ABSTRACT

Title of Document: ANALOG VLSI CIRCUITS FOR BIOSENSORS, NEURAL SIGNAL PROCESSING AND PROSTHETICS.

Alfred M. Haas
Doctor of Philosophy, 2009

Directed By: Professor Martin Peckerar,
Department of Electrical and Computer Engineering

Stroke, spinal cord injury and neurodegenerative diseases such as ALS and Parkinson's debilitate their victims by suffocating, cleaving communication between, and/or poisoning entire populations of geographically correlated neurons. Although the damage associated with such injury or disease is typically irreversible, recent advances in implantable neural prosthetic devices offer hope for the restoration of lost sensory, cognitive and motor functions by remapping those functions onto healthy cortical regions. The research presented in this thesis is directed toward developing enabling technology for totally implantable neural prosthetics that could one day restore lost sensory, cognitive and motor function to the victims of debilitating neural injury or disease.

There are three principal components to this work. First, novel integrated biosensors have been designed and implemented to transduce weak extra-cellular

electrical potentials and optical signals from cells cultured directly on the surface of the sensor chips, as well as to manipulate cells on the surface of these chips. Second, a method of detecting and identifying stereotyped neural signals, or action potentials, has been mapped into silicon circuits which operate at very low power levels suitable for implantation. Third, as one step towards the development of cognitive neural implants, a learning silicon synapse has been implemented and a neural network application demonstrated.

The original contributions of this dissertation include:

- A contact image sensor that adapts to background light intensity and can asynchronously detect statistically significant optical events in real-time;
- Programmable electrode arrays for enhanced electrophysiological recording, for directing cellular growth, for site-specific *in situ* bio-functionalization, and for analyte and particulate collection;
- Ultra-low power, programmable floating gate template matching circuits for the detection and classification of neural action potentials;
- A two transistor synapse that exhibits spike timing dependent plasticity and can implement adaptive pattern classification and silicon learning.

ANALOG VLSI CIRCUITS FOR BIOSENSORS, NEURAL SIGNAL PROCESSING AND PROSTHETICS

By

Alfred M. Haas

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Martin Peckerar, Chair
Joel Cohen
Nicholas DeClaris
Neil Goldsman
Robert Newcomb

Dedication

I dedicate this thesis to my family, to whom I owe everything.

Acknowledgements

To everyone who has supported me, I thank you all.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Overview

Stroke, spinal cord injury and neurodegenerative diseases such as ALS and Parkinson's debilitate their victims by suffocating [1], cleaving communication between [2], and/or poisoning [3], entire populations of geographically correlated neurons. Although the damage associated with such injury or disease is typically irreversible, recent advances in implantable neural prosthetic devices offer hope for the restoration of lost sensory, cognitive and motor functions by remapping those functions onto healthy cortical regions [4]. These prosthetics are remarkable devices, yet for most of these state of the art systems, neural event detection and classification systems remain external; and most implants still consume too much power and occupy too much space to reliably resolve neural events across multiple channels simultaneously [5], [6]. To truly restore lost sensory, cognitive and motor function to victims of debilitating neural injury or disease, a totally implantable neural prosthetic is required [7]. The principal aim of my research, therefore, has been to develop enabling technology for such prosthetics and the rehabilitation of lost sensory, cognitive and motor function.

In order to meet this ambitious aim, I have taken a multi-tiered approach. In one aspect of this research, I studied existing technology and developed novel integrated biosensors for electrophysiological recording and optical sensing of neural activity. In another, I designed, fabricated and characterized ultra-low-power circuits for detecting, extracting the salient features from, and identifying the source of neural action potentials. Finally, in an attempt to build technology for closed-loop

unsupervised control, I implemented a new silicon synapse capable of correlating signals for Hebbian learning. In developing these circuits and systems, I have addressed some of the key limitations of previous generations of biosensor and implantable signal processing architectures – power consumption and reliable encoding of neural data – and I offer some novel approaches for next generation implantable prosthetic devices.

1.2 Research Contributions

This thesis is divided into four technical Chapters, 2-5, which detail the original contributions of this work and provide the context to appreciate the significance of these contributions to the field.

Chapter 2, entitled Contact Imaging, introduces integrated active pixel sensors ("APS") for imaging biological activity, such as neural action potentials. In this chapter, we introduce and lay the theoretical foundations for a deeper understanding of the advantages and limitations of integrated contact imaging [8], and show the simulated and experimentally verified performance of fabricated contact image sensors [8], [9], [10], [11]. The original contributions of this thesis described in this chapter include the design, fabrication and characterization of a contact image sensor that adapts to background light intensity and can asynchronously detect statistically significant optical events in real-time. The architecture of this adaptive sensor ("a-APS") is presented along with experimental data [11], and we disclose an array of a-APS sensors with arbitrated AER readout for asynchronous (and unsupervised) neural spike detection that has also been designed, simulated and submitted for fabrication.

Chapter 3, entitled Neural Recording, details the historical evolution of neural recording techniques and demonstrates the experimental performance of several state-of-the-art integrated microelectrode arrays designed to interface with, manipulate and record from living biological cells and systems [12], [13]. The original contributions of this thesis described in this chapter include the design, fabrication and characterization of programmable electrode arrays for enhanced electrophysiological recording, for directing cellular growth, for site-specific *in situ* bio-functionalization, and for analyte and particulate collection. Fabricated programmable arrays for compensating electrode and amplifier mismatch, process variation and local environmental inhomogeneities have been characterized [13]. When properly packaged, these arrays are suitable for *in vitro* neural recording and also for integration with implantable recording and signal processing devices. Moreover, a variant of these arrays can be used to pattern arbitrary potentials across the sensor surface for directing the growth of developing and possibly damaged nerve cells.

Chapter 4, entitled Spike Sorting, is divided into three parts. First, we provide a concise primer on the relevant aspects of mixed-signal VLSI design. Second, we perform an in-depth review of state-of-the-art spike sorting algorithms and circuits. Finally, we detail the architecture and performance of the ultra-low-power floating gate template matching circuits I designed for the detection and classification of neural action potentials [14]. The original contributions of this thesis described in this chapter include: (a) the overall floating gate template matching architecture; (b) application of floating gate adaptation and template matching to solve the detection and classification problem under competing constraints of low-power dissipation and

high computational precision; (c) novel on-chip variance estimation circuitry; (d) novel asynchronous current-mode weight-update circuits; (e) unique silicon neuron template generation mechanism; and (f) a VLSI implementation of a theoretical non-linear energy operator ("NEO") to threshold incoming signals for unsupervised template generation. Together, these contributions represent a significant step down the path toward next generation neural prosthetics.

Finally, in Chapter 5, entitled The Two Transistor Synapse, we consider biological Hebbian learning, explore a new analog circuit architecture for implementing biologically realistic learning (the two transistor synapse, "2TS"), and conclude with an illustrative pattern recognition application. The original contributions of this thesis described in this chapter include the development and implementation of a novel two transistor synapse that exhibits spike timing dependent plasticity and can implement adaptive pattern classification and silicon learning. This silicon synapse [15], provides the foundation for unsupervised silicon learning which could one day be used for the closed-loop control of implantable neural prosthetics.

Chapter 2: Contact Imaging

In this Chapter, we introduce, lay the theoretical foundations for and demonstrate the experimental performance of integrated contact image sensors for biosensing applications. Original contributions of this thesis to the field include the design, fabrication and characterization of a contact image sensor that adapts to background light intensity and can asynchronously detect statistically significant optical events in real-time.

2.1 Integrated Image Sensors

Since the late 16th century we have been using microscopes as a window into a world that remains hidden to the naked eye [16]. Powerful optical lenses and sub-micrometer precision stages provide an unparalleled view of the molecules, structures and biological organisms that populate that world – microscopes enable us to elucidate surface chemistries [17], to characterize deadly pathogens [18], and to visualize the very mechanisms of conscious thought [19]. However, for all the marvels that modern microscopes can conjure, ownership and operation of these instruments is costly in more ways than one. Conventional "light" microscopes are heavy and take up large volumes of space, e.g. [20]. While this is an overhead that is fine for research, it is a price that first responders in the field and prosthetics engineers cannot tolerate. In the modern era, diagnostic, therapeutic and rehabilitative applications that have historically been anchored to the instruments must be untethered and allowed to follow the need.

Necessity breeds innovation, and forty years after Smith, Boyle and Thompsett first shaped silicon sands into a charge-coupled array for storing patterns of light and dark projected onto the focal plane [21], today's engineers have developed the technology to integrate tens of millions of tiny photosensors onto a silicon chip the size of a fingernail, e.g. [22]. As a result, state-of-the art telecommunications devices can transduce full frame images into a bitstream of millions of 1's and 0's in a fraction of a second [23]. Removing the lenses from such systems and coupling microscopic particles or biological cells directly to the integrated sensor surface, it is possible to leverage the myriad advantages of integrated circuit technology – low cost and power consumption, high speed, and advanced focal plane processing – with the superior efficiency and reduced footprint of imaging systems that do not require the overhead of intervening optics. What follows is an introduction to integrated contact image sensors, and the theoretical and experimentally established performance of such sensors.

2.1.1 Active Pixel Sensors

Although there are many different means of transducing light into electrical signals, we shall focus on the active pixel sensor ("APS"), which is broadly defined as an integrated photodetector and amplifier. Arrays of APS's fabricated in today's standard complementary metal oxide semiconductor ("CMOS") technology can be read out orders of magnitude faster and more efficiently than the charge-coupled devices ("CCD") first fabricated in the early 1970's, compare e.g. the sensors reported in [24] with the technology in [21]. CMOS offers the additional advantages of reliable, low cost manufacturing and the ability to integrate photosensors with

advanced image plane processing on the same chip. The prototypical APS constitutes a reverse biased, or "pinned", photodiode and source-follower readout transistor. A schematic of a typical 3-transistor ("3T") CMOS APS is shown in Fig. 2.1 below.

In this particular configuration, PMOS transistor **M1** is operated as a switch that resets the voltage at the gate of the source-follower, **M2**, to **Vdd** when the **rst** signal is asserted. Incident light (photon flux) generates charge carriers (electron-hole pairs) in the depletion region of the reverse-biased photodiode, and the carriers that do not recombine internally are swept across the photodiode junctions, resulting in a photocurrent that tends to discharge the photodiode junction capacitance to ground. Transistor **M2** is a source-follower single transistor amplifier whose output voltage is proportional to the voltage at its gate, which is equivalently the voltage across the photodiode junction capacitance. **M3** is a simple switch that permits the voltage at the source of **M2** to be read out on a common bus.

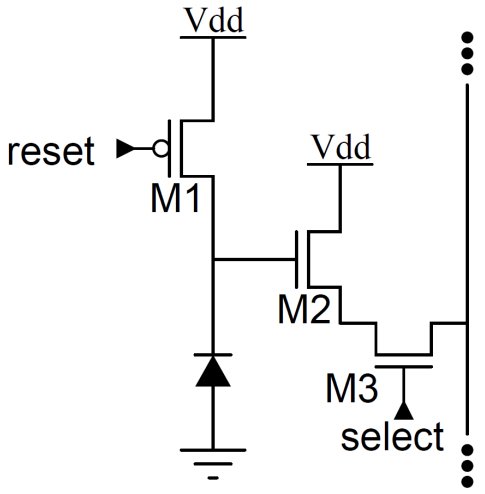


Figure 2.1: Schematic of typical 3-transistor ("3T") active pixel sensor ("APS").

Under ideal conditions, assuming a perfectly linear photodiode junction capacitance and uniform illumination, we would expect both the photocurrent, I, and the photodiode junction capacitance, C, to remain relatively constant, so that according to the characteristic equation: $I = C\frac{dV}{dt}$, we would observe the change in voltage over time to be roughly linear. This holds true to first order approximation

even though the photodiode junction capacitance is not truly linear. Instead it is a function of photodiode junction width, which is proportional to the square root of the difference between the built-in potential and the voltage across the diode, as is shown in Equations 2.1 & 2.2, below:

$$C_{diode\_junction} \approx \varepsilon_{silicon} \cdot A_{diode} \Big/ X_{diode\_depletion} \tag{2.1}$$

$$X_{diode\_depletion} \approx \sqrt{\frac{2\varepsilon}{q}\left(\frac{N_A + N_D}{N_A N_D}\right)(V_{reverse\_bias} - \Phi_0)} \tag{2.2}$$

In these equations, $C_{diode\_junction}$ represents the photodiode junction capacitance, $\boldsymbol{\varepsilon_{si}}_{licon}$ is the permittivity of silicon, A_{diode} is the exposed area of the photodiode, $X_{diode\_depletion}$ is the diode depletion region depth, q is the charge on a single electron, 1.6 x 10^{-19} A, N_A and N_D respectively represent the acceptor and donor concentrations of the p- and n- type silicon regions of our photodiodes, $V_{reverse\_bias}$ is the external voltage drop across the photodiode, and Φ_0 represents the built-in-potential. The voltage drop across the photodiode, in turn, depends on both the reset bias voltage and illumination conditions which give rise to photocurrents that tend to discharge the photodiode junction capacitance. However, for nearly all illumination conditions, dV is so small during the integration period (the time between reset and readout) that the dependency may be relegated to a second-order effect.

Figure 2.2 shows a schematic of a photodiode in a standard, single-well CMOS process. The p-type substrate ("p-sub") is silicon that has been doped with a

group-III acceptor such as Boron, while the n-well has been doped with an electron rich group-V donor such as Arsenic or Phosphorous. As with a typical diode junction, in the absence of any applied electric field across the diode, electrons from the donor atoms on the n-type side will tend to migrate across the junction to fill the vacancies in the group-III acceptors on p-type side. This results in unmasked positive atomic cores on the n-type side and filled valence shells on the p-type side that form the depletion region.

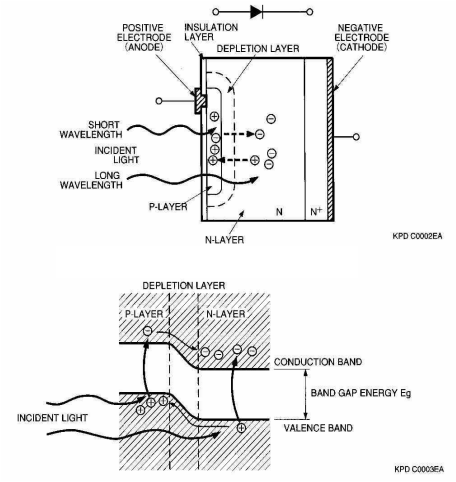


Figure 2.2: Photodiode physical cross-section (top) and energy band diagram (bottom). [25]

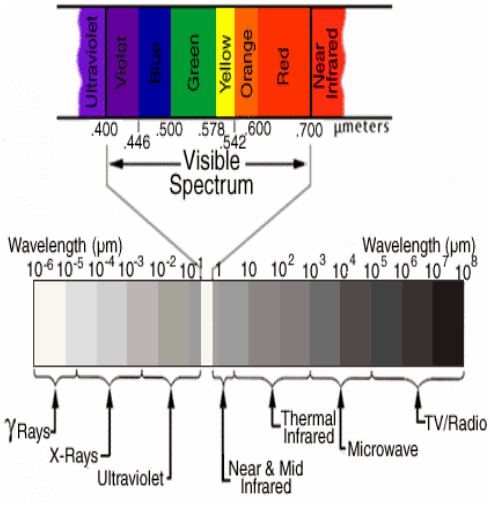


Figure 2.3: Electromagnetic spectrum. [26]

Free charge carriers do not reside in the depletion region, as the built-in potential reflected in the potential energy diagram shown in Fig. 2.2, sweeps them away. The depletion region as bounded by the charged atomic cores acts as a parasitic capacitor in parallel with the photodiode. It is this parasitic capacitance which is charged and discharged by the photocurrent, and thus gives rise to a proportional voltage change across the photodiode. Under reverse bias, the depletion

region is extended and free carriers tend to get swept across the junction if they do not first recombine with each other.

Photocurrents occur when incident photons impart sufficient energy to the doped silicon to generate electron-hole pairs – i.e., when photons impart sufficient energy to valence electrons to jump the 1.12eV bandgap to the conduction band. Since the wavelength and energy, or frequency, of electromagnetic radiation are inversely proportional ($c=\lambda\upsilon$), UV, visible and near-IR all possess such energy, see Fig. 2.3, but electromagnetic radiation with a wavelength greater than about 1100 nm, such as microwaves and radio frequencies, do not and therefore cannot generate photocurrents in Si. The reason for this is described simply by Planck's law: $E = hf = \frac{hc}{\lambda} = \frac{1240_{nm}}{\lambda_{nm}} eV$, where h is Planck's constant, 6.626 x 10^{-34} kgm^2/s, c is the speed of light, 3 x 10^8 m/s, and 1 eV is 1.6 x 10^{-19} J. Given the 1.12 eV indirect bandgap of silicon, individual photons of light with $\lambda > (1240/1.12)$, or approximately 1100 nm (IR), do not possess the requisite energy to, by themselves, excite electrons to the conduction band. If the incident light intensity is high – e.g. with a laser excitation source, it is possible for two higher wavelength, lower energy photons to impact the same atom at the same time to generate electron-hole pairs, and a corresponding photocurrent, but absent high intensity collimated laser light, this occurrence is a rarity.

However, using longer wavelength, lower frequency laser radiation, scientists can probe deeper into tissues with less damage – 2-photon microscopy is a very useful technique [27]. Finally, with respect to the spectral responsivity of silicon – light with a wavelength below about 300 nm does not penetrate the photodiode as

deeply and tends to recombine before it reaches the depletion region to be turned into photocurrent; likewise some much higher frequency (lower wavelength) radiation, such as x-rays, can pass through silicon without being substantially detected at all.

The ratio of electrons collected to incident photons at a given wavelength is referred to as the detective quantum efficiency ("DQE")[1]. Theoretically DQE can approach unity, but for a standard CMOS APS fabricated in a commercial 0.5um process, owing to material imperfections that give rise to local recombination, thermal carrier generation that can become conflated with small photocurrents, and electrical noise, the observed DQE over the visible wavelengths is on the order of 0.33. The spectral responsivity ("SR") reflects the relative DQE of a photosensor at different wavelengths, normalized for optical power. A characteristic silicon photodiode SR curve is shown in Fig. 2.4, alongside measured data from a fabricated APS, shown in Fig. 2.5. Note that the shape of the experimentally observed data on the right, shown in blue, closely parallels the optical power curve, shown in red. Thus when normalized, the experimentally characterized APS SR is relatively flat.

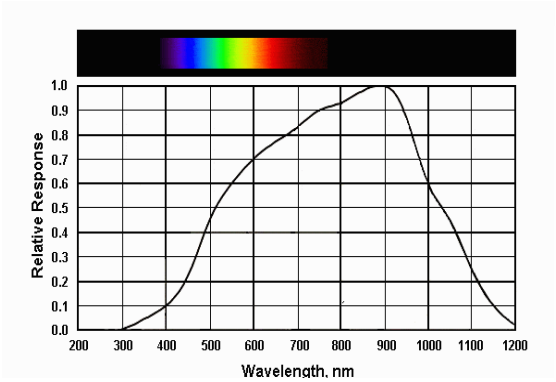


Figure 2.4: Typical silicon photodiode spectral responsivity. [28]

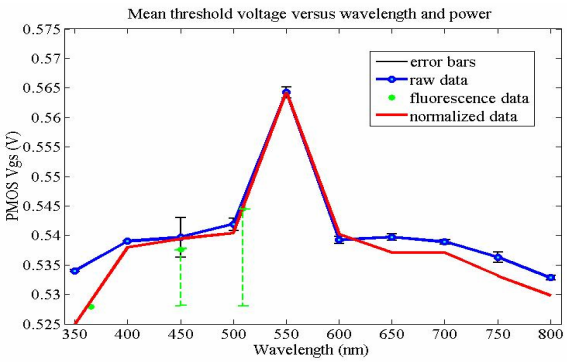


Figure 2.5: Measured spectral responsivity for the n-APS sensor. Blue curve represents raw data; bottom represents measured intensity. Each data point is the mean of 50 trials. [11]

[1] As contrasted with quantum efficiency ("QE") which we use to refer to the number of electron-hole pairs generated by a single incident photon.

The disparity between the characteristic and experimentally observed data is in part a function of the sensitivity of the fabricated photodiode with respect to the dark current noise floor – note that the minimum mean threshold voltage recorded is approximately 527 mV (for the filtered fluorescent light), separated from the highest peak by less than 10% or 38 mV; it is in part proportional to the accuracy and precision of the readout and data acquisition hardware – measured mean threshold voltages are proportional to the photocurrent, but do not represent a direct current measurement – instead they are the bias voltages at which, over a fixed integration time, the generated bias current overwhelmed the photocurrent (more on this later) and invoked a digital trigger – the imprecision in the readout circuitry is further compounded by the data acquisition system which is capable of controlling the bias voltage with approximately 1-2 mV accuracy; finally, and to a lesser extent, the observed SR is in part due to the precision of the optical test apparatus, including such factors as ambient light leakage, and also the limited precision and reliability of the monochromator light source and optical power meter.

In sum, device physics and circuit design both play an important role in defining the capabilities of integrated CMOS. Physical properties of the photodiode set both the DQE and SR and define the noise floor – i.e., the minimum detectable signal beneath which it is not possible to differentiate between the photocurrent and the current resulting from thermally generated carriers. However, numerous circuit techniques have been reported to reduce dark current and to extend the dynamic range of silicon photodiodes, e.g. [29]; device physics alone does not determine the characteristics of the APS. As seen above, much of the difficulty in designing ultra-

sensitive APS image sensors is in crafting the readout circuitry that will preserve the dynamic range of the signal against circuit imprecision – due to mismatch and process variations – and against noise inherent in analog and mixed-signal designs. We will discuss APS design considerations in greater detail later in this chapter. First, in evaluating the fundamental limits of CMOS APS imagers, we consider the resolution limits of these image sensors independent of any intervening optics – coupling the object or cell to be detected directly to the sensor surface.

2.1.2 Contact Imaging

APS imagers are ubiquitous. They are embedded in every cell phone, PDA and laptop computer on the market, they have replaced CCD's in even high-end digital cameras, and they have supplanted film as the medium of choice for most medical and diagnostic applications. However, these imagers typically employ discrete optical lenses to focus images onto the sensor surface, which adds to the cost, size and complexity of these instruments and makes them unsuitable for applications such as implantable neural imaging [30]. In order to establish the fundamental performance limitations of APS technology, we have investigated and characterized contact imagers capable of visualizing microscopic objects simply by coupling them directly to the sensor surface. These sensors leverage the enhanced collection efficiency that proximity to the sensor provides, and are able to perform functional microscopy without the overhead of intervening optics. In this section, we describe the results of simulations and measurement of CMOS APS imaging response, as published in [8].

In assessing the quality of captured images, we used the contrast of an imaged object – defined in terms of the mean object intensity, m_{object}, the mean background intensity, $m_{background}$, and the background variance, σ^2 – as $C = \frac{(m_{object} - m_{background})^2}{\sigma^2}$. Contrast represents the squared, or power, SNR of the imager and was both simulated and experimentally verified.

