitance ratios to be altered to allow for the effects of mismatch and parasitic capacitance from routing wires and switches. The calibration process was iterated until the residual error fell below thresholds chosen by the user, in this case < 50 mV offset and $< 5\%$ difference in gain. A calibration routine could also be devised for cutoff frequency but this has not been implemented. Pulse generator frequency, the basis of filter cutoff frequency and other behaviors, was however calibrated on a component-by-component basis.

To avoid accumulation of offset differences from one stage to the next, input was always to the first filter in the chain and comparison was always with the accumulated effect of all the software filters up to that point in the chain. In case a desired gain could not be programmed because the required capacitance were greater than that allowed for in the placement of CSC components, then the extra gain would automatically be introduced by the calibration in the following stage, correcting the overall behavior of the signal processing chain (the gain of the final HPF is, however, uncorrectably less than unity due to parasitic capacitance on the output node forming a capacitive divider to ground, but this simply results in altered thresholds for detection).

The Inc- Dec circuit was constructed from CLBs and was clocked by the outputs of two PGNs, which were enabled only during plasticity events. A binary -weighted design was used for the DAC, with CSCs emulating resistances. The CSCs were all clocked at the same low rate (since weight changes only slowly) and a calibration phase fine -tuned the capacitance values to maximize the linearity of the conversion given mismatch.

This is a case where accuracy can be traded off against resources; the more CSCs used, the better the linearity that can be achieved, see the discussion on accuracy in Section VI-B1.

IV. **RESULTS**

A. Real-Time Learning

As described above in Section IV-B, data from 240 trials (4 repetitions of 60 trials with the first 120 having paired CS-US events and the rest effectively having CS alone) was streamed to the chip, once it had been programmed to perform event detection and the cerebellar model, and had been calibrated accordingly. Fig. 5 shows the results of selected trials from the experiment. Note the diversity of the signals involved: *CS-detected*, *US-detected*, *LTP-clk-enable* , and *LTD-clk-enable* are all low-starved digital outputs from CLBs, with biases ranging from 0–500 nA; *CR* is the output of an AMP thresholding *PU-activation* biased at 30 nA (smooth upwards slews can be seen); and *PU-activation* and *Weight* are analog traces, with *Weight* being the output of a DAC sub-circuit buffered by an AMP, and *PU -activation* being the output of a linear ramp sub-circuit (driven by a CSC). In an early trial, (a), CS and US were both detected, leading to a period of LTP which lasted for the duration of the detected CS, and LTD was applied for a fixed period after the detection of the US. The net effect on the weight was negative, though almost imperceptible in the graph. Note that a detection of IO activity prior to the CS did not cause LTD. During the detected CS, *PU -activation* gradually declined from its baseline level, although not enough to cause a CR. In (b), after *Weight*, and thus the baseline for *PU-activation*, had de-creased somewhat, *PU-activation* crossed its threshold causing an output in *CR* around time 0.45, too late to anticipate the aversive stimulus. In (c), with the weight slightly lower, the *CR* event occurred prior to the air-puff, and in (d) the CR happened early enough that the US detection did not lead to LTD, because its action was blocked by the modelled effect of DN to IO inhibition.

Fig. 6 shows overall results for the experiment. Fig. 6(a) shows trial-by - trial detection performance for the two nuclei superimposed, as well as the CR events produced. Most CS events were detected shortly after their onset, and in addition there was a low rate of false alarms. Those correctly detected stayed active for an average of 0.46 s. US onsets were detected during the air-puff with a frequency $\approx$3 times the background rate. The noise inherent to the system is evidenced by the fact that the pattern of detections of CS and US events was similar but not identical from one block of 60 trials to the next, although the inputs were identical. Nevertheless the modelled neural system achieved the acquisition and extinction of a well-timed response to the CS; Fig. 6(b) zooms graph (a) in the region of the acquisition of a well-timed response; the first well-timed CR (excluding one produced due to a false detection at trial 59) occurred at trial 69, and from then until trial 120, 88% of CS events caused a well -timed response. There was a period until trial $\approx$70 in which it descended, after which it remained buffeted around the same level. Then from trial $\approx$120 onwards the weight ascended until it reached its maximum level, to which it thereafter stayed close. For comparison with Fig. 6(d) and (e) shows the evolution of the weight variable during the software simulation of the experiment. Although differences are visible, the broad behavior is the same.

B. Adapted Model

A demonstration of the utility of the programmable system is provided by an alternative experiment. Electrical stimulation of the FN to elicit an eye-blink can introduce large artifacts into the recordings from PN and IO which, unless cancellation techniques were applied, would result in detections of CS and US events for the duration of stimulation, corrupting the action of the model. To avoid this without developing artifact cancellation, an alternative form of the model was implemented on the chip, as in [4], in which there was no delayed inhibition of LTD based on the production of a CR, but rather, both forms of plasticity, LTP and LTD, were inhibited for the duration of a CR (this departs from the biomimetic roots of the model for the sake of practicality). In this case, when CRs are well-timed, US events will be blocked by this mechanism and the weight should stabilize in any case.