Simulations were performed using the commercial software simulator LightTools™, by fixing a 2 mm x 2mm Lambertian light source a uniform distance from a virtual opaque circular disk with diameter of 5 um and thickness of 1 um. The simulated image plane constituted one quarter of a 60 µm x 60 µm 2D array of 4900 square bins, each bin corresponding with a pixel of an image sensor. Fig. 2.6 illustrates the simulated system. Since the simulated image plane captured every photon incident upon it, the stochastic nature of the simulation is a function of the random photon generation of the Lambertian light source.

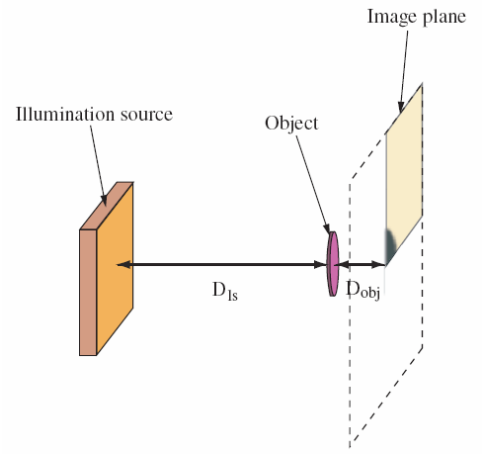


Figure 2.6: Schematic of theoretically modeled contact imaging system. [8]

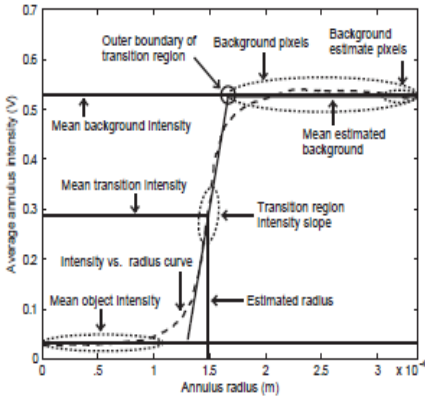


Figure 2.7: Plot of average annulus intensity as a function of radius, illustrating the computation of contrast parameters. [8]

In order to quantitatively assess contrast, we needed to specify the regions corresponding to the captured image of the object and the background. We defined the object as comprising all pixels of the simulated image plane within a specified radius of the fixed object center. To ensure that we could reliably segregate background from object, we defined a transition region between the two whose midpoint was determined iteratively by identifying the smallest radius at which the mean value of an annulus two pixels wide was greater than the calculated mean intensity of the transition region. We determined the outer boundary of the transition region by finding the intersection between the line approximating the intensity in the transition region and the estimated background intensity. All pixels outside the transition region are background pixels. The method is illustrated in Fig. 2.7.

Fig. 2.8 represents simulated data for the virtual quarter image plane as a function of increasing object distance from the surface. Notably, even at relatively large distances it is possible to resolve the object with the human eye, a feat that would prove exceedingly difficult for an integrated image processor. Fig. 2.9 illustrates the simulated data across all measured distances; the solid line is the LMS fit between the function $f(D_{obj}) = \frac{a}{(1+(\frac{D_{obj}}{d})^n)}$ and the logarithmically weighted contrast data. For the simulation data shown in Fig. 2.8, a = 1341, the characteristic distance, d = 0.1305 mm, and shape parameter, n = 3.076.

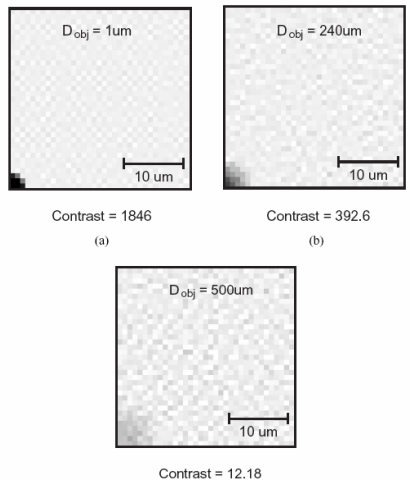


Figure 2.8: Simulated contact images of a quarter disk formed on image planes at: (a) 1 µm, (b) 240 µm, and (c) 500 µm away from the disk. [8]

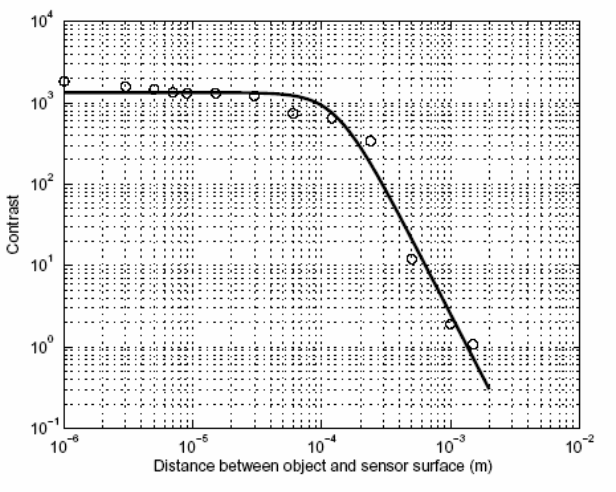


Figure 2.9: Simulated image contrast as a function of distance between the object disk and sensor surface. [8]

In order to measure imager contrast and confirm the theoretical results shown above, we performed two sets of experiments on a 96 x 96 APS, 8.4 µm pitch imager fabricated in a commercial 0.5 µm process.[2] The test setup I devised is shown in Fig. 2.10, alongside the fabricated contact imager.

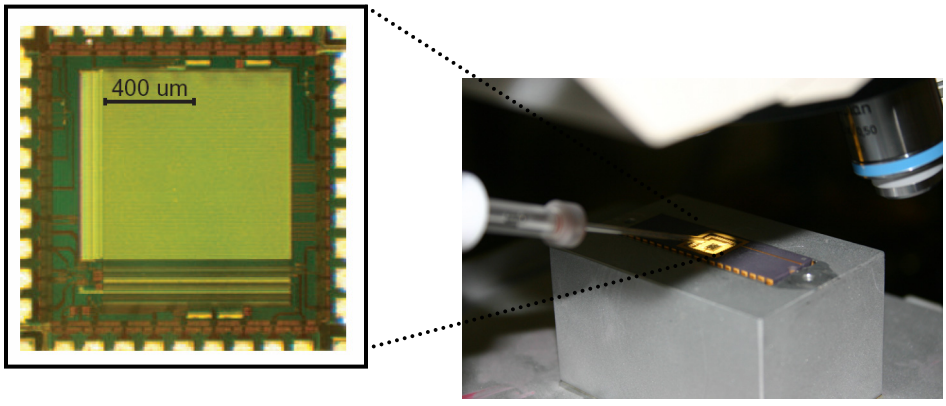


Figure 2.10: Photomicrograph of fabricated image sensor, [8], alongside a photograph of experimental contact imaging setup with micropipette, light source and chip shown.

[2] This imager array was designed and submitted for fabrication by Mr. Honghao Ji while he was a student at the University of Maryland Department of Electrical and Computer Engineering.

For the first experiment that I formulated, we represented the Lambertian surface between the light source and the focal plane using a 48 μm polystyrene microbead; for the second, we employed a 284.5 μm stainless steel microball. In both instances, the microparticles were attached to the tip of a pulled (using a Flaming Brown P-97 micropipette puller) borosilicate pipette with a clear UV-curable polymer, Loctite™ 3340. The bead-pipette fixture was then affixed to a Sutter MP-285 micropositioner (at an angle of 29 degrees from the horizontal) and the imager chip and board were then positioned onto a custom stage for an Axiotron microscope. Using the micropositioner, we first aligned the anchored bead to a position near the center of the imager and then raised and lowered the beads to take images at varying distances from the focal plane The sample and imager were illuminated using halogen light provided by the built-in light source of an Axiotron microscope, projected through an empty socket in the nosepiece which was approximately 45 mm above the sensor surface.

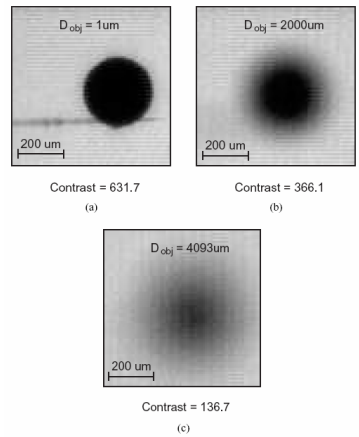


Figure 2.11: Measured images of a 284.5 um bead formed on image planes at (a) 1 μm, (b) 1950 μm, and (c) 3950 μm away from imager surface. [8]

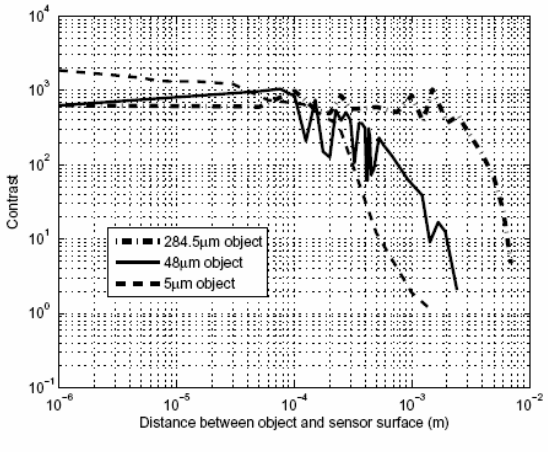


Figure 2.12: Image contrast as a function of increasing D from simulation (5μm) and experimental results (48 and 284.5 μm). [8]

The images were analyzed using the algorithm described above, except that the location of the object center was determined by inspection, and we: (a) attempted to reduce fixed pattern noise by subtracting a reference frame containing no bead from the captured data; and (b) cropped four pixels on each edge of the frame to eliminate edge effects. Once these tasks were completed, we computed the contrast ratios for each frame according to the formula provided above. Captured images for the stainless steel microball, as well as compiled data for all experiments and simulations are presented in Figs. 2.11 & 2.12.

For the polystyrene bead (not shown), the fit has amplitude, a=643.1, characteristic distance, d=0.2679 mm, and shape parameter, n=2.150; the metal microbead shown above has a=603.6, d=3.283, and n=5.8087. The characteristic distance at which the contrast begins to degrade increases as the size of the object increases. The higher noise content of the measured images, including the pipette tip and real-world illumination conditions, resulted in an expected but nonetheless significant increase in variance versus theoretical results, accounting in part for the observed discrepancy in contrast values between the three sets of data. Nonetheless, experimentally observed data qualitatively match the theoretical predictions, and proves the paradigm for assessing contact image quality.

2.2 Biosensing

Having explored the device physics, integrated circuit realization and performance of integrated contact image sensors, we next consider contact imagers as biosensors. In particular, we focus on two principal applications: (1) localizing cells and other microscopic particles coupled to the sensor surface for handheld cytometry,

cell sorting and other diagnostic and integrated feedback applications; and (2) detecting optical transients, or spikes, that correspond with, e.g., localized metabolic or neural activity, for identifying pathogens or visualizing cellular neural activity.

2.2.1 Sensing cells

An APS contact imager can detect microscopic particles or biological cells coupled to the sensor surface, provided that the pixel size is on the scale of the particles or cells to be visualized. Numerous integrated APS sensors for biological sensing have been reported, e.g. [8], [9], [11], [30]. A common problem is that cells, being semi-transparent, are difficult to detect without some means of enhancing contrast. Typically, this is accomplished by staining the cells with a dye such as neutral red, which can be introduced into live cells without significantly impacting their health. As an example that we presented in [8], taking a maximum safe dye concentration of 0.1 M, we can use the extinction coefficient for neutral red of 39000 cm-1 M-1 to compute the transmission rate through a 2 μm thick monolayer of stained cells as $T = 10^{-E_C \cdot conc \cdot l} \approx 0.17$ so that approximately 83% of incident light will be blocked by the cells. Therefore, to register the location of individual cells on the sensor surface, it is simply necessary to identify the dark, or occluded pixels – this forms the foundation of the spike detecting imager that we shall discuss in the next section. The dynamic range of a typical n-well, p-sub APS imager in a commercial 0.5 μm process, is about 54 dB (500:1), [8] allowing us to register 83% occlusion easily.

To demonstrate proof of principle cell localization, it is first necessary to insulate the exposed electrical areas of the packaged chip (bond wires and pads), and

also to keep the biological cells from coming into contact with any electrical connections or toxins. To accomplish both aims, we encapsulate the bondwires with a biocompatible UV-patternable polymer, Loctite 3108 in this case, and affix a custom media well using silicon glue, as is shown in Fig. 2.13 [8].

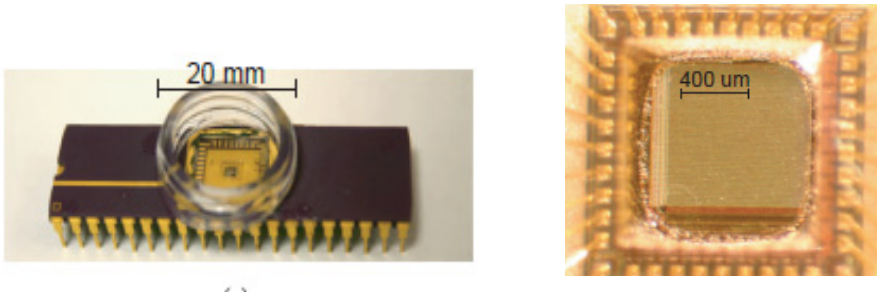


Figure 2.13: Photographs of (a) test fixture ready for cell plating, and (b) a close-up view of packaged contact imager. [8]

The surface of the chip is washed with deionized water and culture media to remove any harmful residues, and for the experiment whose results are shown in Fig. 2.14, bovine aortic smooth muscle cells ("BAOSMC") that had been stained with neutral red dye were plated onto the surface of the chip. A digital photograph of the sensor surface was taken through a microscope lens, and images were captured using the contact imager. Fig. 2.14 (a) is an enhanced photograph, Fig. 2.14 (b), is the image acquired by the APS sensor array, and Fig. 2.14 (c) is the overlay of (a) on top of (b). The stained cells are clearly seen in all three images [8], [9].

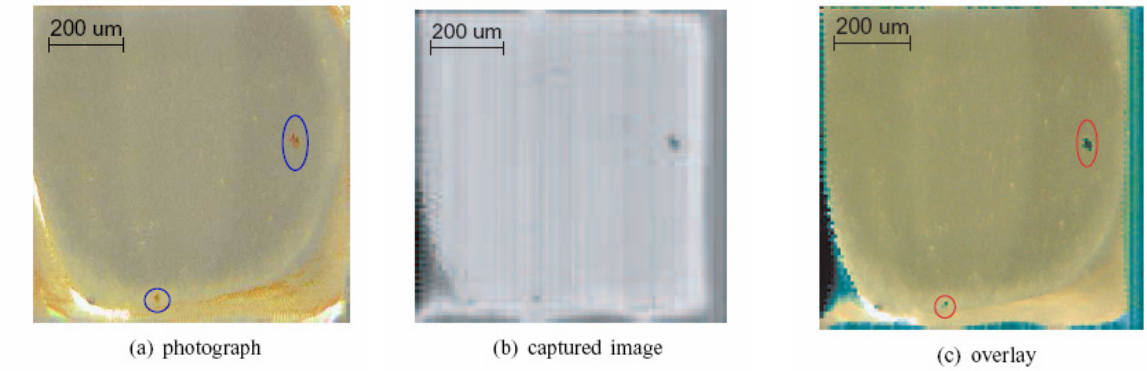
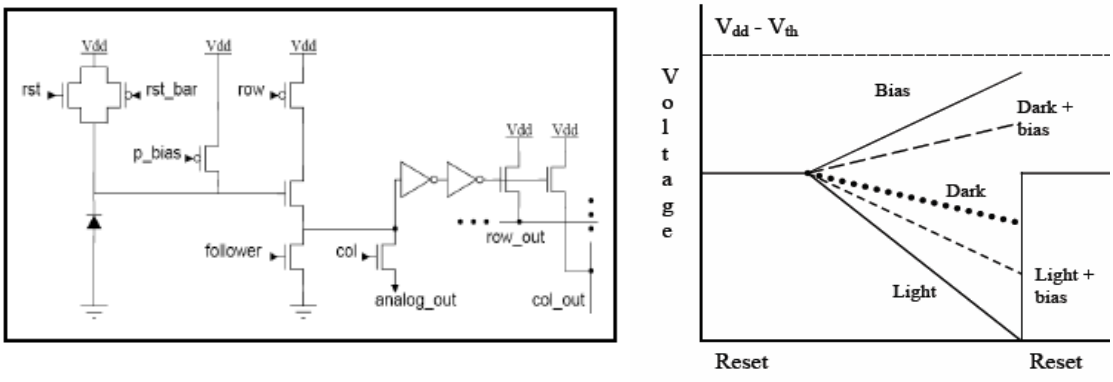


Figure 2.14: Pictures of live cells coupled to chip surface are taken using (a) a camera and (b) the contact imager. The overlapped view is shown in (c). [8],[9]

It is possible therefore, to image cells coupled directly to the sensor surface. However, in order to identify the location of the cells using a standard CMOS APS imager, an entire imaging frame must be read out first so that a user or computer program can segregate object(s) from background. To enable unsupervised detection of dynamic cellular behavior, it would be desirable for the imager itself to be able to identify and report the location of sparse distributions of cells without the necessity and computational cost of full frame readout. That is precisely the problem that the dark address event imager seeks to address.

Specifically, we designed a hybrid dark-active address-event representation ("AER") image sensor whose active pixel elements operated in both: (1) a conventional imaging mode; and (2) a "dark AER" mode wherein the individual pixels asynchronously generate voltage pulses, or spikes, when the incident light on a pixel falls below a user defined threshold [10]. Thus, in the dark AER mode, individual pixels sense whether they are occluded ("dark"), or not, by integrating the difference between the photocurrent and a tunable bias current onto the photodiode junction capacitance. Charge accumulates on an occluded photodiode because the photocurrent is smaller than the bias current. The capacitor voltage is then buffered by a source follower and, for a dark photodiode, increases until it crosses the threshold of a CMOS inverter, whereupon it is converted into a logic "high" event. For a full frame image sensor, "dark" digital events can be queued using a standard arbiter structure as outlined in [31], or may priority encoded in real-time and then multiplexed together for serial readout.

Fig. 2.15 (a) illustrates the fundamental sensing unit of the dark AER imager – it is a standard APS cell, with two essential modifications: (a) an in-pixel current source controlled by the voltage **p_bias**; and (b) the digital **inverter buffers** which convert the analog follower voltage to a digital signal. As laid out in a commercial 0.5 μm, 3-metal 2-poly process, the pixel is approximately 32 μm x 30 μm in size, and achieves a fill factor of 19%. The operation of the circuit can be seen in Fig. 2.15 (b), which represents the voltage across the photodiode junction capacitance as a function of time, under different bias conditions.



(a) (b)

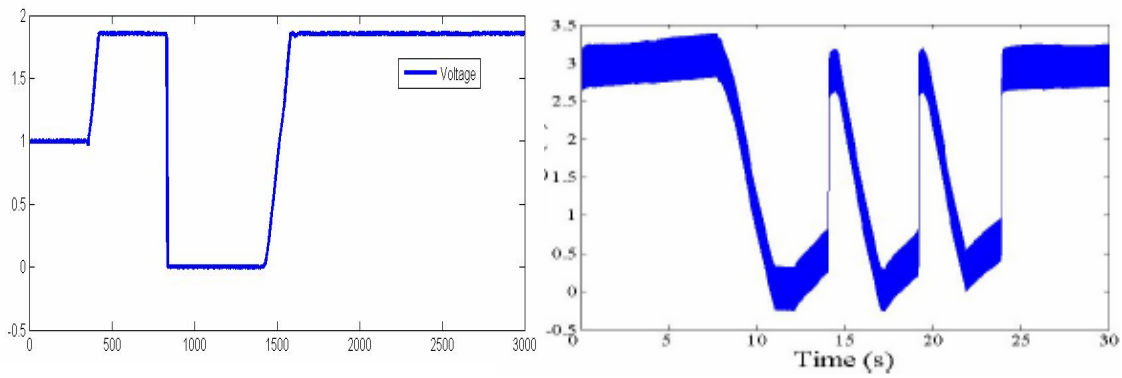
Figure 2.15: (a) Non-adaptive APS ("n-APS") schematic; (b) n-APS principle of operation. Incident light generates a photocurrent that discharges the photodiode junction capacitance, while an opposing, user-tuned, current source supplements the thermal (dark) current and charges the node. [10]

During the reset period, the voltage is fixed at a DC level. When the **rst** switch is turned off, then the incident light generates a photocurrent that tends to discharge the pixel. However, the opposing bias current tends to charge the photodiode, resulting in a competition between the two. In the dark, the bias current is stronger than the photocurrent and so the pixel will charge and the voltage rises.

By contrast, in bright light, the photocurrent is more powerful than the user-fixed bias current, and so it tends to discharge the pixel, thereby reducing the voltage across the photodiode. Note that owing to the characteristic relationship between voltage and current described above, the $\frac{dV}{dt}$ for a given illumination intensity is essentially linear. In asynchronous mode, when the voltage rises above or falls below a certain threshold, the competition between light and dark is concluded; the voltage triggers the inverter buffers which convert the analog output into a binary decision – logic high or logic low, and thereby signal an optical occlusion or spike event. Thus, instead of reading out an entire image frame and performing costly post-processing on the data, the image sensors themselves flag the location of migrating objects or cells that occlude light from the sensor surface. Furthermore, since the spatial location of each APS is fixed in an array, the address of each flag can be deduced and transmitted by integrated readout circuitry. For sparse distributions of cells, this address-event representation is an efficient and compact alternative to conventional techniques.

Several experiments were performed to validate the operation of the dark AER APS. Results from two of these experiments which illustrate analog and digital transitions between light and dark for a pixel of the fabricated image sensor are shown in Figs 2.16 (a) & 2.16 (b), respectively. Voltage peaks correspond with the dark, whereas troughs represent bright incident light. Note that the digital transitions are rail-to-rail, whereas the analog transitions do not rise above 2 V. Likewise, it is important to observe that while the analog transitions are crisp, the digital readout

suffers from considerable jitter, owing in part to parasitic capacitances and leakage through the analog readout architecture.



(a) (b)

Figure 2.16: (a) Non-adaptive APS ("n-APS") source follower output in response to slow changes in ambient light intensity (30s total time scale); (b) n-APS digital output in response to slow changes in ambient light intensity. Plateaus represent periods of static ambient light intensity, while high voltages represent dark or occluded signals, and low voltages represent the incident light. [10], [11]

In addition, owing to the very small photo- and bias- currents (a fraction of a pF), and the size of the photodiode junction and parasitic capacitances (100's of fF), the full-swing transition time is on the order of hundreds of ms to seconds. Their slow speed, while useful for detecting the natural migration of cells that occurs over minutes, hours and days, does not permit rapid physiological measurement of optical activity. Furthermore, the thresholds for these transitions are fixed, user defined and subject to circuit mismatch and process variation; although each pixel can distinguish light from dark, none can adapt to changes in ambient illumination. As a result, in order to: (a) enable reliable unsupervised optical event detection under different ambient lighting conditions; and (b) enhance the discrimination speed of the detector, we developed a new adaptive image sensor for optical spike detection.

2.2.2 Sensing biological activity

With sufficient spatial, optical and temporal resolution, it is possible to measure not only static information about a cell, such as the location to which it may have migrated, but additionally to observe some aspects of a cell's dynamic behavior and physiology as well. There are several reported integrated architectures that perform just such measurements [30]. Photodiodes fabricated on the scale of biological cells possess the requisite sensitivity[3], integrated fluorescence filters permit us to employ optical dyes and coated fluorescent microbeads to illuminate chemical potential spikes [32], and detector speed is enhanced so that sensing dynamic biological activity with integrated contact imagers is rapidly becoming the new paradigm.

As our first step down that path, we note that the ambitious objective of observing salient biological activity using a contact imager is predicated on the notion that we can visualize such activity – i.e. that physiology can be mapped to fluctuations in optical intensity that can be captured by the image sensor. Inasmuch as most biological processes do not exhibit optical signals naturally, this is no mean feat. The most common conventional manner in which to visualize such processes fluorescence microscopy – this process requires a microscope fitted at a minimum with both an excitation filter and source [33]. Several important biological compounds, principally NADH, will fluoresce when excited, and emit light at a

[3] Many promising new technologies with superior sensitivity are being developed – avalanche photodiodes in standard CMOS still suffer a high noise floor, but offer much promise in this area. Likewise, techniques for enhancing the signal quality, such as surface plasmon resonance and optical waveguides suggest that the future of this field is indeed bright.

different wavelength; other biological compounds, metabolic processes and potential shifts can trigger fluorescent dyes or probes.

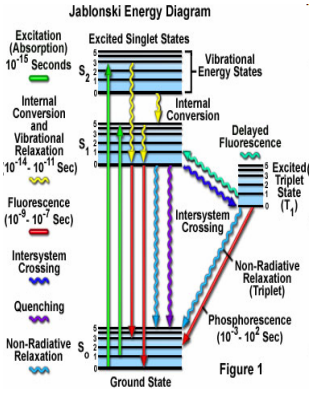


Figure 2.17: Jablonski diagram, [34].

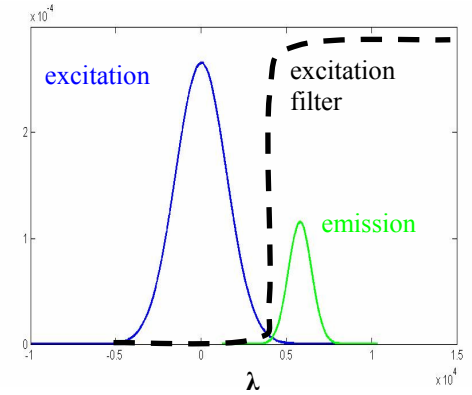


Figure 2.18: Stokes shift and principles of fluorescence detection.

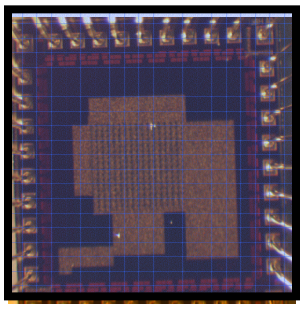


Figure 2.19: Photomicrograph of n-APS sensor array, shaded with a drawn blue filter.

Fig. 2.17 is a schematic Jablonski diagram illustrating the fundamental physics underlying fluorescence, and Fig. 2.18 displays an idealized representation of excitation and emission spectra, along with the concept behind an excitation filter. Although there are many commercial ways to implement these filters, none of these fall within the province of this dissertation. For the fluorescence experiments conducted here, the integrated fluorescence filters were prepared by adding a UV chromophore, benzotriazole, to poly-dimethyl-siloxane ("PDMS") in order to form a patternable absorption filter that absorbs over 99 percent of light below 400 nm, while passing nearly all of the light above that wavelength [35].[4] Fig. 2.19 displays a photomicrograph of the n-APS shaded with a drawn blue filter.

[4] The filters were prepared by Mr. Marc Dandin of the University of Maryland Department of Bioengineering.

To the extent that there exist several well-known reported techniques for monitoring biological activity by fluorescence imaging, we have not attempted to substantially modify or add to these methods. Although we address two such means for neural signal detection – the methods in and of themselves do not represent a contribution of the present work. Instead, we establish basic theoretical performance characteristics of a contact imaging system for the fluorescence detection of neural signals, and report a novel adaptive contact image sensor with enhanced detection speed suitable for measuring optical spikes that correspond with biological metabolic and/or neural activity.

We shall focus our theoretical discussion on two classes of neural signals: (a) action potential propagation along a nerve axon; and (b) synaptic transmission between neurons. For each of these classes of neural signals we shall consider the required: (1) spatial resolution; (2) sensitivity; and (3) speed to capture the signal. Spatial resolution depends principally on neuroanatomy – action potentials propagate along nerve axons that can be smaller than 1 μm in diameter, but are typically hundreds of microns to millimeters in length [36]. Thus, sensors with dimensions on the order of tens of microns are capable of resolving action potentials that traverse their photodiodes. Likewise, although synaptic terminals are often less than 1 μm in diameter, with collections of Ca^{++} channels and synaptic clefts on the order of 10-20 nm [37] – a sensor 50 μm on a side can still capture aggregate synaptic activity from a population of synapses. As with such dynamic imaging techniques as fMRI, quorum sensing of neural activity can still provide meaningful information about dynamic neural activity, and as integrated microlenses become available, it is possible

to leverage the next generation of optics together with the advantages standard CMOS brings.

Sensitivity is a function of both: (a) sensor dynamic range; and (b) the transduction chemistry – i.e., fluorophore excitation and emission spectra. As we have seen, the dark current floor and associated shot noise determine the minimum detectable signal, and dynamic range of an APS imager is also affected by circuit noise [38]. Although it is not possible to characterize the dark current directly for the a-APS, plain-vanilla APS imagers fabricated in a commercial 0.5 μm process using the same n-well, p-sub structure have a reported dynamic range of in excess of 53 dB [8]; techniques exist for further extending this range [39]. Under low-light conditions, it is often difficult to distinguish thermally generated carriers from photocurrents, but from a systems perspective, since we illuminate the objects of interest with an excitation source of our choosing, we can engineer appropriate ambient illumination. The real limitations are imposed by the selection of viable transduction chemistries – reported potentiometric dyes embedded in the nerve membrane register a shift of between 0.1% and 10% for a 100 mV depolarization [40], [32]; while a 10% shift is well within the capabilities of standard APS imagers, 0.1% would need to be amplified for reliable APS detection. This is equally true for monitoring *in vivo* synaptic activity – although there are reported APS detectors capable of monitoring free Ca^{++} concentrations *in vitro* in the nM range, only the coarsest detection of optical events using an implantable contact imager has been demonstrated *in vivo* [30]. However, synaptopHlourin ("spH") is a pH sensitive GFP which exhibits a 20-fold increase in fluorescence when deprotonated in the neutral

extracellular fluid of the synaptic cleft versus its quiescent state inside the acidic environment of the pre-synaptic cell [41]; transfecting cell lines with this GFP derivative should render suitable signals for APS detection. In addition, there are other technologies for amplifying a weak biological signal, like surface plasmon resonance, and avalanche photodiodes – both of these pose unique advantages and challenges that go beyond the province of this dissertation. The critical point is that CMOS APS imagers are capable of resolving the signals generated by potentiometric dyes and proteins from background noise and so are suitable for monitoring dynamic biological electrophysiology.

With respect to speed, neural spike trains saturate at frequency of around 1-2 kHz, so that even high end systems that would seek to sample the transient components of an individual action potential require limited bandwidth – a conservative estimate of the required Nyquist sampling rate suggests that 10-20 kS/s would be sufficient to ensure perfect reconstruction. However, to capture 30 frames of a 128x128 array per second, while serially recording 10kS per pixel for each frame would demand a minimum clock speed of close to 5 GHz – pushing the outer limits of present day integrated technology and consuming far too much power for an implantable system. The solution for recording sparse neural activity is far simpler – employ a more modest (power and speed) imager to identify pockets of neural activity, and then focus 20kS/s attention on those active regions of the frame. In this capacity, a standard APS imager operating asynchronously performs the required operations beautifully. It is to that end that we have designed an adaptive image sensor for optical spike detection.

As a first step down the path towards integrated contact imaging of neural activity, I developed a novel APS with adaptive in-pixel thresholding for optical spike detection. The adaptive threshold is set by ultra-low-power current-mode CMOS circuits which continuously compute the mean and standard deviation of the photocurrents generated by eight representative pixels in real-time. Sensor pixels discriminate between light and dark by integrating onto the photodiode junction capacitance the difference between the photocurrent and an opposing bias current whose magnitude is set by the mean and standard deviation circuits. I have characterized the active pixel sensor with and without an integrated fluorescence filter for biosensing applications and measured results agree with theory and simulations.

Fig. 2.20 reflects the basic principle of on which the adaptation is premised – the magnitude of photocurrents generated across an array of APS under uniform ambient illumination should approach a normal distribution, so that it should be possible to distinguish transient optical events from background light levels by applying a simple statistical threshold, in this case, some user-set number of standard deviations above the mean.

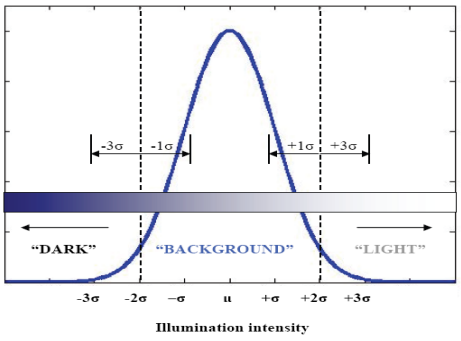


Figure 2.20: Principle of adaptive thresholding. [11]

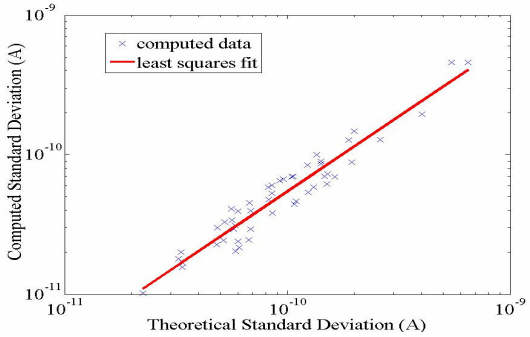


Figure 2.21: Computed standard deviation from measured variance estimation circuit data. [11]

Custom circuits compute the mean and standard deviation of N (= 8 for this design) representative photodiodes in real-time. Fig 2.21 illustrates the computed standard deviation from measured variance estimation circuit data.

As shown in Fig. 2.22, each pixel contains a current source whose magnitude is set by the mean and standard deviation circuits. As with the n-APS, this current source opposes the photocurrent, but in this case it charges in ambient light that is not focused on the center pixel – i.e. it represents the background light. A spatially localized transient optical event of sufficient power will overwhelm the current source and drop the photodiode junction capacitance to ground. When the transient optical event has concluded, the feedback transistors [42] ensure a rapid reset to logic high. Fig. 2.23 shows simulation data.

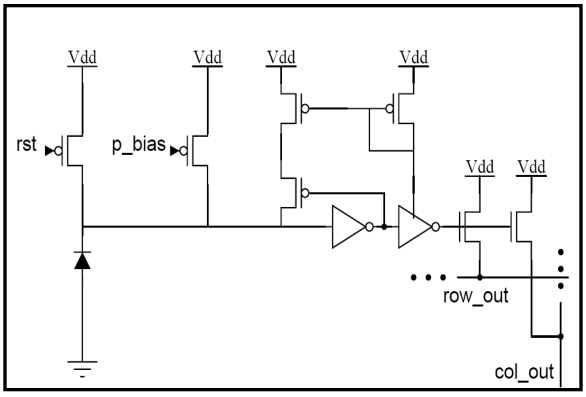


Figure 2.22: a-APS sensor. [11]

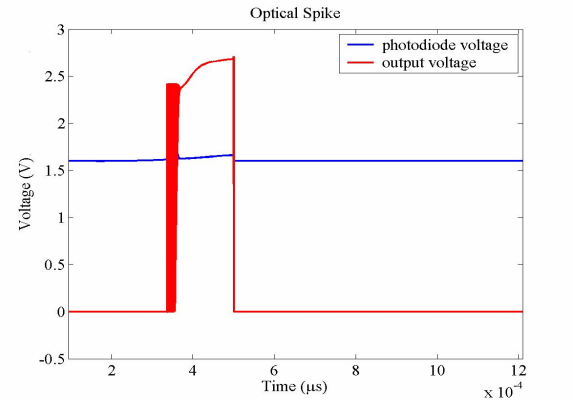


Figure 2.23. Simulated single a-APS spike. [11]

The layout for a pair of these image sensors, with surround pixels and mean and standard deviation computation circuitry is shown below in Fig 2.24, alongside a photomicrograph of the fabricated image sensor in Fig. 2.25. I have previously shown the spectral responsivity for the n-APS, and demonstrated that measurements

using a weak excitation source (nW) in the sub-400nm cut-off regime register no higher than dark current on the fluorescence-filter-coated sensor.

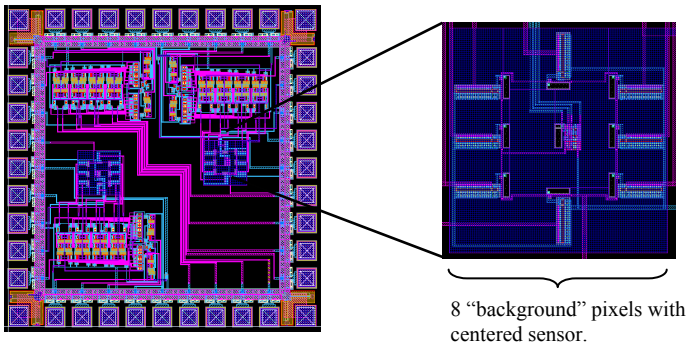


Figure 2.24: Layout of two a-APS sensors and one standard deviation circuit, [12].

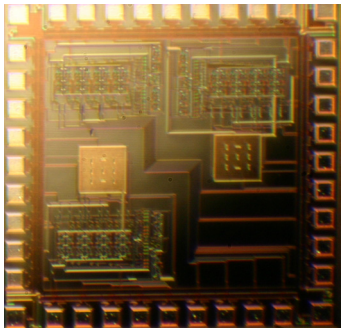
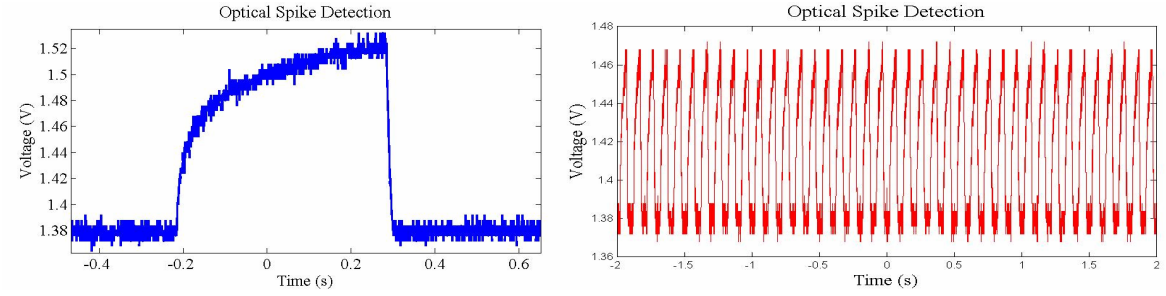


Figure 2.25: Photomicrograph of fabricated image sensor, [12].

To illustrate proof-of-principle optical transient detection, I performed 2 additional sets of experiments. For the first, whose results are presented in Fig. 2.26 and 2.27 below, a blue LED light (~465 nm) was pulsed at 1 Hz (left, blue) and 10 Hz (right, red) onto the image sensor. For these tests, all pixels – the local sensor, as well as the 8 representative pixels used to set the threshold – were illuminated simultaneously with the pulsed LED source. As a result, the real-time adaptation mechanism sets a bias current that overwhelms the local pixel sensor current and charges the photodiode in the light; this powerful adaptation mechanism also maintains a higher voltage floor in the dark.



(a) (b)

Figure 2:26: Measured data from the a-APS when uniformly illuminated with a pulsed blue LED. (a) single spike measured at 1 Hz; (b) spike train measured at 10 Hz, [11].

In order to further characterize the operation of the sensor, I considered two different methods of coupling light directly to the center pixel and keeping the remaining surround relatively dark: (a) fiber optic cable; and (b) focused laser beam. For ease of setup, I chose to work with a focused visible laser at ~ 633 nm; this class II laser delivers approximately 1 mW of power in typical operation. The experimental setup is shown in Fig. 2.27, below. A chopper wheel was used to pulse the laser signal onto the pixel of interest while keeping the surround shrouded only in ambient room light. Fig 2.28 shows an approximately 1 kHz signal as measured by the a-APS; at the highest available chopper setting, ~ 4kHz (not shown), only about 10% peak attenuation was observed in the amplitude of the measured signal.

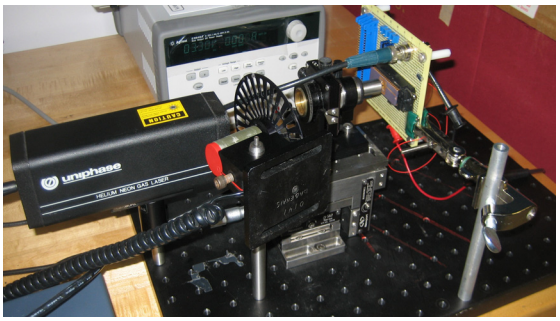


Figure 2:27: Laser experimental setup, including 633 nm class II laser, chopper wheel, prototype board and sensor.

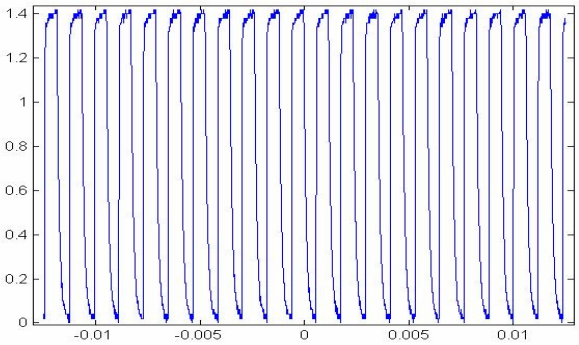


Figure 2:28: Measured data from the a-APS when red laser light (633 nm) is focused directly onto the sense pixel and chopped at 1 kHz.

Laser light exhibits a normal distribution. Fig 2.29 shows a MATLAB simulation of laser light intensity as a function of position in the upper right quadrant of the image sensor. In this simulation, the location of the center and surround pixels is identified with black boxes, the pseudo-color gradient reflects the relative intensity of the distribution, with red being normalized to unity and blue approaching zero. For this particular distribution, the focus is principally on the center pixel, but the

surround receives sufficient light to theoretically balance the center. A graphical illustration of simulated relative current intensities of center vs. surround as a function of the spread or focus of the laser light is provided in Fig 2.30. The intersection of the two approximately represents the distribution at which the center pixel exceeds the 3σ threshold; small variations in the spatial arrangement of the surround photodiodes shift this slightly.

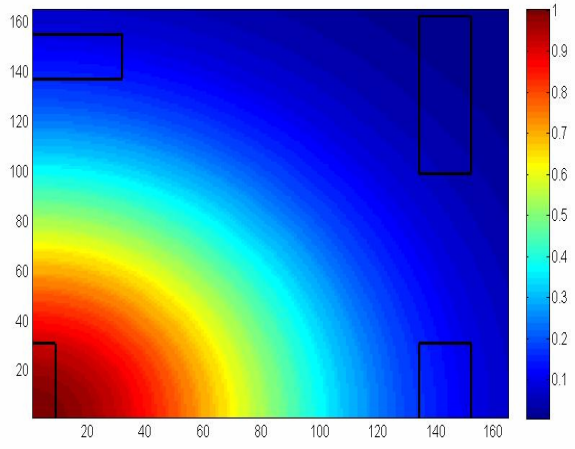


Figure 2:29: MATLAB pseudo-color plot of laser light intensity as a function of position in the upper right quadrant of the image plane. Black boxes mark the locations of the pixels, the pseudo-color gradient reflects the relative intensity, with red normalized to one and blue approaching zero.

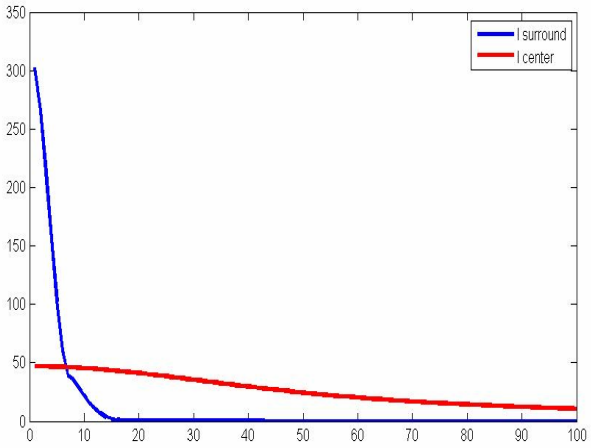


Figure 2:30: Simulated center versus surround current intensities as a function of the spread or focus of the laser light. Units on the x-axis represent signal spread; the y-axis is proportional to the photocurrent.

As a result of these figures, we can establish a basic framework for the experimental design of a system to detect optical transients. New blue voltage sensitive fluorescence dyes are less toxic than older red dyes, and when excited with laser light at a wavelength of approximately 633 nm exhibited ΔF/F of 10-13% [32]. Using such a potentiometric dye, it would be possible to couple the laser light in close proximity to the sensor surface using fiber optic cable, and with a suitable excitation filter, perform real-time adaptively thresholded neural imaging.

To realize this objective, I have designed and submitted for fabrication a complete imager incorporating this novel adaptive sensor. Full simulations and experimental data from the a-APS indicate that the array possesses the sensitivity, speed and spatial resolution to monitor neural signals.

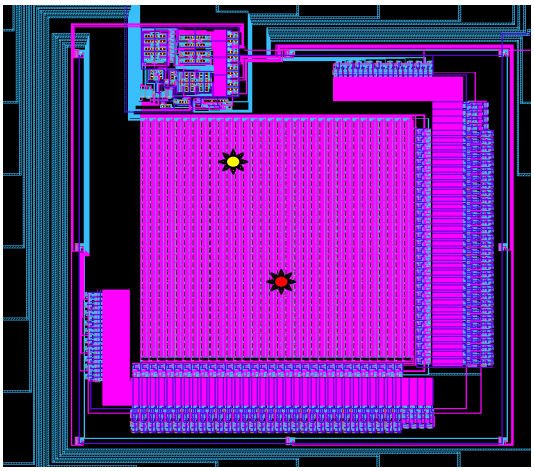


Figure 2:31: (above) Layout of 32 x 32 a-APS imager array with adaptive thresholding and arbitrated AER readout. Yellow and red suns indicate optical events.