C. Power Consumption

Power consumption is presented as measurements of cur-rent (at room temperature; with Keithley 6487 picoammeter, Keithley Instruments Inc., USA) into outer *vdd* (thus dropping 3.3 V through the outer power rails) or into inner *vdd* (drop-ping 2.9 V through the inner power rails). With the chip powered but all components disabled, 0.9 nA passed through inner *vdd*; this has $\approx$8000 entry points to the matrix and the components through back-biased transistors, implying $\approx$110 fA per transistor and demonstrating very low leakage. Outer *vdd* cur-rent through the FPMA should be comparably low but interference from other cores on the prototype precludes accurate measurement. The bias generators of Section III-B leak 2 $\mu\Lambda$ internally, in order to generate currents which may be orders of magnitude lower, and since the core uses 60 of these elements which cannot be enabled separately in this prototype, the quiescent current would be no less than 120 $\mu\Lambda$. In fact, due to further

biases in other cores and to additional buffering, the cur-rent through outer *vdd* when they were switched on was 420 $\mu\Lambda$. Thus biasing overheads are unnecessarily high. During the main experiment (Section V-A), current increased by 94 $\mu\Lambda$ (outer and inner *vdd* contributed similar currents to this total and are hereafter combined for simplicity) . This was dominated by 26 amplifiers in constructed filters and the DAC circuit, which were biased at full strength. 6.4 $\mu\Lambda$ of this was due to switched capacitor operation i.e., to the state machines within CSCs which create non - overlapping clocks, the PGNs (of which 16 were used), and the routing capacitance leading from these to the CSCs; therefore, switched capacitor machinery had a significant but not dominant power cost. Most of the current consumption was due to the fastest processes, i.e., the rectifying filters and the initial summation of inputs from PN electrodes, which operated at $\approx$50 kHz. A separate experiment was performed in which only the PU activation part of the model was implemented (i.e., third trace in Fig. 5). The bias currents used are stated in Section III-C; this caused $<$ 20 nA total additional cur-rent during operation.

V. CONCLUSION

An FPMA specialized for neural signal processing and neural modeling has been designed and fabricated as a core on a chip prototype intended for use in an implantable closed-loop prosthetic system aimed at rehabilitation of a function internal to the brain. Novelties in the design of the FPMA include: the intimate mixing of SC analog techniques with current-starved digital computation and power saving innovations within this frame-work; and the adaptation of components for use within a switch-leakage-resistant framework employing inner- and outer -power rails. The utility of the system has been demonstrated by the implementation of classical conditioning of an eye-blink reflex, resulting in the acquisition of well-timed responses to paired conditioned and unconditioned stimuli, which have been detected in real -time from multichannel data recorded simultaneously from two sub-cerebellar nuclei, and the extinction of those responses given unpaired trials constructed from the same data.

ACKNOWLEDGEMENT

This paper is just the survey of the paper "**A VLSI field-programmable mixed signal array to perform neural signal processing and neural modelling in a prosthetic system**" by **Simeon A. Bamford, Roni Hogri, Andrea Giovannucci, Aryeh H. Taub, Ivan Herreros, Paul F.M.J. Verschure, Matti Mintz, Paolo Del Giudice**. Hence, we would like to thank all the authors of that paper for giving us a platform for carrying out survey on this extremely good topic.

REFERENCES

[1] W. House, "Cochlear implants," Ann. Otol. Rhinol. Laryngol., vol. 85, pp.1–93, 1976.
[2] R. Kumar, A. Lozano, Y. Kim, W. Hutchison, E. Sime, E. Halket, and A. Lang, "Double-blind evaluation of subthalamic nucleus deep brain stimulation in advanced Parkinson's disease," Neurol., vol. 51, pp.850–855, 1998.
[3] D. Taylor, S. Tillery, and A. Schwartz, "Direct cortical control of 3D neuroprosthetic devices," Science, vol. 296, pp. 1829–1832, 2002.
[4] R. Prueckl, A. Taub, R. Hogri, A. Magal, I. Herreros, S. Bamford, R. O. Almog, Y. Shacham, P. Verschure, M. Mintz, J. Scharinger, A. Silmon, and C. Guger, "Behavioral rehabilitation of the eye closure reflex in senescent rats using a real-time biosignal acquisition system," in Proc. Int. Conf. IEEE Eng. Med. Biol. Soc., 2011, pp. 4211–4214.
[5] T. Berger, R. Hampson, D. Song, A. Goonawardena, V. Marmarelis, and S. Deadwyler, "A cortical neural prosthesis for restoring and en-hancing memory," J. Neural Eng., vol. 8, 2011.
[6] ReNaChip project [Online]. Available: http://www.renachip.org/
[7] D. Marr, "A theory of cerebellar cortex," J. Physiol., vol. 202, pp. 437–470, 1969.
[8] D. Woodruff-Pak, M. Papka, and R. Ivry, "Cerebellar involvement in eyeblink classical conditioning in humans," Neuropsychol., vol. 10, pp. 443–458, 1996.
[9] V. Bracha, M. Webster, N. Winters, K. Irwin, and J. Bloedel, "Effects of muscimol inactivation of the cerebellar interposed-dentate nuclear complex on the performance of the nictitating membrane response in the rabbit," Exp. Brain Res., vol. 79, pp. 453–468, 1994.
[10] A. Taub and M. Mintz, "Amygdala conditioning modulates sensory input to the cerebellum," Neurobiol. Learn. Mem., vol. 94, pp. 521–529, 2010.