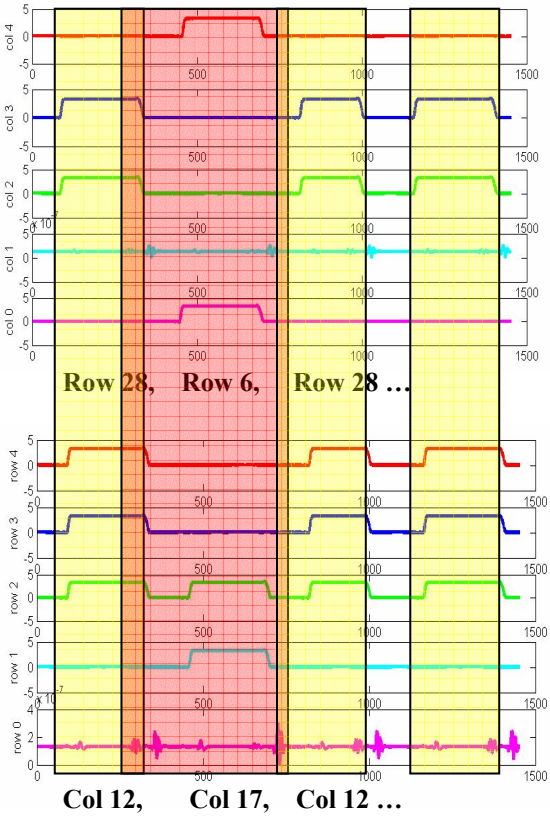


Figure 2.32: (right) Simulated AER readout from four optical events. Event timing is represented by the shaded color-coded columns.

In particular, the imager array comprises a 32 x 32 array of a-APS elements, with adaptive thresholding circuitry and mixed-signal arbitrated address-event-readout ("AER") circuitry. It represents a first step down the path to low-power, unsupervised optical spike detection for implantable neural prosthetics.

Chapter 3: Neural Recording

In this Chapter, we detail the historical evolution of neural recording techniques and demonstrate the experimental performance of several state-of-the-art integrated microelectrode arrays designed to interface with, manipulate and record from living biological cells and systems. Original contributions of this thesis to the field include the design, fabrication and characterization of programmable electrode arrays for enhanced electrophysiological recording, for directing cellular growth, for site-specific *in situ* bio-functionalization, and for analyte and particulate collection.

3.1 Neural Signals

Neurons are the fundamental building blocks of our perceptual and cognitive systems – they transmit and shape sensory information and collectively give rise to consciousness. There are on the order of 100 billion neurons in the average human brain – members of PhD committees tend to have a few more – and roughly 10^{15} synapses [37]. Incapacitate a few hundred thousand with a chemical inhibitor such as ethanol and the rest rally to compensate; damage millions of geographically correlated neurons, e.g. by stroke or spinal cord injury, and it is possible to irreversibly impair sensory, cognitive and motor function. As we collectively begin to take the first steps down the path to restoring lost function to the victims of neural injury and disease, we must first examine the behavior of individual neurons as they interact with one another. Over the years engineers have devised a host of different imaging techniques to meaningfully measure neural activity; MEG, fMRI, cat scans, EEG recordings, to name a few. All of these methods can resolve neural activity in

some detail and, remarkably each does so without piercing the skull. However, despite recent advancements in the field it remains impossible to measure individual neural electrical signals non-invasively. Instead, to obtain single unit data, it is necessary to record directly from, or in close proximity to the nerve cells of interest. Further, in order to conserve the signal strength of extracellular action or local field potentials, which are typically on the order of 50 – 500 μV, and to mitigate against signal degradation across long distances, recording electrodes should be connected as closely as possible with the hardware that will encode the incident neural events. Thus, implantable electrode arrays are preferred. Ideally such systems would be fully implantable, in order to obviate the need for hard-wired connections between microelectrode and prosthetic that are susceptible to both signal attenuation and infection.

In building the foundations for such implantable arrays, we shall first introduce some fundamental neurophysiology, including the mechanisms of neural action potential ("AP") generation and salient AP characteristics. Next we shall describe conventional mechanisms for *in vitro* and *in vivo* neural recording, including the patch clamp technique and conventional sharp electrode arrays. We report several new integrated electrode arrays for *in vitro* recording and for cellular manipulation. Finally, we lay the groundwork for extensions to *in vivo* and integrated implantable spike sorting systems. We begin with the neuron.

3.1.1 Neurophysiology

Roughly speaking, one can divide the neuron into four primary functional signal-processing compartments: (1) the **dendrites**; (2) the **cell body**; (3) the **axon**; and (4) the **synaptic terminals**, as shown in Fig. 3.1, below.

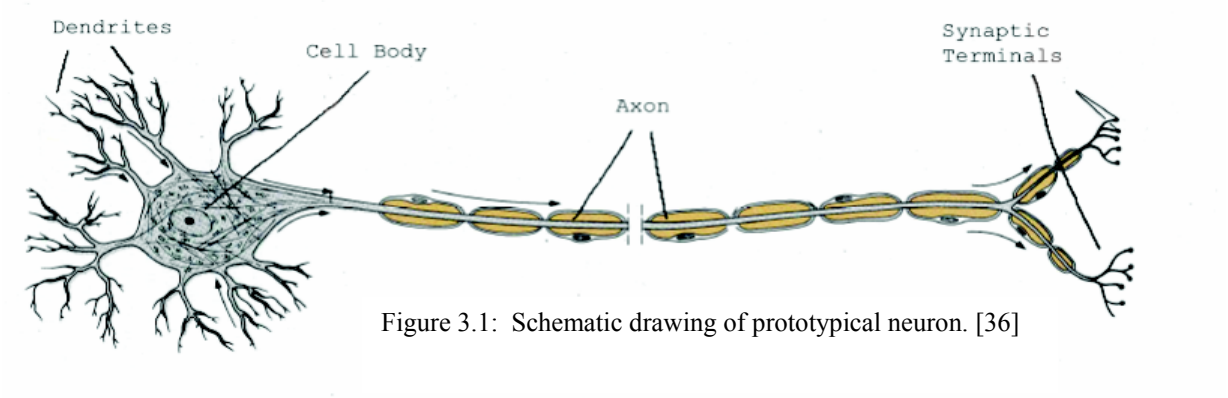


Figure 3.1: Schematic drawing of prototypical neuron. [36]

The (1) **dendrites** transduce electrochemical signals from the synapses of other nerve cells into graded post-synaptic potentials (**PSPs**) which travel passively down to (2) the **cell body**, which integrates the dendritic inputs and instantiates one or more action potentials at the axon hillock; (3) the **axon** acts as a transmission line, or cable along which action potentials are transmitted until they reach the synapses of the neuron; and (4) the **synaptic terminals** chemically propagate the signal to other cells. [36]. At a higher level of abstraction, the neuron may be considered as an information theoretic channel, with certain coding format – the stereotyped AP – and capacity [43].

Although it is often useful to regard biological action potentials as digital spikes, it is important to remember, especially for detection and classification purposes, that action potentials remain in fact analog signals. Fig. 3.2 shows the first

published recording of a neural action potential. Even though some information is lost in these early recordings due to slow response time and poor current detection limits, we nonetheless observe the characteristic analog peak and refractory period trough existing in all neural action potentials. These salient features always occur, but the precise shape and timing vary from neuron to neuron and in response to changing environmental conditions and stimuli [44]. In addition to AP shape, patterns of neural activity can range from quiescence to intermittent or tonic AP firing, to bursting activity in response to intense stimuli with an inter-spike-interval ("ISI") on the order of ms; the rate of firing is typically regarded as an encoding mechanism that researchers use to decipher these patterns [43].

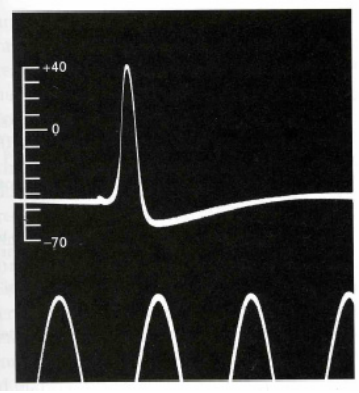


Figure 3.2: First published AP recording, 1939. [36]

However, before we attempt to divine meaning from the frequency at which a neural cell fires, we need to take a step back and examine the analog nature of APs, and in particular, the biological mechanisms by which they are generated. In the early 1950's, Hodgkin and Huxley performed a landmark series of experiments on the squid giant axon using a voltage clamp apparatus shown in Fig. 3.3 to elucidate those functions.

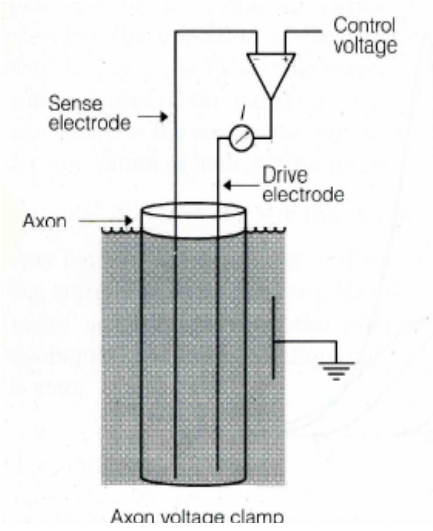


Figure 3.3: Voltage clamp apparatus drawing. [45]

The **voltage clamp** allowed Hodgkin and Huxley to fix the membrane potential using negative feedback: the sense electrode measured the potential across the squid axon relative to the bath ground and fed this potential into one input of an operational amplifier whose other input is a fixed control voltage; the output of the op amp drives a current injection electrode located inside the giant axon. This output injection compensates for any changes in sense input to bring the sensed and control voltages together, and in so doing provides a measure of membrane conductance at different fixed membrane potentials.

From the experimental data they obtained, Hodgkin and Huxley postulated the following explanation for how an action potential is generated:

> (1) At rest, the nerve cell membrane is selectively permeable to the ions that comprise the fluid media of the cerebral cortex and, through active and passive mechanisms, maintains an equilibrium potential of approximately -70mV with respect to the exterior of the cell; (2) stimuli fed by the cell's dendrites to the cell body exceeds a certain threshold voltage (approximately -50mV) which causes gated ion channels in the cell membrane to open and permits Na^+ ions to rush inward and K^+ ions to rush outward along their respective concentration gradients; (3) the net influx of positively charged ions (Na^+ ions are at a higher potential gradient than K^+ ions) depolarizes a localized portion of the axon membrane, and this depolarization causes neighboring sites along the membrane to also depolarize; (4) the previously depolarized portion has been flooded with positive ions, so that the membrane approaches the Na^+ ion equilibrium potential (about +50mV), which ultimately closes the Na^+ channels, while simultaneously K^+ ions continue to egress until local equilibrium is established again and the membrane repolarizes; (5) the signal is continually regenerated with almost no appreciable attenuation along the length of the axon, until it reaches the synaptic terminals, at which point it is propagated chemically to another cell. [36], [45]

Thus, the action potential is thus a traveling local depolarization that moves like a wave along the axon. The depolarization and subsequent repolarization cycle is so sharp and abrupt that it is often referred to as a "spike." The shape and frequency

of biological action potentials are both functions of the ionic conductances across the membrane, as is shown in Figs. 3.4 & 3.5.

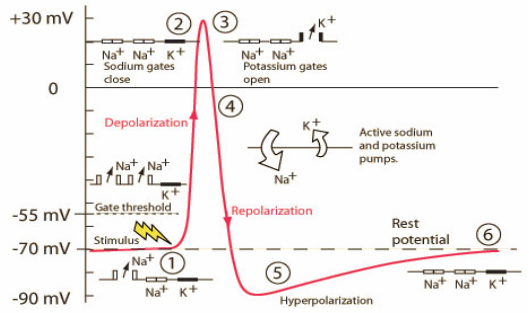


Fig. 3.4: Action potential as a function of ion channel activity. [46]

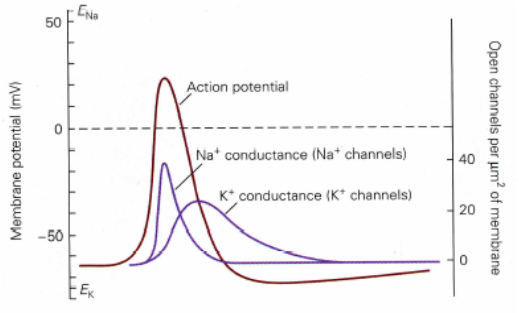


Fig. 3.5: Action potential as a function of Na^+ and K^+ channel conductances. [36]

As the figures indicate, resting potential for a typical cortical neuron while proportional to the ionic gradients of the local environment, tend to be around -70 mV; likewise, the stereotypical action potential has a peak amplitude on the order of 30-50 mV, along with a repolarization trough that dips slightly below resting potential. The entire cycle of a neural AP is typically on the order of 1 ms; this period can be extended by K^+ channel blockers such as tetraethylammonium ("TEA"), or shortened as a function of environmental conditions and intense stimuli [36].

3.1.2 Modeling Neural Action Potentials

Drawing on their experimental data, Hodgkin and Huxley proposed a simple circuit model to represent the nerve axon and its signaling function, as shown in Fig. 3.6. The model represents the cell membrane as a capacitor, and the gated ion channels as parallel conductances across the capacitor connected to batteries that represent ionic equilibrium potentials:

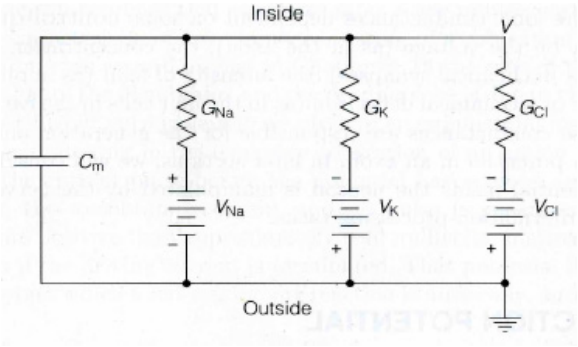


Fig. 3.6. Hodgkin-Huxley cell membrane model. [47].

Quantitatively we obtain: $I_M = C_m \frac{dV}{dt} + I_K + I_{Na} + I_L$; where, I_M = membrane current; $C_m \frac{dV}{dt}$ = capacitor current; I_K represents the potassium current; I_{Na} = sodium current; and I_L = passive leakage current.

The model is plain, and although subsequent authors have extended this model (Fitzhugh Nagamo, etc.), the precise algorithm used to compute the ionic conductances that give rise to the action potential are not of primary importance. What is critical is: (1) that the action potential shape, including spike amplitude, peak-width, trough duration and rise and fall slopes, are ***determined,*** at least to first order, by ***known*** functions; and, (2) as Mahowald [48] and others have demonstrated, these functions can be reproduced in silicon thus providing a rudimentary template for matching spike form to function. Figures 3.7 (a) and 3.7 (b) are, respectively, actual and simulated neural data reproduced from Mahowald's silicon neuron paper and a course paper containing original PSpice simulations.

The imperfections in the modeled neuronal behavior shown above, while interesting from a theoretical standpoint, are not significant from the circuit design perspective; if spike shape is all that matters[1], there are simpler and more precise ways of constructing an action potential template in silicon (*See*, *e.g.*, [49]).

[1] Of course, spike shape is not all that matters; spike frequency is of significance as well.

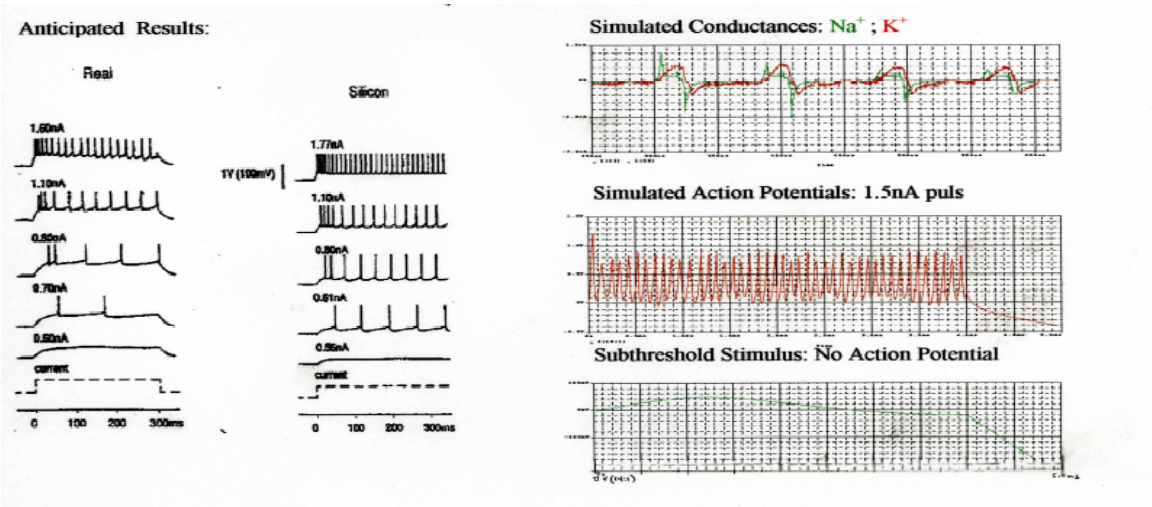


Figure 3.7. (a) Real action potentials versus silicon neuron APs [12]; (b) PSPICE simulations.

Subsequent authors have significantly improved on the prototypical conductance-based silicon neuron model, implementing circuits that generate tunable, biologically realistic action potentials in real-time or faster (*e.g.*, [50]). What matters is that the mapping can be done, and it can be done efficiently in silicon, thus giving rise to a tunable template generation mechanism that is considerably simpler than conventional means for accomplishing the same goal. In addition, having characterized the mechanisms and elucidated fundamental models for AP generation, we can now proceed to resolving the means by which these signals are measured and recorded.

3.1.3 Conventional Neural Recording

While there are several ways to record extracellular action potentials, the "gold standard" for measuring the channel currents that define these potentials is the patch-clamp method. Though there are many variants on this method (for which Neher and Sakman won the Nobel Prize), all of them entail suctioning the micron-sized tip of a hollow, fluid-filled glass micropipette onto a portion, or "patch", of the

cell membrane to form a seal and monitor channel currents or cellular potentials. In the case of whole cell patch-clamp recording, the seal between the micropipette tip and cell membrane joins the intracellular and microelectrode salt solutions and separates them both from the bath media, imposing a gigaohm membrane resistance between the recording electrode and the extracellular media. The voltage or current is "clamped" using negative feedback through an external amplifier, and the desired signal is measured. The gigaseal is the hallmark of whole cell patch-clamp recording, and is required for reliable monitoring of aggregate channel currents and potentials across the membrane. [51].

I assembled our patch-clamping apparatus from: (1) an Axiotron confocal microscope fitted with water immersion and long distance objectives corrected for biological work; (2) a custom-designed[2], rigid, locking integrated stage and mount for an MP-285 micromanipulator; (3) an MP-285 micromanipulator; (4) a Multiclamp 700B patch-clamp amplifier; and (5) a data acquisition system using a Measurement Computing PCI-DAS 1602 data acquisition card, and custom software developed in MATLAB and C++. Fig. 3.8 illustrates the experimental patch-clamp setup.

In developing my recording acumen, I learned the culture and care of several different cell lines and developed a multi-phase protocol for whole-cell patch clamping to govern: (a) the preparation of intracellular pipette filling solution; (b) the program parameters required to reproducibly generate optimal pipette tip resistances and geometries; (c) pipette filling; (d) electrode chloriding, assembly, placement and manipulation; (e) cell preparation; (f) patching onto a cell; and (g) neurophysiological recording. Following this extended protocol, I performed

[2] Jay Pyle of the IREAP machine shop fabricated the stage I designed.

repeated whole-cell patch-clamp recording on cultured bovine aortic smooth muscle cells ("BAOSMC").

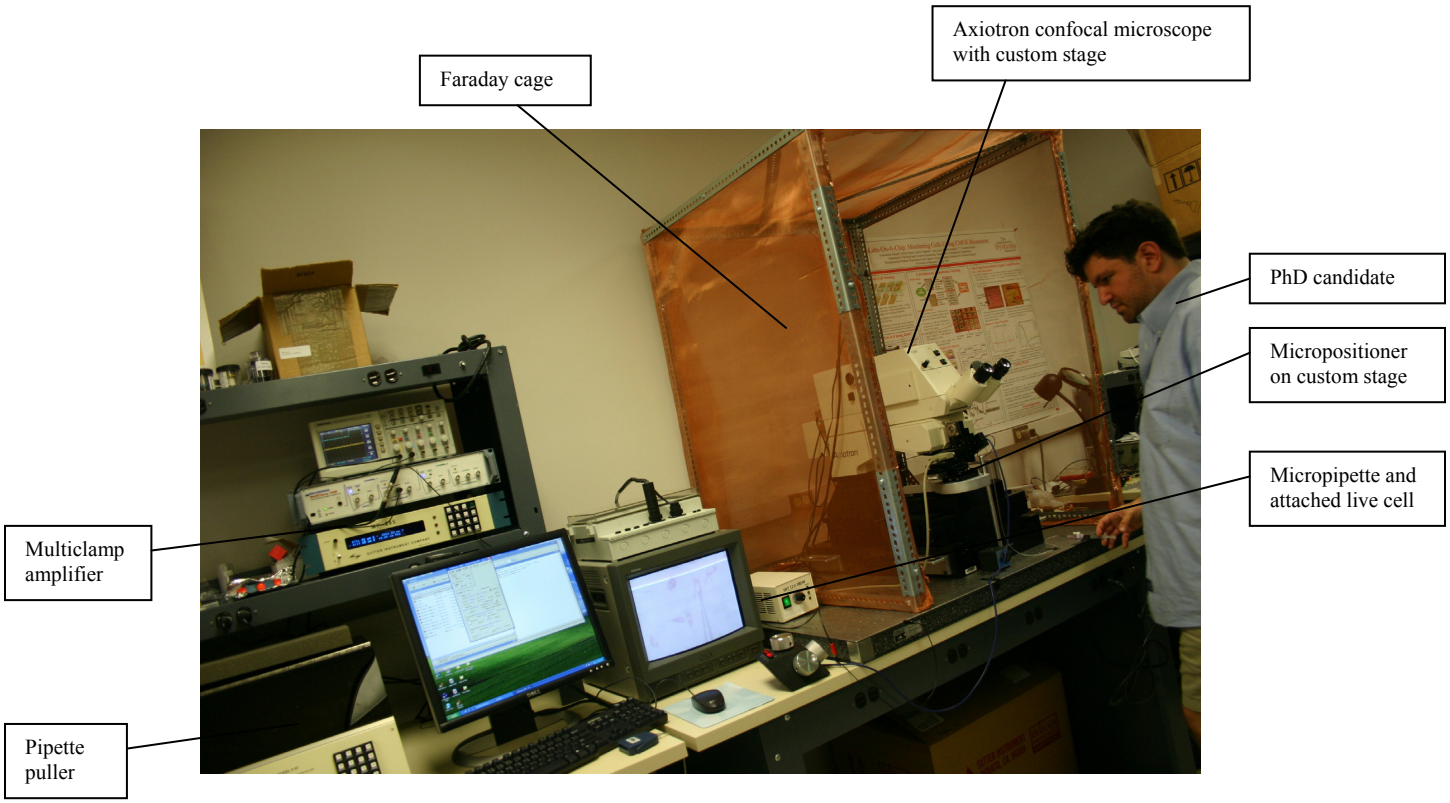
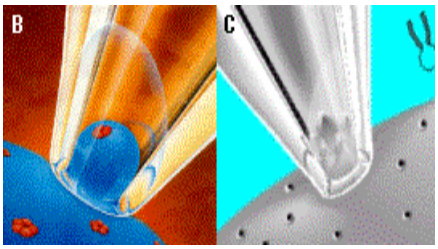


Figure 3.8: Author performing a whole-cell patch-clamp experiment using the custom-built rig.

Cartoon schematics of the whole-cell method are provided in Fig. 3.9, alongside a video capture of one of my own whole-cell patch-clamp experiments in progress, as seen in Fig. 3.10. The cartoons illustrate gigaseal formation by patching onto and subsequently rupturing the cell membrane, and show a schematic of the whole-cell electrophysiological configuration. Figs. 3.11 and 3.12 on the following page show recordings from BAOSMC experiments demonstrating successful application of whole-cell patch-clamping protocol resulting respectively in: (a) gigaseal formation; and (b) (noisy) signal recording.



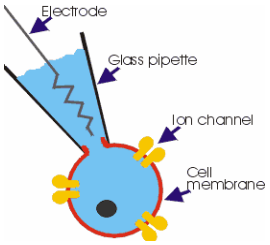


Figure 3.9: Cartoons of the whole-cell patch clamping procedure. [62][63]

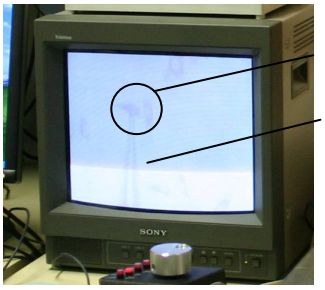


Figure 3.10: Video capture of whole-cell patch-clamp experiment.

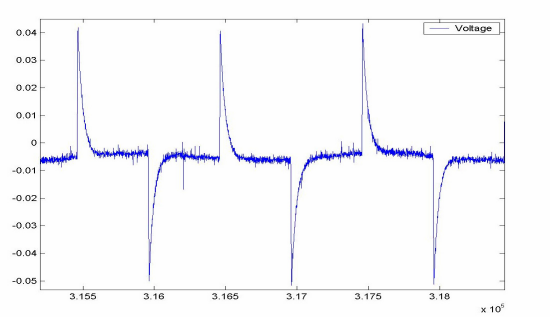


Figure 3.11: Seal test on BAOSM cell.

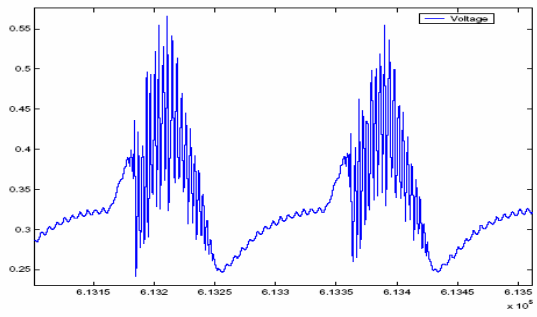


Figure 3.12: Whole cell patch-clamp recording from cultured BAOSM cells.

The seal-test reflects the transient voltages resulting from pulsed current injections into the sealed cell membrane. The exponential voltage decay is a function of membrane, pipette tip, and parasitic capacitances, as well as the amount of injected current. The measured features of the whole cell recording roughly approximate that of a prototypical cardiac AP, in both amplitude, duration, and ISI.

Initially, the patch-clamping work was intended to characterize the relationship between sensor (planar microelectrode) recordings of extracellular potentials and cellular depolarization events, to enable more accurate detection and classification of these potentials.[3] However, we quickly discovered that BAOSMC, while particularly robust to the iniquities of multi-party cell culture and care, only

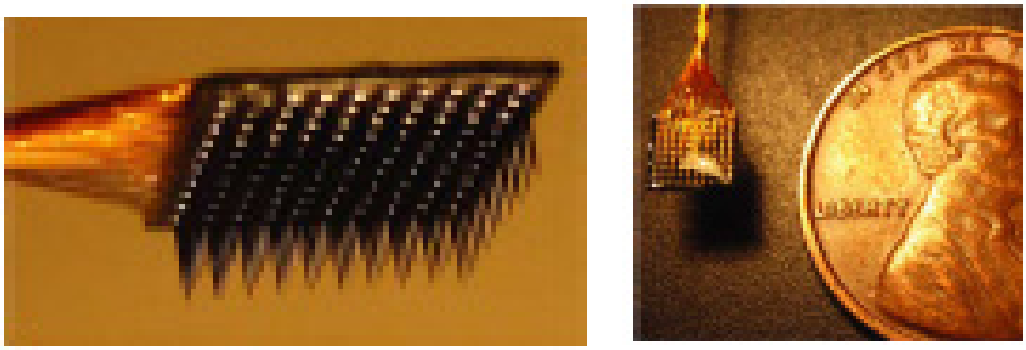
[3] Although a real recorded action potential waveform is distorted in non-linear fashion as it travels through the biological impedance network that separates cell membrane from recording electrode; for adaptive detection and classification of spikes, this distortion can be characterized over time.

rarely exhibited spontaneous depolarizations and were thus not the ideal surrogates for networks of spiking neural cells. It is my understanding that to this day the only published recorded action potentials from BAOSMC are reported in [12] and [52]. Widely-reported experimental successes using cardiac myocytes [53], and actual neural cells [54], suggest that these would be more appropriate cell lines for the intended experiments.

Patch-clamp recording remains the gold standard for measuring channel currents and corresponding depolarizations of individual neural cells *in vitro*. However, whole-cell recording is a destructive technique which inevitably results in cell death. In addition the size of the instrument prohibits clamping more than 2 or 3 neurons at a time, and so is unsuitable for recording from populations of neurons, or for chronic implantation. As a result, engineers seeking to make meaning out of signals generated by tens or hundreds of neurons, as opposed to two or three APs, must employ a different technology to measure and record such data – the paradigm thus shifts from direct measurement of channel currents of a single neuron to indirect recording of transient shifts in extracellular electrical potentials. Moreover, instead of recording from two or three sharp glass electrodes filled with electrolyte and wire, engineers seeking to capture population data have fabricated integrated arrays of sharp microelectrodes for chronic recording and stimulation. The industry standard is the Utah array, shown in Fig. 3.13 (a) and (b).

In vivo measurements are beyond the province of this dissertation, so I obtained single unit ferret neural recordings from Professors Asaf Keller of the University of Maryland Medical School and Shihab Shamma (and Dr. Ping-Bo Yin)

from the Neural Systems Lab ("NSL") of the Institute for Systems Research at UMCP.



(a) (b)

Figure 3.13: (a) "Utah" microelectrode array; (b) Utah array next to penny for comparison. [56]

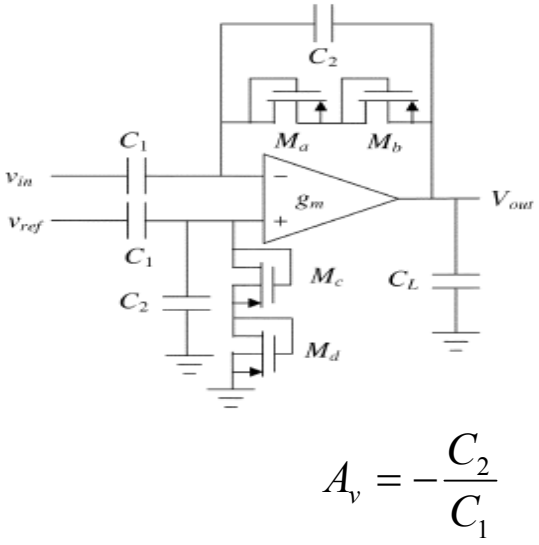
The data from the NSL was obtained using custom microelectrodes chronically implanted into the auditory cortex of live ferrets and as we shall see in the spike sorting section of this dissertation, the recordings are excellent. However, as with the patch-clamp technique, the quality of their recordings comes with a price – the size of the discrete external instruments that perform the signal amplification, filtering and conditioning. As an added overhead, the ferrets must be immobilized during testing to avoid introducing extraneous noise to the tethered system. It is neither desirable, nor in the author's opinion, ethical, to tether human subjects to such extraneous instrumentation. To truly restore lost cognitive, sensory and motor function to the victims of neural injury and disease, a fully implantable system is required.

3.2 Integrated Electrode Arrays

The Utah array has served as the closest thing to emerge as a standard for sharp integrated microelectrodes, but there has been an explosion of new microelectrode designs over the past 10 years. Some electrodes are simple spikes [55], others are planar arrays [54]; some employ metal interfaces, others use conducting polymer [56]; some are hardwired to external instruments, others are integrated with amplification, and recording capabilities [5]. It is to this end that we aspire – leveraging the density and power efficiency of standard CMOS technology, to implement several novel integrated electrode arrays for monitoring neural activity *in vitro*, for demonstrating proof of principle integrated spike sorting systems on-chip, and for manipulating nerve cell growth, orientation and network formation.

To be clear, we have not developed any novel interface chemistries or electrode geometries as part of this dissertation. Nor do we look to reinvent the wheel in terms of integrated amplifiers, although we have trimmed one model to our specifications. We seek to build ultra-low-power integrated systems for neural signal processing on-chip, but first and foremost, our systems must be capable of accurately and reliably recording incident extracellular potentials. More specifically, they need to be able to record signals on the order of 100's of μV to 100's of mV, across tens to hundreds of channels, all while not exceeding several mW of chronic power dissipation. As a starting point down that path, we have scaled down Harrison's classic neural amplifier [57], Fig. 3.14, to a commercial 0.5 μm process, and integrated this design, with high density electrode arrays, and also with integrated spike sorting architectures. Harrison's basic design has a gain of ~ 100 (39.5 dB),

bandwidth of 7.2 kHz, operates on +/- 2.5 V rails and a remarkably low input referred noise of 2.2 μV, rms. The one published in [12], operates on +/- 1.65 V rails, exhibits a gain of approximately 100 (> 38.4 dB measured) , has a bandwidth of ~ 3 kHz, low-noise design, and an area just over half that of Harrison's (.16 mm^2 versus .09 mm^2). Figure 3.15 shows my own variant on the Harrison bioamplifier, which is again more than halved in size, and laid out using common centroid techniques for matching.



$$A_v = -\frac{C_2}{C_1}$$

Figure 3.14: Harrison bioamplifier schematic. [57]

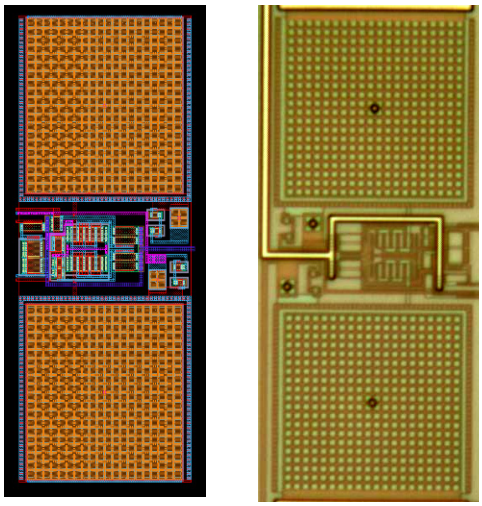


Figure 3.15: Haas scaled version layout and photomicrograph

As in Harrison's original work, both scaled amplifiers meet all of the requirements for neural signal recording: power consumption for an array of these amplifiers comes in at well under the 80-100 mW/cm^2 budget; bandwidth is sufficient to capture the salient features of AP shapes; and the gain set by the ratio of feedback capacitors, C1 and C2, is sufficient to shift small extracellular signals well above the noise floor for subsequent on-chip processing. Having outlined the basic neural

recording requirements and technology, we next turn to completely integrated systems.

3.2.1 Neurite outgrowth

In 2006, we reported data from a design with ten integrated bioamplifiers, with an ideal gain of 40 dB, and planar commercially cut electrodes, Fig.3.16 [12]. As previously noted, these experiments and those found in [52] represent the only known reported BAOSMC action potentials in the literature, Fig. 3.17.

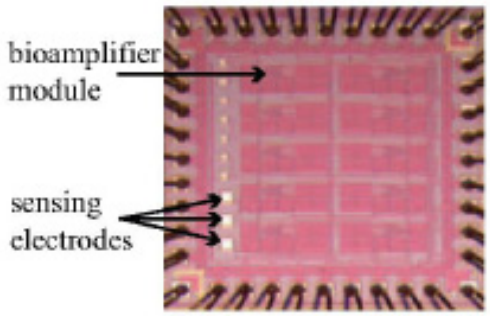


Figure 3.16: Array of first generation scaled bioamplifiers, [12].

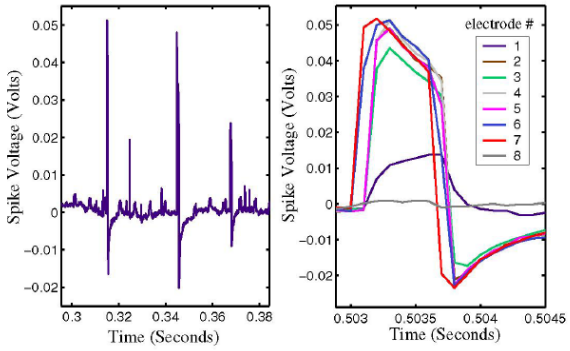


Figure 3.17: Recorded signals from cultured BAOSMC, on one channel, and across all eight channels. [12]

Although these sensors were fabricated in an integrated CMOS process, each channel was implemented with its own dedicated bioamplifier; as a result, it was only possible to place ten recording channels onto a 1.5 mm x 1.5 mm MOSIS tiny chip. Inter-electrode distances can be rendered almost arbitrarily small (even violating design rules), but sensor resolution is limited by the density of the amplifier circuits. Thus, in order to enable higher resolution recording from, *e.g.* cultured neural cells, we designed and fabricated a suite of 128 x 128 electrode arrays at 7-14 μm pitch, on 2.5 x 2.5 mm^2 dies, with simple pre-amplifier circuits at each electrode, to buffer the

weak neural signals and minimize the effects of readout noise. High input impedance MOSFETs buffer the signals near their source, thereby reducing attenuation and improving signal integrity [58], [59].[4] As a result, a single bioamplifier can be used to drive 16,384 high fidelity, multiplexed signals off-chip.

Multiple copies of each design were submitted so that we could compare the performance of electrodes defined by commercial versus in-house glass cuts. The 7-14 μm pitch of our electrodes was comparable to the then state-of the-art MEA reported in [54]; our principal point of novelty was obviated when it was decided not to purchase the photomasks required to cut sub-micrometer electrodes in the array. However, the array served a testbed for the two high density pre-amplifier architectures: (a) a single transistor common-source buffer; and (b) a differential sensor to compensate for bath and localized DC electrical potentials. Fig. 3.18 shows the schematics and layout for the two preamplifier architectures:

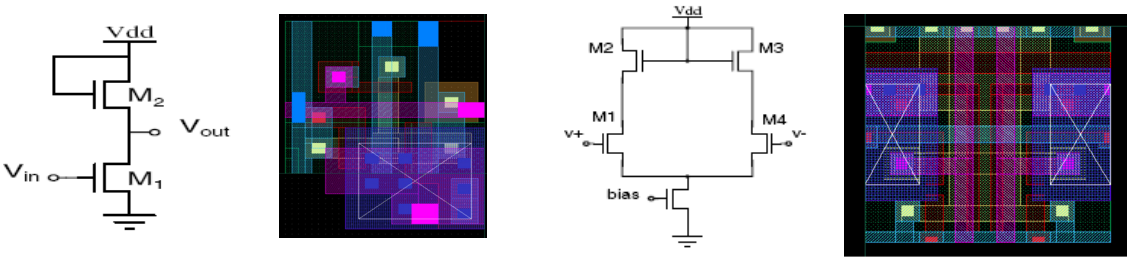


Figure 3.18: (a) single-ended common-source pre-amplifier schematic and layout; (b) differential common-source preamplifier schematic and layout.

The single transistor common-source amplifier acts as a unity gain buffer driving the load imposed by the readout circuitry. Monte Carlo simulations performed with respect to the published corner parameters suggest that the common-

[4] Interfacial capacitance would not dramatically perturb measurements of extracellular potentials due to the relatively high Helmholtz layer capacitance (approximately 0.1 pF per μm sq. [60]), and the comparatively low series gate capacitance (on the order of 1fF per μm sq.).

source preamplifier can drive a 100 μV, 1 kHz signal onto a 128 element column bus with less than a tenth of one percent attenuation; although actual circuit performance is attenuated somewhat as a function of the bioamplifier performance and also non-ideal parasitic switch capacitances.

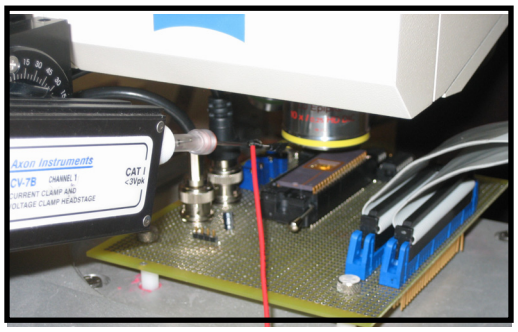


Figure 3.19: Bench-testing apparatus for probing microelectrodes using signal generator.

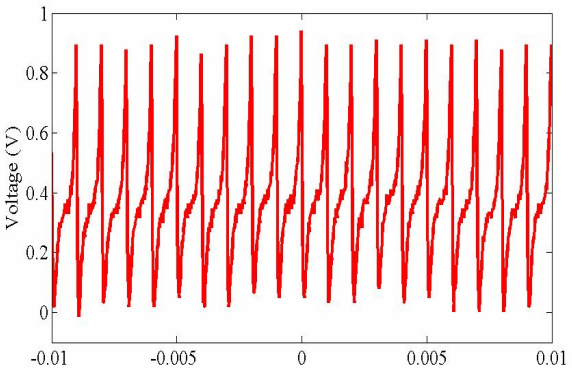


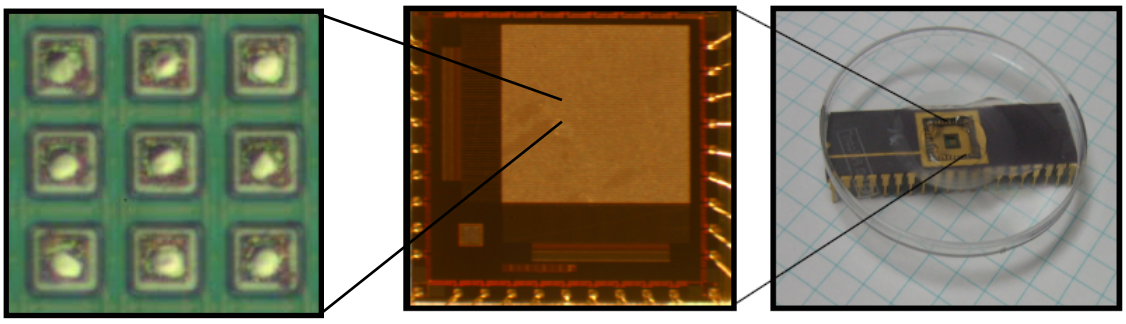
Figure 3.20: Data from single-ended pre-amplifier, fed by arbitrary waveform generator and buffered by bioamplifier from [12].

This is illustrated in Fig. 3.20, which represents the bench-test output of the CS preamp and bioamplifier as buffered by a unity gain op-amp, in response to a synthetic AP. In this case, I connected an artificial AP signal directly to a pinned-out electrode; it had an amplitude of ~20 mV, peak-to-trough, and a frequency of 1 kHz. Note that although the ISI is precisely preserved, the fast transient components of the simulated AP are attenuated by approximately ½ as they push into the roll-off freq. (~3kHz) of the bioamplifier. Additional bench-testing using the probe-station featured in Fig.3.19, demonstrated the performance of the buffered CS amplifier when a signal was coupled through directly to the aluminum pad on the sensor surface. By comparison with the directly pinned out electrode, a probed pad contact

yields a significantly greater degree of attenuation; this is believed to be a function of interfacial capacitances and process imperfections.

By contrast, the NMOS-only differential amplifier was designed to provide a small gain while minimizing common-mode noise. Gain for the differential pre-amplifier is a function of the ratio of the widths of the input transistors to the active loads: in the ideal case of perfectly matched process parameters and taking the lengths of all transistors as equal, $A_v = \sqrt{W_i / W_L}$, where W_i is the width of each of the input transistors, and W_L is the width of each of the active load transistors. With a fabricated W_i of approximately 9 μm, and W_L of about 1 μm, we expect a gain of about 3. As with the common-source amplifier, whose theoretical frequency response is nearly identical, there is should be no appreciable frequency roll-off in the range of interest for this preamplifier configuration.

Having characterized several sensors on the bench, and desiring to conduct biological experiments with these sensors, it was first necessary to package them – to passivate the toxic aluminum electrodes, to encapsulate the bond wires, and to form a culture well for cell media. First, the chips were affixed into ceramic DIP40 packages and wirebonded through MOSIS. Next, we electrolessly plated the aluminum electrodes with gold. Then, we insulated the bondwires using a biocompatible UV-patternable polymer, Loctite 3340, leaving the electrodes exposed. Fourth, we affixed a custom-milled culture well to the packaged surface with silicone glue. Fig. 3.21, below, shows: (a) a close-up of electrolessly-plated 14 µm planar electrodes; (b) a photomicrograph of the fabricated 128x128 differential sensor array; and (c) a fully packaged chip.



(a) (b) (c)

Figure 3.21: Left, photograph of electrolessly plated planar commercially cut electrodes; center, photomicrograph of fabricated 128x128 differential sensor array; right, packaged electrode array.

I conducted several sets of biological experiments using the packaged sensors. In one, I plated BAOSMC onto the surface of the sensor, added media to the culture well and incubated them overnight to permit the cells to adhere to the sensor surface. The following day, I would visualize the adherent cells using an Axiotron confocal microscope, and record from selected channels of the array. To spur the naturally quiescent BAOSMC to generate and sustain APs, I would occasionally dose the bath media at selected points with tetraethyammonium ("TEA") and/or salts.

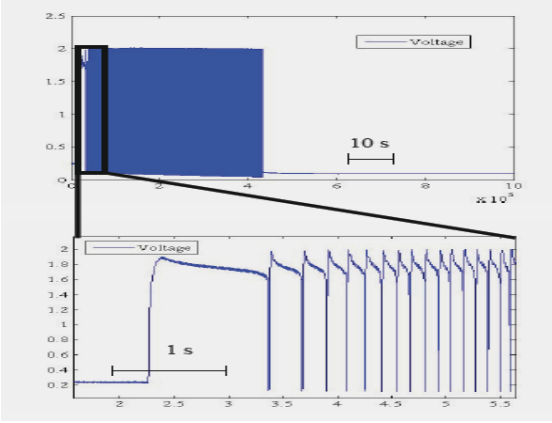


Figure 3.22: Recorded activity of cultured BAOSMC in HBSS dosed with ~2mM TEA.

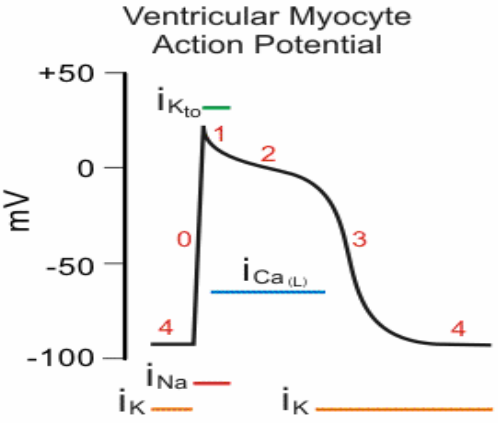


Figure 3.23: Stereotyped cardiac action potential. [61]

Data from one such experiment is shown in Fig. 3.22. For comparison, Fig. 3.23 shows a schematic representations of a cardiac AP. For subsequent generations of chips, in future collaborations with other faculty and institutes, I intend to pursue neural recording on-chip.

3.2.2 Programmable electrode arrays

Integrated electrode arrays provide high density, low power neural recording capabilities and permit sophisticated on-chip signal processing circuitry. However, standard CMOS was designed for digital circuits – device mismatch and process variation that would not significantly affect digital designs can have a pronounced impact on analog circuits, particularly CMOS sensors used to decode very small neural signals. For high-density arrays of neural sensors, it is essential to be able to reliably correlate signals measured across different sites in order to render an accurate representation of the environment sensed. However, even state of the art MEAs using complex compensation circuitry exhibit variation in excess of 30 percent [54]. This problem is further compounded when random, but quantifiable environmental inhomogeneities associated with *in vitro* cell culture are introduced. Thus, in order to address these problems and enable more reliable comparison of signals across the MEA, I developed a series of programmable VLSI sensor arrays for compensating these offsets and accurately monitoring electrogenic cells *in vitro*. In addition, I have fabricated and performed bench tests on a completely passivated planar electrode array that is designed to exploit the tendency of biological cells to grow along electric field vectors, a phenomenon known as galvanotropism. In particular, these arrays are

designed to manipulate nerve cells, including neurites of developing axons, to enable user-controlled patterning of custom neural (and other cellular) networks. In addition to the specific examples described, programmable array technology may also used to collect and concentrate charged analytes and particulate matter, and to implement site-specific *in situ* bio- and chemical-functionalization for such applications as DNA arrays, and other assays.

A. Programmable High Density CMOS Microelectrode Array [13]

For the first proof-of-principle design, the principal aims were electrophysiological recording, so that it was important to not exceed the pitch of existing microelectrode arrays (14 μm), while demonstrating basic programming of individual array elements. In striving to accomplish this programming within the confines of a cell-sized sensor, I stripped the circuits down to the bare minimum, removing as much of the extraneous architectures as possible and attempting to make circuits multi-functional where feasible. As a result, I developed the circuit shown in the following schematic and layout of Fig. 3.24, which is a current-mode amplifier

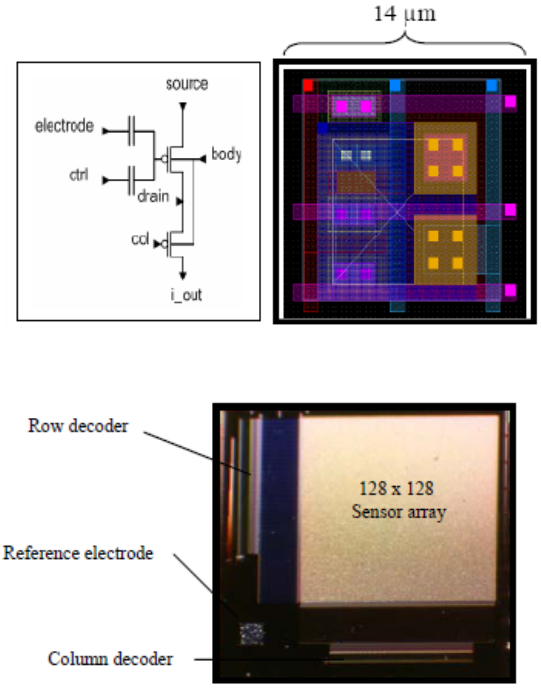


Figure 3.24: Schematic, layout and fabricated 128 x 128 array of programmable electrodes.

with non-volatile analog memory built into each element of the sensor array. The control gate aids in programming and is also useful in characterizing the performance of the array because it is in parallel with the cell-sensor interface ("CSI") electrode. The CSI electrode is a commercially cut exposed aluminum electrode which is capacitively coupled to the floating gate of the programmable PMOS transistor. Slow-moving DC offsets are rejected and small-signal AC transients such as those from cells coupled to the electrodes on the packaged sensor surface are transduced into drain currents that are buffered through a digital column switch to a common bus. These signals may be further buffered using, *e.g.*, integrated operational amplifiers, but for this prototype they are read out directly through user-controlled switches with adjustable rails.

A photomicrograph of the fabricated array is also shown in Figure 3.24. The large golden square that fills the bulk of the 3 mm x 3 mm die comprises a 128 x 128 array of programmable sensors, and the remaining elements are the readout and programming circuitry. In attempting to reduce the footprint of the overall design, I employed digital row and column select decoders to control the source and drain voltages and thereby regulate both programming and readout across the array. For ease of testing, each decoder was fabricated with its own independent, user-tunable rails so that programming would not be constrained by fixed biases; this helps avoid undesired forward biasing any of the diodes in the array and avoid latch-up. Source voltages are asserted column-wise, while drain voltages are controlled by the row decoder. This enables a user to access any element of the array desired for readout or programming, and effectively isolates other elements because no current will flow

absent the required source voltage and no programming will occur without the appropriate voltages at the source and the drain of the programming transistor.

Each channel of the array may be programmed independently using hot electron injection, and can be erased by UV light or Fowler Nordheim ("FN") field-induced tunneling. Hot electron injection in MOSFETs is a function of transistor source- ("S") and gate- ("G") drain ("D") voltages, has been empirically shown to obey the following relationship [64]:

$$I_{inj} = \alpha \cdot I_s \cdot \exp\left(-\frac{\beta}{(V_{gd} + \delta)^2} + \lambda \cdot V_{sd} \right) \tag{3.1}$$

where α, β, and δ are experimentally derived process dependent constants, *Is* is the S-D current flowing through the transistor, and *Vgd* and *Vsd* are the voltages across the gate-drain and source-drain regions, respectively. These parameters have previously been characterized for the commercial 0.5 μm process in which the sensor array has been fabricated.[5] Thus, for a controlled S-D voltage, which may be asserted through user-controlled tunable switches, the injection current is a function of: (a) the current flowing through the programming transistor, and (b) the floating gate voltage which falls with injection and rises with tunneling. This current results in a stored charge on the floating node which modulates the gain and offset of the sensor element.

The array may be uniformly erased by exposure to strong UV light for several hours, and it is also theoretically possible to erase these structures by field-induced Fowler-Nordheim tunneling. In MOSFETs, FN tunneling exhibits an exponential dependence on the voltage across the oxide barrier that is given by [64]:

[5] by Eric Wong while he was a student and colleague in the ECE department.

$$I_{tun} = -I_{tun0} \cdot WL \cdot \exp\left(-\frac{V_f}{V_{ox}}\right) \tag{3.2}$$

where I_{tun0} is a pre-exponential current, Vf is a process dependent constant, and Vox is the voltage across the oxide barrier. In either case, the programming mechanism operates to change the charge and thereby the voltage stored on the floating node between the CSI and the gate of the PMOS sensing element.

Simulated and measured data from the array is presented in Figs. 3.25, 3.26, & 3.27. The simulation data shown below in Figure 3.25 reflects the output current of individual sensors with mismatch variation: (a) before ideal programming; and (b) with the offsets mathematically removed. However, this simplistic view of mismatch compensation does not fully reflect the realities of programming single-transistor amplifiers that behave differently across many regions of operation, and with different biases. In order to more fully understand these real world issues, prior to receiving the fully fabricated sensor array, I tested some other analog front ends that I had previously fabricated which share the design, if not the pitch of the sensors in the microelectrode array. Figure 3.26 shows the results of those preliminary tests on actual circuits with architectures nearly identical to those on the fabricated array. More specifically, those figures are the results of programming arbitrary voltages onto the floating node of individual, previously fabricated sensors to illustrate how programming affects gain and offset for this design. In the top box of Fig. 3.26, programming results only in an offset shift of the input sine wave, as can be seen in the inlay when these offsets are mathematically subtracted using MATLAB. In the

bottom box, programming arbitrary voltages onto the floating node affects both gain and DC level, as can be seen in the second inlay, which again is the result of mathematically subtracting out the offsets.

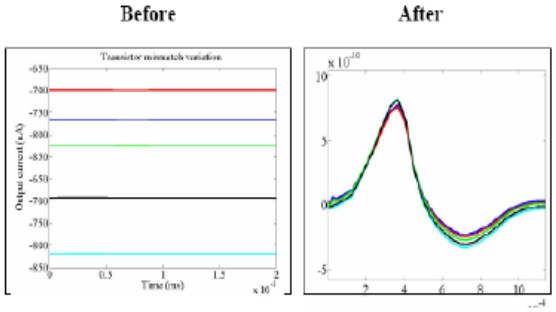


Figure 3.25: Before: simulated swamp the signal; After: the offsets are mathematically removed.

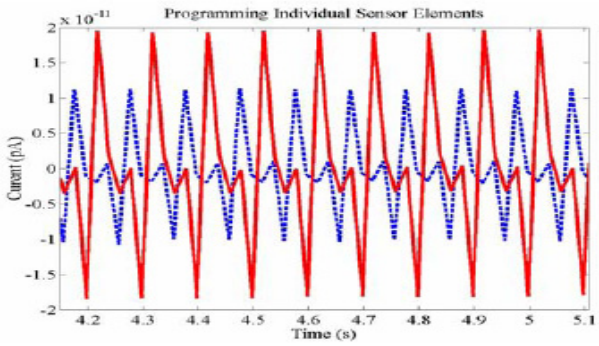


Figure 3.27: Sensor array element before and after programming the floating node by injection. Both gain and offset are affected. [13]

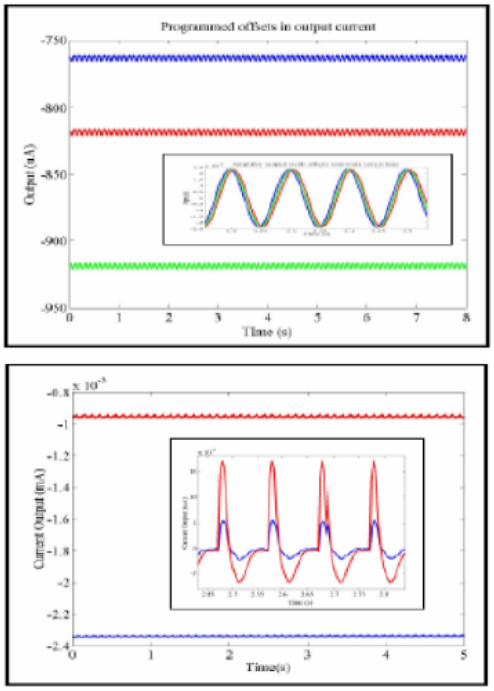


Figure 3.26: Programming arbitrary offsets onto floating gates in order to shift I-V curves and tune gain. Within each box, input signals are identical. Inlays show signals with DC offsets subtracted. [13]

These figures illustrate both the problems and the promise of the programming techniques. In the former case, I was able to control injection well enough to avoid pushing the single transistor amplifiers into a different mode of operation, and hence the programming voltage simply shifted the I-V curve with a DC offset. On the other hand, I believe in the latter case, the programming voltage actually pushed the transistor deeper into saturation, resulting in a higher overall

current (the axes are negative on that figure), with a lower signal gain. I observed similar phenomena when I bench tested the fabricated floating gate array. Fig. 3.27 shows the output from a single sensor element of the fabricated array in response to a synthetic spike train input asserted directly at the control gate. The dashed line is the output before programming, and the solid line is the output afterwards; the output spike trains are normalized to their respective mean voltages for comparison. In this instance, programming impacts gain considerably more than offset. Fig. 3.28 is a reproduction of the packaging flow, with a photomicrograph of the fabricated sensor array.

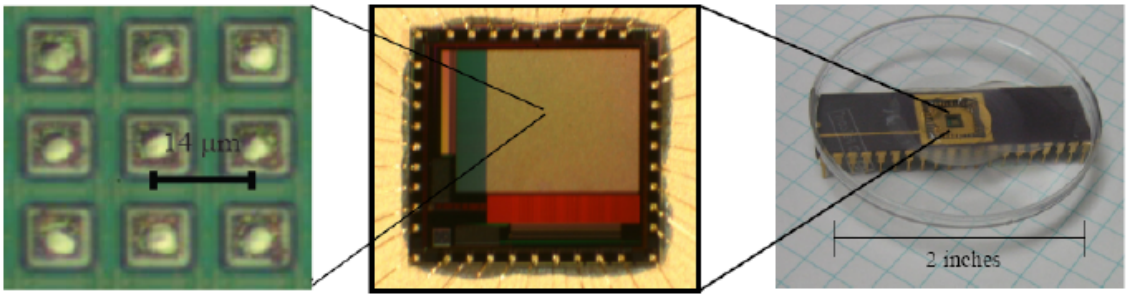


Figure 3.28: Previously fabricated microelectrode array with identical pitch and die size to the fabricated sensors, plated and packaged.

B. Galvanotropism

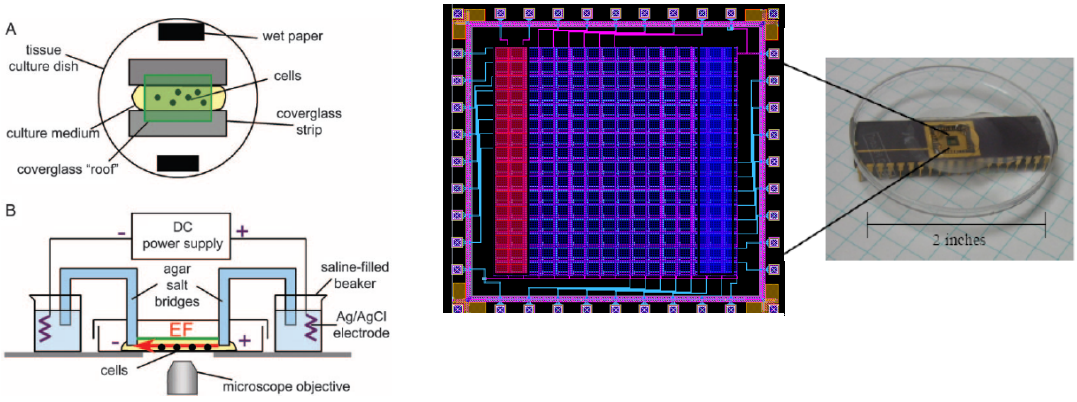
The ability to program analog offsets onto the electrodes of an array is valuable to many applications in addition to neural recording. Stroke, spinal cord injury and neurodegenerative diseases such as ALS and Parkinson's debilitate their victims by suffocating [1], cleaving communication between [2], and/or poisoning [3] entire populations of geographically correlated neurons. Although the damage associated with such injury or disease is typically irreversible, studies have shown

that neurite outgrowth itself can be directed and enhanced *in vitro* by subjecting developing cells to applied electric fields [65]-[74]. Together with *in vivo* studies implicating endogenous electrochemical potential gradients in neuronal differentiation and development [71], and clinical research showing that applied electric fields ("EF") can stimulate recovery of severed neural pathways [74], this research offers the possibility of substantial therapeutic advances in the treatment of nerve injury and neurodegenerative diseases. While promising, the systems used to realize these therapeutic applications are typically macro-scale implementations with fixed EF vectors [71], coarse controls [74], and poor spatial resolution [75], which limits the potential for such applications. In view of these limitations, the present research is directed to developing a programmable, high-resolution micro-system for inducing EF-directed neurite outgrowth, and to testing this system on several different cell lines. In performing this research, I hope to develop a tool which researchers can use to obtain deeper insights into the phenomena underlying neural development and growth, and which clinicians might use in the treatment of damaged and diseased nerve cells.

In so doing, I designed, fabricated and characterized an integrated planar electrode array capable of generating localized, programmable EF gradients for stimulating and manipulating neuronal growth. The electrodes are paired with integrated readout circuitry, and employ floating gate electrodes to pattern EFs and potential gradients. The fabricated microchip has been packaged using a photo-curable polymer to provide a culture well for cell growth and to insulate the electrical components of the chip from the cells and fluid media, as with the previous work.

The passivated glass surface of the chip remains exposed and, in conjunction with the polymer packaging, serves as the culture well substrate which may be coated with an adhesion or neurotrophic factor to promote neurite outgrowth. For the prototype, controls are discrete off-chip components in order to permit better characterization of the electrode array and the generated EFs. In characterizing the performance of the system, local field strength is both modeled and must be experimentally measured.

The prototype chips are comprised of a 16 x 16 passivated array of electrodes coupled to analog floating gate memories that store precise analog voltages on each electrode instead of a digital bit. Circuits were modeled in part using the Cadence Spectre circuit simulator; layout of this design as shown in Figure 3.29 (b) was done with the Cadence software suite, Virtuoso.



(a) (b)

Figure 3.29: (a) Typical set-up for two electrode galvanotropism, [72]; (b) layout of fabricated 16 x 16 array for performing arbitrary field pattern galvanotropism, and photograph of typical bio-packaging.

The tiled pattern of voltages set on the array of electrodes will define the DC EF gradients used to direct neurite outgrowth; a fixed-potential counter-electrode may be mounted atop the array to ensure proper behavior. Electric field vectors will be

studied using three-dimensional (3-D) finite-element-analysis ("FEM"), taking each planar electrode in the array, including the passivation (insulation) layer and corresponding fluid volume as a single element of the tiled array. EFs will also be measured indirectly with potentiometric fluorescent probes. Theoretical calculations suggest that the prototype FG system currently in fabrication is capable of generating EF strengths up to several orders of magnitude higher than that reported in the literature. Specifically, taking the passivation layer as several hundred microns thick and the top plate counter-electrode as being situated 1-2 mm above the surface of the chip, given supply rails which may be set to ± 5 V, the tiled vertical EFs could be as high as 10 V / mm and, owing to electrode geometry and pitch, the lateral growth-inducing fields could be up to 200 times greater. The entire array may be reconfigured within minutes using only electric currents and voltages (and not physical manipulation of switches or other apparatus). Initial experiments will attempt to program relatively uniform and unidirectional gradients, to reproduce the experiments reported in the literature. Subsequent investigation will focus on: (a) the effects of novel patterned DC EFs on neurite outgrowth; and (b) the impact of time-varying EFs on neural development.

Having characterized a suitable micro-scale system for stimulating cells, future efforts will be directed at conducting a series of experiments designed to assess the effects of different applied DC EF gradients to cells cultured in the fabricated, encapsulated systems. For the first series of experiments, I intend to work with a clinical research partner to culture mouse nasal explants directly atop the partially encapsulated chips and expose them to a unidirectional EF gradient over several days.

These cells are obtained from colleagues at the NIH and also in the Bioengineering department of the University of Maryland, College Park. To affect the cell growth, we first generate the unidirectional EF by programming each column of the array to a specific voltage, generating a columnwise potential gradient from left to right comparable to the 10-150 mV/mm used to induce directed neurite outgrowth in several of the reported studies [65]-[73]. Once the EF pattern has been stored on the array, we clean the glass substrate and add any required adhesion factors (such as laminin). After preparing the microsystem, harvested cells will be introduced to the culture well by conventional techniques along with culture media sufficient to support 6-7 days growth. The entire apparatus will be placed in an incubator under standard conditions (37^0C, 5% CO_2), and we will visualize the emerging growth cones using conventional microscopy. For the preliminary experiments, periodic digital images will be taken over the course of several days, although depending on the observed rate of growth, this sampling period may be increased or lowered. These digital images will be registered, sorted and evaluated, first by hand, and then by custom software (MATLAB) to automatically track neurite outgrowth for a particular neuron.

Subsequent experiments will employ (a) collagen; (b) poly-L-lysine; and (c) NGF to coat regions of the glass substrate to ascertain whether any EF induced outgrowth is subject to a substrate-induced dependency. The rate, direction and morphology of neurite outgrowth will be monitored during this period using conventional microscopy. Subsequent experiments will investigate the effects of different EF spatial, and temporal patterns on neurite outgrowth of the nasal explants. Finally, we will conclude the proposed work by attempting to extend the results to

other cell lines, such as NGF-differentiated PC12 cells and dissociated hippocampal cells. Previously reviewed work has shown *Xenopus* motor neurites turning toward the cathode, while PC12 neurites turn the other way [72]. Such definitions may further be stretched by the unconventional electric fields generated across the proposed array; virtual "anodes" and "cathodes" will simply be the most positive and negative regions on the chip.

In summarizing this section, I have designed and fabricated a novel architecture for stimulating *in vitro* cell cultures with programmable DC EFs. Future work will include more sophisticated experiments to elucidate the mechanisms of EF-induced growth by selective modulation of environmental conditions, such as extracellular Ca^{++} concentration or actin polymerization (using, e.g., latrunculin A).

C. Other Applications

Finally, it should be pointed out that the programmable electrode arrays presented here have many variants, including different geometries (such as nano-electrodes), materials (e.g. carbon nanotubes and organic electrodes), circuit architectures (different trimmable amplifiers), and modes (current, voltage, charge, etc…). One set of experiments indicate its suitability for concentrating analytes or particulate matter into desired regions of an array. The possibilities are practically limitless.

3.3 *EMG*

Finally, we briefly discuss surface electromyography ("EMG") for biofeedback and rehabilitation. From 2006-2008, I mentored a team of Gemstone undergraduates and helped them to develop a closed-loop surface-EMG feedback system for the rehabilitation of hemiparetic stroke victims.

Surface EMG is the measurement of electrical potentials generated by muscles when they contract. More specifically, motor nerves that terminate in muscle fiber give rise to local depolarizations of the muscle fibers innervated by those nerves – the aggregate activity of a group of muscle fibers results in a synchronized depolarization called a motor unit action potential, or MUAP. For a given muscle, such as the hamstring, MUAPs superpose and give rise to a collective waveform whose amplitude reflects the intensity of a muscular contraction and whose period coincides with the frequency of stimulation. Fig. 3.30 illustrates the basic principles of signal acquisition and decomposition. MUAPs are important in the characterization and rehabilitation of the victims of neurophysiological diseases and injury such as stroke – they become a barometer for the strength of motor innervations, which in turn are a function of cortical vitality.

Together with the Gemstone team, previously acknowledged, I coordinated with clinicians to develop a system that would measure these MUAPs and provide biofeedback for the rehabilitation of stroke victims. We developed miniature instrumentation amplifiers and filters for recording the very small surface potentials, ~10uV-10mV, representing subcutaneous MUAPs. Gel surface electrodes were connected to the amplifier inputs to minimize interfacial impedance, and

characteristic signals were recorded from, e.g. arm-wrestling students. For biofeedback, our clinical partners at UMBC School of Medicine, devised a therapy regiment and assisted us in identifying appropriate recording sites – ultimately we chose 6 muscle groups involved to assess gait. Fig. 3.31 shows representative surface-EMG data that we recorded from a student's calf muscle.

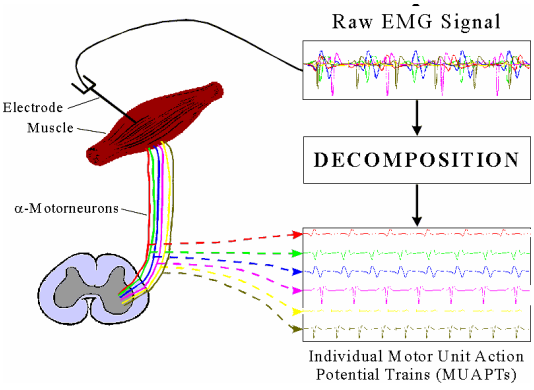


Figure 3.30: Basic principles of EMG acquisition and signal decomposition. [76]

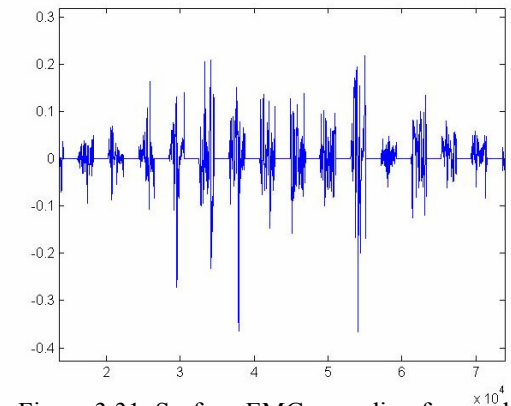


Figure 3.31: Surface-EMG recording from calf-muscle.

Future work will involve measuring the impaired gait of hemiplaegic stroke victims, for rehabilitative biofeedback, and next generation systems integration.

Chapter 4: Spike Sorting

The principal aim of this work has been to develop mixed-signal VLSI circuits that can process real-time neural signals recorded from cultured (*in vitro*) and living (*in vivo*) biological systems. In particular, I have designed, fabricated and characterized circuits for detecting, extracting the salient features from, and identifying the source of neural action potentials. In so doing, I have evaluated the performance of these systems against other mixed-signal spike sorting circuits, comparable digital systems, and theoretical performance metrics. This work has provided me with a deeper understanding the advantages and limitations of low-power analog computation, and enabled me to test some novel stochastic and adaptive architectures which are designed to enhance the precision and accuracy of analog VLSI systems generally. More particularly, I have addressed some of the key limitations of previous generations of implantable signal processing and biosensor architectures – power consumption and reliable encoding of neural data – and offer some novel approaches for next generation spike sorting systems.

I undertook this research to meet a perceived deficiency in the broad development of neural prosthetics. Specifically, I observed that first generation neural implants typically recorded and amplified localized time-varying extracellular potentials, but performed limited additional signal conditioning on-chip [77]. State-of-the-art systems fared little better: although microelectrode arrays have been chronically implanted into the cortex, many such "implants" are essentially just wires, [78], onto which low-resolution, high-power ASICS have been mounted [5], [6], or mixed-signal threshold detectors that are incapable of discriminating between spikes

from neighboring neurons [79], [80], [81]. Each of these systems suffered from the same general deficiency – they did not reliably encode the vast amount of incident neural data, and therefore required external hardware to make any sense of it. Furthermore, while cultured cells may be able to endure the heat generated by clocked, high-powered digital processors (see, e.g. [77]), neurons *in situ* cannot. As a result, neural implants continued to rely on hard-wired connections between microelectrode and PC that pierce the skull and can succumb to noise, corrosion, signal attenuation, and infection. Thus, a principal motivation for the present work has been to develop low-power VLSI architectures that will accurately classify incident neural data for low-power RF transcutaneous transmission (or additional on-chip processing for control, etc.), and thereby to enable a new generation of implantable cognitive and cortically controlled neural prosthetics. Such prosthetics could be implemented to restore lost or impaired vision, hearing, and motor control, among other possibilities.

In addition, when implemented in conjunction with densely populated microelectrode arrays and low-noise bioamplifiers for monitoring the neural activity of cultured cells, the spike sorting circuits I developed can also provide a robust platform for low-false-positive cell-based sensing. In this context, small footprint and integrability are the key constraints – size and power-efficiency matters. Because the fabricated design takes advantage of the efficiencies of real-time analog signal processing and obviates the need for, e.g., A/D conversion and vast memory stores, it is a fraction of the size and corresponding power cost of comparable DSPs. Leveraging the efficiency of analog VLSI systems thus enables a new platform for

low-power cell-based sensing that is both portable and precise. As one particular example of how such a system could be implemented, olfactory nerve cells are known to respond to different classes of chemical moieties with different electrophysiological signatures. If such cells were cultured atop a microelectrode array whose outputs were encoded by one of the spike sorting circuits, it should be possible to detect changes in individual neural signatures by observing the incident action potential frequencies. Ultimately, the accuracy of action potential detection and classification will determine the number of false positives a sensor reports, and so another aim of the research has been to enhance the receiver operating characteristics ("ROC") of cell-based sensors by increasing the precision of the sorting circuits.

In furtherance of this research, I have studied cellular neurophysiology and mixed-signal VLSI design. I have developed and tested novel cellular sensor architectures with embedded signal processing capabilities, and have experimentally investigated the cell-CMOS interface.

The original contributions of this thesis with respect to the field reside in the design, fabrication and characterization of ultra-low power programmable floating gate template matching circuits for the detection and classification of neural action potentials. Specific contributions to the field include: (a) the overall architecture; (b) application of floating gate adaptation and template matching to solve the detection and classification problem under competing constraints of low-power dissipation and high computational precision; (c) novel on-chip variance estimation circuitry; (d) novel asynchronous current-mode weight-update circuits; (e) unique silicon neuron template generation mechanism; and (f) a VLSI implementation of a theoretical non-

linear energy operator ("NEO") to threshold incoming signals for unsupervised template generation.

This chapter of the dissertation is organized into three parts. In the first, we explore the relevant foundations of mixed-signal stochastic computation. Next, we perform an in-depth review of existing spike sorting algorithms, methods and architectures. Finally, I present my own contributions to the field, represented by ultra-low-power floating gate template matching circuits for neural spike sorting, and place this system into the context of existing and future work.

4.1 Mixed Signal Stochastic Computation

Almost twenty years ago, Carver Mead proclaimed his "conviction that the nervous system … contains computing paradigms that are orders of magnitude more effective than are those found in systems made by humans." [45]. Drawing inspiration from nature, nurture and necessity, I now share Mead's conviction. As a result, the systems investigated, designed, fabricated and tested are all informed by the knowledge that the laws of physics apply equally to silicon and sensory cells, to nerves and to NAND gates. In mapping biological functions onto silicon substrates we exploit the analog subthreshold regime that MOSFETs provide, exchanging some precision for the ability to operate within biological power constraints.

In addition, beyond simply confining my designs to run cool, I have attempted to optimize in a very real sense the balance of analog versus digital computational blocks used, investigating both mixed-signal and mixed-mode (voltage, charge, current) computation. Using novel stochastic and adaptive feedback circuitry, and incorporating floating gate technology to implement multiple input and multiple input

translinear elements ("MITE") and analog memories, I have devised means to not only adapt away offsets, but to compensate for a primary source of imprecision in analog computation – device mismatch. As a result, system and circuit noise can be more accurately characterized and reduced.

The following subsections serve as a focused primer, detailing the theoretical underpinnings that informed my designs.

4.1.1 Analog VLSI

All silicon computation is ultimately analog. Although digital designers can attain arbitrary resolution by confining signal states to a series of binary bins, in so doing they necessarily ignore the computational cost of confinement. Real world signals, such as neural action potentials, are NOT digital and quantizing them costs – it costs precision, it costs speed, and it costs power. Moreover, with respect to the particular application presented here, digital architectures cannot at present attain accuracy sufficient for unsupervised sorting (Compare, e.g., [82] with [83], [84]). Therefore, in this dissertation, I aimed to press the limits of analog precision by adaptively tuning transistor and process mismatch to the noise floor. I also applied conventional noise reduction and elimination techniques, where applicable, to obtain superior analog precision while maintaining a very low power budget. The guiding hypothesis has been that properly calibrated analog circuits can achieve higher precision at a given power consumption than comparable digital circuits.

4.1.2 Low Power Design

Implementing low-power analog design invokes three primary constraints: (1) reduce the power-supply rails as low as possible; (2) incur as little short-circuit

dynamic dissipation as possible; and (3) perform required computations in low-voltage, low-current regime. Thus, most of the circuits implement analog computation using MOSFETs operating in the subthreshold regime. This limits operating current and dynamic power dissipation accordingly. Likewise, although the rails are presently set at 0 and 5 volts to permit floating gate adaptation, this operating voltage can be scaled with the fabrication process. Dynamic power dissipation is further minimized by ensuring rapid digital transitions and using low-power amplifier and current steering circuits where feasible.

In principle, the subthreshold power savings is largely a function of the log-linear relationship between MOSFET current and applied gate-source voltage. The basic equations that govern the operation of MOSFETs in this regime are given below:

$$I_{subthreshold} = \lambda \cdot I_0 e^{\kappa V_{GS}/V_T} \tag{4.1}$$

$$\lambda = (1 - e^{-V_{DS}/V_T}) \approx 1 \qquad \text{for} \quad V_{DS} \geq 4V_T \approx 100mV \tag{4.2}$$

$$I_0 = q\frac{W}{L} t D_n N_0 e^{-q\varphi_{bi}/kT} \tag{4.3}$$

$$\kappa = 1 - \frac{C_{tot}}{C_{ox}} = subthreshold\ \ slope \tag{4.4}$$

For the commercially available 0.5 μm process used to prototype circuits, subthreshold gate-source voltages of several hundred mV yield currents in the pA – nA range, which minimizes static and dynamic power dissipation.

4.1.3 Floating Gate Basics

Floating gates are implemented as polysilicon ("poly") gates which are electrically isolated from both the MOSFET body and direct electrical contact with any input node by layers of gate- and field oxide, respectively. Floating gates are electrically floating nodes, and so can: (a) store charge and thus serve as analog memory elements; and (b) couple multiple gate inputs into a single gate. I employ floating gates for both of these purposes. We shall discuss each in turn:

a. FLOATING GATE ADAPTATION

Floating gates can serve as analog memories that can be used as dynamically configurable tap weights, template bases, wavelet kernels, etc… The mechanisms by which the values stored on these nodes are updated are (1) hot electron injection; and (2) Fowler-Nordheim tunneling. Taken directly from Rahimi's 2002 ISCAS publication, Figs. 4.1 & 4.2 are schematic drawings of a floating gate node, with tunneling implant, and an energy level diagram illustrating the potential gradients that must be overcome to invoke the physical charge injection and tunneling processes.

Floating gate adaptation is readily implemented to nullify threshold mismatch and process length variation in analog VLSI designs. Conventional floating gate trimming circuits employ voltage comparators to adaptively inject and tunnel charge onto and off the relevant gates, thereby providing a means of achieving enhanced precision in such circuits [19]. Simple current-mode trimming circuits have also been

reported [148]. I have included circuitry to nullify current mismatch in my filters and computational blocks to arbitrary precision, using such techniques.

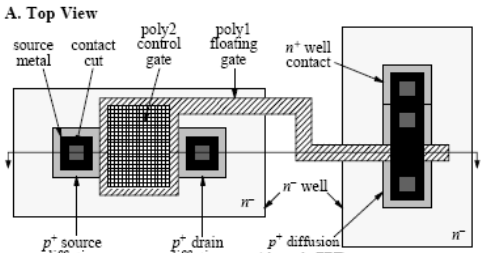


Figure 4.1: Floating gate layout with control, injection and tunneling nodes. [19].

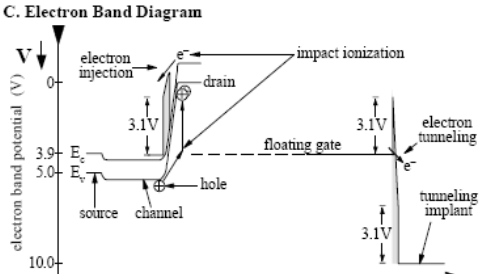


Figure 4.2: Schematic of injection and tunneling mechanisms. [19].

Although the particular details will be addressed in subsequent sections of this dissertation, one way to correct threshold voltage mismatch is the indirect programming method reported by Graham, et al. in [149]. A schematic of the basic architecture copied from this paper is shown below in Fig. 4.3, right. Several fabricated designs include variations on this architecture.

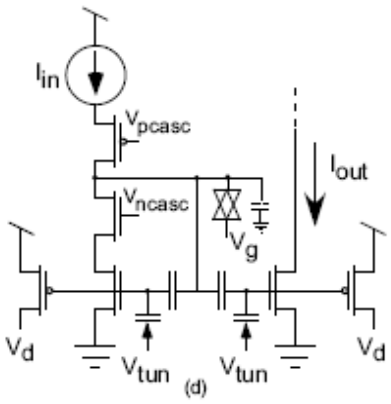


Figure 4.3: Floating gate schematic with differential injection and tunneling nodes. [149].

In one aspect, I have developed circuits that use tunneling only for resetting template values, and prefer carefully balanced hot electron injection to bidirectional updates. In these cases, negative feedback is used to ensure stability. I have also designed architectures which take full advantage of both positive charge increments through controlled tunneling; and negative decrements by continuous hot electron injection. The full details of these architectures are discussed later in this chapter.

b. MULTIPLE INPUT TRANSLINEAR ELEMENTS ("MITES")

Subthreshold MOSFETs display a logarithmic relationship between applied gate-source voltage and drain current, and so may be ideally regarded as translinear elements, which can be used to implement low-power, current mode analog calculations [158]. The basic formulation follows from standard KVL mesh analysis:

$$\text{KVL: } \sum V = 0 \text{ , or } \sum V_{CCW} = \sum V_{CW} \tag{4.5}$$

For the ideal translinear element ("TE"): $$\sum V_{CCW} = \sum \frac{V_T}{\kappa} \log_e \frac{I}{\lambda \cdot I_0} \tag{4.6}$$

assuming for MOSFETs operating in subthreshold that (i) the bodies are tied to common voltage; and (ii) that we operate in saturation with $V_{DS} \geq 4V_T$. Controlling V_{DS} by implementing cascodes is optimal. We can compensate for V_T mismatch by injection at an indirect programming floating gate PMOS node. That should take care of variable component of κ also. Then you have

$$\sum V_{CCW} = \sum \frac{V_T}{\kappa} \log_e \frac{I_{CCW}}{\lambda \cdot I_0} = \sum V_{CW} = \sum \frac{V_T}{\kappa} \log_e \frac{I_{CW}}{\lambda \cdot I_0} \tag{4.7}$$

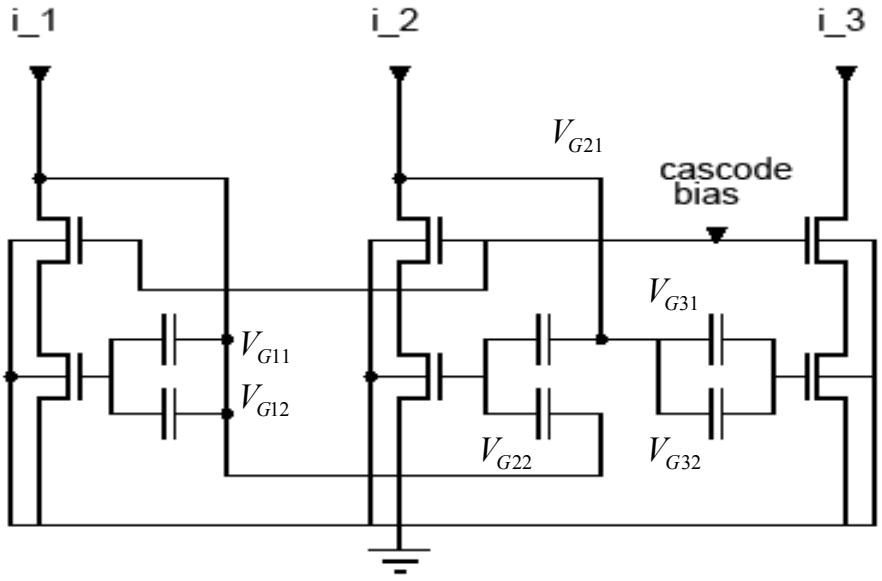
Dropping the common V_T and κ terms, turning the log sum into a product, and exponentiating both sides while assuming a common λ and I_0 yields:

$$\prod I_{CCW} = \prod I_{CW} \tag{4.8}$$

Instead of using multiple transistors in various looped configurations, it is sometimes more convenient to use multiple inputs to the same floating gate to generate MITE structures which perform the same computation more compactly.

In this case, we begin with: $$I_{subthreshold} = \lambda \cdot I_0 e^{(\sum \kappa V_G - V_S)/V_T} \quad (4.9)$$

Taking as a typical example, the subthreshold current squaring circuit used in one of the variance estimation designs, we have:



Fiigure 4.4: Multiple input translinear element ("MITE") current squaring circuit. [158]

Using standard subthreshold analysis, we write:

$$I_3 = \lambda \cdot I_0 e^{V_{G31}/V_T} e^{V_{G32}/V_T} e^{-V_S/V_T} \quad (4.10)$$

$$I_2 = \lambda \cdot I_0 e^{V_{G21}/V_T} e^{V_{G22}/V_T} e^{-V_S/V_T} \quad (4.11)$$

$$I_1 = \lambda \cdot I_0 e^{V_{G11}/V_T} e^{V_{G12}/V_T} e^{-V_S/V_T} \quad (4.12)$$

For the circuits I have implemented, $V_S = 0 \rightarrow e^{-V_S/V_T} = 1$ (4.13)

And, based on the circuit:

$V_{G31}=V_{G32}=V_{21}=$ "V_2" (4.14)

$V_{G22}=V_{G11}=V_{G12}=$ "V_1" (4.15)

Then we have:

$$I_3 = \lambda \cdot I_0 e^{2V_2/V_T} \tag{4.16}$$

$$I_2 = \lambda \cdot I_0 e^{(V_1+V_2)/V_T} \tag{4.17}$$

$$I_1 = \lambda \cdot I_0 e^{2V_1/V_T} \tag{4.18}$$

Solving for V_1 and V_2 in terms of I_1, I_2 and I_3, we obtain:

$$V_2 = \frac{V_T}{2} \log_e(\frac{I_3}{\lambda \cdot I_0}) \tag{4.19}$$

$$V_1 = \frac{V_T}{2} \log_e(\frac{I_1}{\lambda \cdot I_0}) \tag{4.20}$$

$$V_1 + V_2 = V_T \log_e(\frac{I_2}{\lambda \cdot I_0}) \tag{4.21}$$

Assuming identical V_T's, we now equate the three equations and solve for I_1, I_2 and I_3

$$V_1 + V_2 = V_T \log_e(\frac{I_2}{\lambda \cdot I_0}) = \frac{V_T}{2} \log_e(\frac{I_1}{\lambda \cdot I_0}) + \frac{V_T}{2} \log_e(\frac{I_3}{\lambda \cdot I_0}) \tag{4.22}$$

With a little manipulation (not shown), and assuming identical λ and I_0, we find that:

$$I_3 = \frac{I_2^2}{I_1} \tag{4.23}$$

We are now prepared to examine the state of the art in spike sorting systems.

4.2 Spike Sorting Literature Review

Implantable neural prosthetics and portable, precision biosensors must satisfy a host of different bandwidth and power constraints as they detect, classify and decode incoming neural signals, but functionally it is the precision and accuracy of spike sorting that defines the quality of the implant or sensor. Spike sorting is term that broadly encompasses four overlapping and intertwined problems: (a) spike detection; (b) feature extraction; (c) source classification; and (d) decoding of spike trains. The thesis research, like the majority of the hundreds of reported algorithms, methods and means of performing spike sorting, is confined to the first three tasks outlined above. The following literature review addresses the limited subset of the reported spike sorting systems that have informed the present work.

4.2.1 Spike Sorting Methods and Algorithms

Although this research has been directed toward mixed-signal spike sorting architectures for implantable neural prosthetics, much of the inspiration for my designs has come from the following sources:

A. SUMMARIES AND REVIEWS

The seminal review in the field was prepared by Michael Lewicki in 1988 [106]. In *A review of methods for spike sorting: the detection and classification of neural action potentials*, Lewicki introduces the intertwined problems of (a) detecting spikes against background neural noise and (b) classifying the source of overlapping spikes; and then identifies the primary methods for resolving those problems. At a fundamental level, spike detection is simply a thresholding problem, and Lewicki

observes that simple voltage thresholding is a common way of identifying the most prominent feature of a neural action potential: its amplitude. However, simple thresholding is susceptible to classification error owing to overlapping spikes and noise signals, and also suffers a ROC that is a function of the threshold set. Therefore, to the extent that my designs rely on thresholding, those thresholds are set adaptively so as to optimize the detector ROC.

One way to enhance detection accuracy beyond simple thresholding is to incorporate a greater number of features into the classification scheme. Lewicki identifies three ways of classifying spikes based on this notion: (1) principal components analysis ("PCA"); (2) cluster cutting; and (3) template matching.

Principal components analysis involves finding the set of orthogonal basis vectors which represent the greatest variation in the recorded neural signal. Once the basis vectors are identified, action potentials may be classified by convolving them with each of the principal components and assigning a score to each spike according to the convolution. In this case, the pictures shown in Fig. 4.5 provide significant insight:

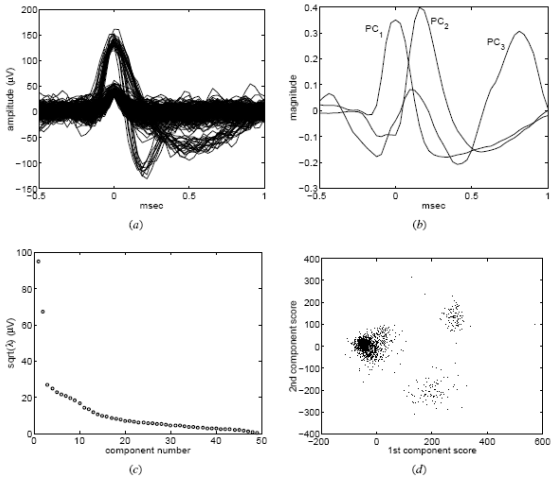


Figure 4.5: Lewicki's PCA analysis [106]: (a) shows raw data; (b) illustrates the first three principal components; (c) provides the standard deviation of the scores for each component; and (d) clusters the component scores.

Since each successive component represents a smaller degree of variation in the data, it is possible to resolve a signal to arbitrary levels of precision by incorporating additional components into the description. For discrete sampled signals, the principal components correspond with the eigenvectors of the covariance matrix of the data. However, while noting that the first three components account for approximately 76% of the variation in the neural data, Lewicki observes that only the first two have latent roots, or scores, above that of the background noise. As such, he proposes using the first two components to classify spikes from this data set. This method can be very powerful, however it is also expensive, and citing a 1982 study by Wheeler and Heetderks, Lewicki comments that this computationally pricey method yielded superior results to simple feature analysis, but were nonetheless "not as accurate as template matching."

Lewicki goes on to address various methods of classifying spikes based on their features or principal components: generally speaking, he describes various algorithms for cluster cutting. The nearest neighbor or k-means algorithm classifies spikes based on their Euclidean distance from one of several cluster means. Bayesian sorting is a more sophisticated method that defines clusters stochastically: a multivariate Gaussian centered about the cluster mean is often used to describe each cluster, and spikes are sorted based on the marginal probability that they belong to a particular class, using Bayes' rule.

$$p(c_k|x, \theta_{1:K}) = \frac{p(x|c_k, \theta_k)p(c_k)}{\sum_k p(x|c_k, \theta_k)p(c_k)}. \tag{4.24}$$

The spike features, such as amplitude and firing frequency used to sort the action potentials into classes are optimized using a maximum likelihood algorithm to define bounded clusters for the feature set. Fig. 4.6 illustrates the method:

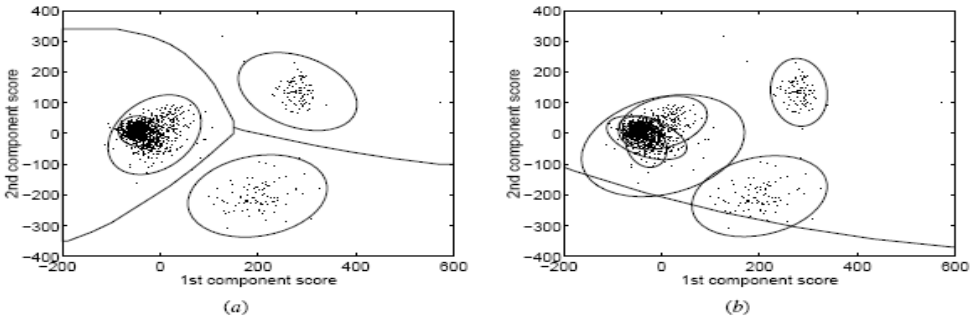


Figure 4.6: Gaussian clustering by Lewicki [106]; (a) shows the Bayesian decision boundaries for the four clusters; (b) shows the same data with ninc clusters.

After surveying a number of related spike sorting problems and advertising his own method for overlap decomposition, Lewicki refers back to the Wheeler and Heetderks study that found "template-matching methods yielded the best classification accuracy compared to spike-shape features, principal components, and optimal filters." Noting that non-adaptive template-based methods suffer during bursting neural sequences and as a result of electrode drift, Lewicki concludes that perhaps the best hope for separation of overlapping spikes is the tetrode, which enables superior source localization.

A subsequent comparison of spike classification techniques in detecting tonic and phasic action potentials generated by caterpillar taste organs was set forth in [123]. In that review, the authors showed that classical template matching and principal components analysis ("PCA") techniques performed nearly perfectly in

detecting tonic spikes generated by various stimuli. However, the neural network classifier that the authors studied performed significantly better at distinguishing phasic, or transient, responses than either template matching or PCA. Although these algorithms were employed offline, and thus do not directly inform the investigation into analog and mixed-signal spike sorting circuits, they do suggest that alternatives to the classic template matching scheme ought to be examined. That is precisely what I have done.

B. TEMPLATE MATCHING

There is an appreciable dearth of recent template matching algorithms for spike sorting, presumably owing to the maturity of the basic algorithms, and the computational complexity of such a system. The following sources address recent advances in the field:

In [127], *A New Template Matching Method using Variance Estimation for Spike Sorting*, Cho et al. describe a novel algorithm for performing template matching by estimating the variance of each purported class of spike shapes and using that variance to define the distribution for that class of spikes (*See*, *e.g.* [131]). Detection is then accomplished by a weighted Euclidean distance metric. This is one way in which training may be accomplished using template matching architectures, by employing the variance estimation circuitry across a band of tuned template matching filter banks. Using my novel variance estimation circuit to tune a Gaussian formed by, e.g. a bump circuit, we can translate this algorithm into hardware – a result which can be applied much more broadly than for the purpose of spike sorting.

Another offline system of note was reported in [86], in which the authors develop a multistage template-matching algorithm for spike detection and classification. In particular, templates are evaluated for fit against ostensible "spike events"; the best fit is the one which results in the minimum residue variance according to the chi-squared test. And preceding that paper, in [141], Zouridakis and Tam developed a method for fuzzy clustering to generate templates that were nearly identical to the signals from which they were learned. Although somewhat difficult to implement in mixed-signal hardware, the principles underlying this algorithm inspired some of the adaptive template learning algorithms developed below. And the earliest reported real-time template matching system used an eight-point matched filter system and was implemented primarily in software, [138], [139].

C. WAVELET AND MULTIRESOLUTION ANALYSIS

Wavelet decomposition and multiresolution analysis methods gained favor in the late 1990's, and since that time, the signal processing community has virtually exploded with wavelet sorting algorithms. The theory is fairly straightforward: a few signal subbands or discrete coefficients characterize the basic action potential and may be classified accordingly; detection accuracy is improved inasmuch as noise is implicitly discounted. Many of the algorithms proposed employ an arbitrary resolution discrete wavelet transform; however, the continuous time analog wavelet transform possesses the distinct advantage of not requiring storage for discrete coefficients. The references that follow represent the state of the art wavelet sorting algorithms:

In the most recent publication on point, [112], Robert Brychta and his colleagues evaluate *Wavelet Methods for Spike Detection in Mouse Renal Sympathetic Nerve Activity*. Brychta, *et al*. compare the performance of the discrete and stationary wavelet transforms against one another at detecting the spikes which comprise mouse renal sympathetic nerve activity. Noting that conventional algorithms either integrate spikes indiscriminately or rely upon static and error-prone thresholds to detect action potentials, the authors proposed to employ and contrast two wavelet schemes for decomposing a neural signal: the DWT and the SWT. In evaluating their data, the authors conclude that, in spite of its computational complexity the SWT is a more reliable alternative to either the DWT or simple amplitude thresholding for denoising and spike detection and discrimination.

A couple of years earlier, in [150] Zoran Nenadic and Joel Burdick had proposed an analogous method for *Spike detection using the continuous wavelet transform*. Rather than simply decomposed thresholding the spikes, the Nenadic and Burdick used Bayesian hypothesis testing at different scales to find and classify spikes. According to the authors, their ROCs are better and their false positives closer to real neural spikes than those identified by previous methods of spike sorting. The authors extol the virtues of template matching techniques, but complain that template generation requires supervision – the reports cited above cast some doubt on this position. Regardless, the authors of [150] demonstrate (through simulation) the theoretical power of wavelets in discriminating between signal and noise.

Owing to this ability to reliably separate relevant signal components from noise, a litany of authors have employed the continuous and discrete-time wavelet

transforms: as pre-processors, [111], for unsupervised spike detection and sorting, [109], [110] to characterize action potential features for template generation, [87], to detect overlapping extracellular neural signals with a low SNR, [101], [104], [134], [142]. These are but the most recent in a surfeit of wavelet algorithms to perform or assist in spike sorting. Likewise, multiresolution non-linear time-frequency classifiers have been proposed – see, e.g., [115]. Despite this flush of algorithms and methods, there are, however, less than a handful of custom wavelet transform circuits.

D. NEURAL NETWORK CLASSIFIERS

Many alternatives to the wavelet technique exist. One promising method is reported by Kyung Kim and Sung Kim, who demonstrated *Neural Spike Sorting Under Nearly 0-dB Signal-to-Noise Ratio Using Nonlinear Energy Operator and Artificial Neural-Network Classifier,* [82]. The authors achieved classification of extracellular spikes from Aplysia abdominal ganglia at better than 90% with an SNR as low as 1.2 (0.8 dB). They critique thresholding, Haar-basis wavelet analysis and matched filters as poor means of detecting spikes against an SNR that approaches the zero level. They further observe that such methods can require *a priori* knowledge of the waveform and background noise which is often not available during the training phase. They assert that neural networks can achieve superior blind performance but may be computationally costly, and to the extent that they rely on, *e.g.*, thresholding during training, that such methods are susceptible to the same shortcomings as the others. By contrast, the authors' method trains an NN using a non-linear energy

operator (NEO) proportional to the instantaneous product of frequency and amplitude for both training and detection. Specifically, they relied on supervised classifiers comprised of multilayer perceptrons (MLPs) and radial basis function networks (RBFNs) (*See* [114]). The authors conceded that adaptation based on learning the actual distribution function through repeated training could yield better results. However, while the authors set their threshold level manually, I have designed circuits to do it adaptively – an NEO operator is used to detect neural spikes and compared against amplitude thresholding.

Classical neural network ("NN") detectors have been applied to multiunit spike detection and sorting for well over a decade. Such classifiers were particularly in vogue during the mid-to late 1990's when a bevy of authors reported enhanced ROCs using real-time digital NNs [133], [136], [137], [146]. Difficult problems in resolving superposed spikes from one another were handled significantly more reliably by trained NNs than by simple matched filters, [146]. It is on the basis of this and other groundbreaking work in mapping prototypical frequency transforms onto a NN architecture by Professor Martin Peckerar and colleagues [135], that I have modeled a NN classifier to compete with the adaptive floating gate template matching and wavelet transform circuits.

E. AUTOMATED AND UNSUPERVISED

Among the several hundred other recent spike sorting papers, those following authors' work have had the most significant impact on this work. Wood and Black at

Brown University have co-authored several papers in the field, [100], [103], [153] beginning with an illuminating article on the variability of manual spike sorting. In [100], they report that even among purported experts, false positive and false negative spike detections exceeded 25% on average for a synthetic signal; and that the variability in separating spikes from noise and other action potentials was even greater when real data was evaluated. The authors concluded that this data pointed to the need for an automated sorting method that would yield more consistent results, [100], and their work corroborates the conclusion by Shamma in [108], that it is important to focus on reliable or repeating spikes, rather than errant or transient firings. Wood and Black developed two methods for automatic spike sorting, in [153] and [103], the latter assuming an infinite mixture model ("IMM") of possible spike classes and partitioning this space according to a Bayesian computation of maximum a posteriori ("MAP") probabilities. The authors correctly postulate that their offline method may be extended to an online classifier that estimates the posterior probabilities sequentially; the floating gate filter bank, the wavelet decomposition circuit and the neural network classifiers can adapt to these probabilities as they are trained to operate on real data.

A number of authors have also reported sophisticated techniques for unsupervised sorting: projection pursuit based on negentropy maximization, [88], [98]; Markov chain Monte Carlo classification based on spike timing and generation dynamics as well as amplitude features, [79]; a purportedly information theoretic maximum likelihood algorithm, [127]; a principal components analysis ("PCA") based method with automatic overlap decomposition employing an iterative

algorithm to optimally cluster spikes according to a spectral cost function, [90]; and clustering spikes and assessing the quality of classification based on noise characteristics [107].

In addition many authors continue to employ more conventional methods of performing unsupervised sorting, including: PCA plus self organizing feature map neural network, [89]; spike-sorting with a multivariate t-distribution expectation maximization algorithm, [144]; modeling spike waveforms by ordinary differential equations with perturbations and characterizing their phase space features to classify them [147]; using support vector machines, [92]; resorting to tetrodes to differentially localize the source of independent action potentials, [95]; software-based feature extraction and template learning plus detection and classification, [140].

F. NEURAL PROSTHETICS

Neural prosthetics are remarkably advanced, yet the field remains in its infancy. The most recent work detailed in [5], [6], [116], and [79],[80] represent the state of the art with respect to truly implantable prosthetics, yet each of these systems does little more than record and identify peaks against the noise. In [96], [117], [121] and [125] and a host of other publications, Nicolelis and his colleagues report on a remarkable motor prosthetic they have built which is driven by a computer that records from populations of neurons via a chronically implanted microelectrode array. While incredible progress has been made on this front, the hardware responsible for detection, classification and decoding of neural signals remains off-

chip, so that the "implant" is essentially just a bundle of fine wires. For a glimpse into some of the earlier contributions to the field, and in particular population coding, see [118], [119], [120] and [122].

4.2.2 Spike Sorting Circuits

Despite advancements in the field, it remains impossible to measure individual neural electrical signals non-invasively. Thus, neural prosthetics that would seek to ascertain the state of the system demand a direct brain-machine interface; it is necessary to record directly from, or in close proximity to the nerve cells of interest. Further, in order to conserve the signal strength of the extracellular potentials measured, which are typically on the order of 50 μV peak-to-peak, and to mitigate against noise corruption along transmission lines, the required recording electrodes should be connected as closely as possible with the hardware that will sort the incident spikes. To maximize SNR and mitigate against external interference, implantable signal processing architectures are therefore preferred. This section reviews reported and proposed implantable integrated architectures which primarily serve as preamplification, conditioning and detection stages for spike sorting. A table summarizing the various architectures is presented at the end of this section.

In 2003, Reid Harrison, who despite his young years we will call the grandfather of this field, reported *A Low-Power Integrated Circuit for Adaptive Detection of Action Potentials in Noisy Signals* [102]. In this paper, he observed that it would not be possible within power operating constraints to directly transmit the

raw electrophysiological data from an implanted multi-electrode array transcutaneously to a processing unit outside the brain. Consequently, he proposed to encode the measured signals using an implicit AER scheme to represent the measured action potentials so that a low-power RF transceiver could serially send the addresses of firing neurons off-chip. Harrison implemented his novel encoding architecture using the following circuit:

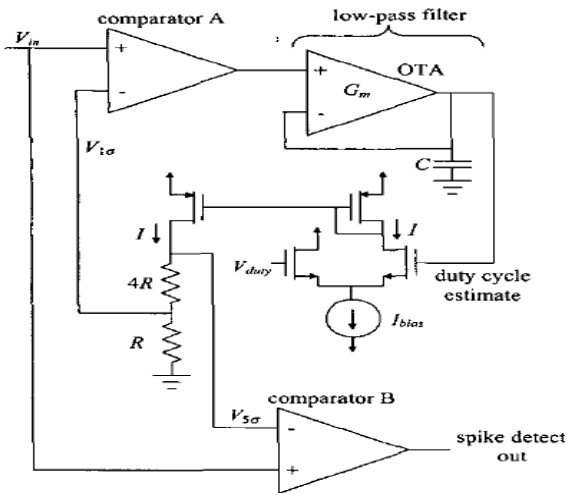


Figure 4.7: Harrison's adaptive threshold detection circuit. [102]

Harrison's architecture operates by adaptively tuning the detection threshold to some multiple of the standard deviation of the noise (approximately 5σ). To compute σ, the circuit relies on the output of the integrating low-pass filter to estimate the duty cycle, or fraction of time the noise component of the signal will spend above the threshold potential. Since Gaussian noise will result in a theoretical duty cycle of approximately 0.159 with the threshold set at σ, σ can be found by forcing the estimated and theoretical duty cycles to match. Using a simple resistive divider, Harrison ensures the reliability of his peak detector by setting the threshold to $5\sigma^{+}$.

Harrison's circuit is clever, but has several drawbacks: (a) first, while it does well to discriminate peaks against Gaussian background noise, it is still possible for transients or other artifacts to give rise to false positives – incorporating other spike features can offer more accurate and precise detection with only incremental costs; (b) second, the adaptation relies heavily on the performance of the Gm-C filter to give a reliable estimate of the duty cycle, on noise theory to ensure that it matches the actual noise encountered, on the differential amplifier to accurately compare the two, and on IC resistors to set the threshold – the precision of each of these elements is severely constrained by device and process mismatch, and also by limited computational precision in the operating regime – acting in concert, these effects result in either an underestimate of the noise and a surfeit of false positive detections, or an overly conservative estimate of the noise floor, and missed spikes; (c) third, the comparators used are comprised of latches which must be clocked, resulting in additional circuitry and sources of error; (d) fourth, the architecture consumes (admittedly in a 1.5 μm process with 5 V supply) almost 60 μW of power in operation. Multiply that by 100 to perform parallel encoding of 100 channels, and nearly 6 mW of power per 10 square mm is consumed – this is an entirely unacceptable level for implantable devices; (e) Fifth, the architecture only allows for the detection of spikes – it would be confounded by the classification problem – i.e., how to resolve the identity of the neurons responsible for concurrent spikes.

The following year, in [93], Horiuchi and Abshire published: *A Low-Power CMOS Neural Amplifier with Amplitude Measurements for Spike Sorting*. The reported architecture is an evolution of Harrison's earlier low-power, low-noise

amplifier for neural recording. Working in conjunction with their students, Horiuchi and Abshire added peak, trough and level detection circuits to the output of the Harrison amplifier, thus enabling the authors to discriminate between classes of firing neurons based on spike amplitude features. Although the ROC for the detector are not reported, the authors did confirm low-power amplifier operation (<1uW) and demonstrated proof-of-concept feature extraction on chip. The major drawbacks to this particular architecture are related to its performance during bursting sequences of potentials, and at low SNR where many conventional thresholding methods fail.

Later that same year, Rogers and Harris, published *A Low-Power Analog Spike Detector for Extracellular Neural Recordings* [91], tackling the same bandwidth bottleneck noted by Harrison. These authors implemented a more sophisticated "onset detection" scheme for detecting neural spikes.

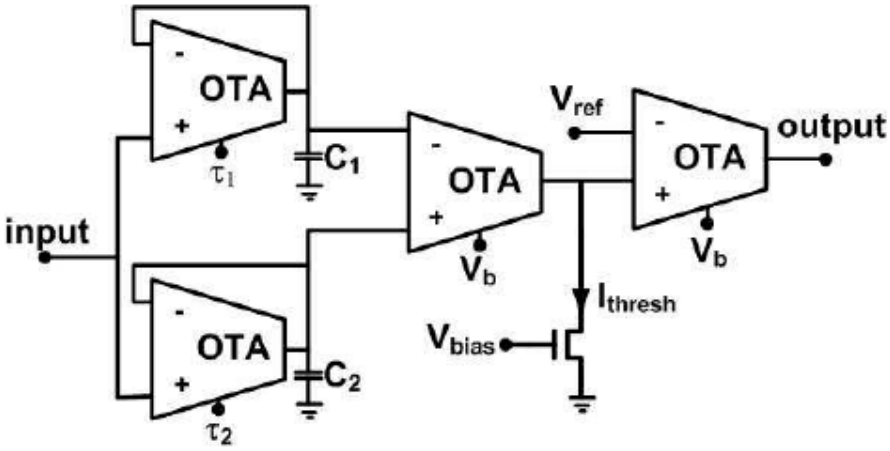


Figure 4.8: Frequency selective onset detection architecture. [91]

The two OTAs which measure the input signal are tuned to different frequencies: the high pass filter is used to remove high frequency noise from the signal of interest and the low pass filter is used to estimate generate a local average

background signal. The outputs of those two are compared first against one another, and then against a fixed threshold to generate a spike output. Notably the capacitance values for C_1 and C_2 were on the order of 10 pF. By operating their circuit in the subthreshold regime, the authors were able to achieve total circuit power dissipation of under 1 uW.

Although the circuit is more robust in some ways than a simple threshold detector, it has several significant drawbacks: (a) first, both the cutoff frequencies for the input low-pass filters and the thresholds themselves are set off chip, according to the authors to account for mismatch, etc. – as a result this circuit is not suitable for unsupervised or implantable spike sorting; (b) second, although the MATLAB simulations (and subsequent data) yield a nearly perfect ROC, the precision and hence accuracy with which the filters can actually detect spikes is largely a function of the quality and precision of the filters, which can vary dramatically as a result of process and mismatch variation – without adaptive biasing, these circuits cannot accomplish their intended purpose; (c) third, MATLAB simulations do not provide compelling evidence that the circuit is robust against a near 0 dB SNR, suggesting other features might be considered; (d) fourth, and finally, the circuit shares the deficit of its predecessor – namely, it cannot classify spikes.

Taking a more complex theoretical approach to the problem, Y. Suhail and K.G. Oweiss, published *A Reduced Complexity Integer Lifting Wavelet-Based Module for Real-Time Processing in Implantable Neural Interface Devices*, that same year. They introduced a "design methodology for computing the DWT with the WL scheme for arbitrary number of channels." In order to conserve memory and

bandwidth, the authors proposed to use the integer wavelet transform (IWT) on signal data sampled with 10 bit precision, and to quantize the filter coefficients. They chose the DWT because the discrete representation is computationally efficient, and because the wavelet transform tends to encode action potentials in a few large principal components and the noise in many smaller ones, thus performing an implicit thresholding operation in the process. Although the authors did not implement the architecture they proposed, they did demonstrate theoretical accuracy comparable to that of conventional digital signal processing hardware.

In principle, this architecture performed admirably and offered cost savings versus conventional DSPs. However, the drawbacks are a matter of conjecture inasmuch as neither an architecture, nor power consumption and footprint for the DWT are disclosed. A compact, precise, analog variant on the wavelet decomposition algorithm could surely be made smaller and with significantly reduced power requirements, though computational precision remains an issue for existing analog implementations.

In a parallel attempt to shoehorn Pentiums into the low-power regime, in [84], Zachary Zumsteg and his colleagues published a study on the *Power Feasibility of Implantable Digital Spike-Sorting Circuits for Neural Prosthetic Systems*. Taking the standard heat dissipation for a 0.13 μm digital system at 1uW / GOPS (Giga-operations per second), the authors estimated the computational cost of each of the steps involved in two spike sorting algorithms, including training and classification, concluding that a 5mm x 5mm chip would consume less than 3mW/cm^2 power, well below the 80mW/cm^2 thought to cause tissue damage. Their study presents a

compelling argument that we should not completely discount digital spike sorting systems, however, the authors fail to account for a number of factors which could inflate their consumption estimates. First, they purportedly perform the desired operations in the digital domain, but do not appear to account for the cost of A-D sampling and conversion – neural signals may be comprised of trains of all or nothing spikes, but their waveforms are decidedly not digital. Second, to the extent that their A-D conversion is in fact a simple thresholding operation, the authors do not report the accuracy of this stage, nor in fact do they discuss the overall performance of their proposed architectures against a specified set of SNRs. Also neglected is the practical degree to which a custom ASIC can replicate the performance, i.e. power efficiency, of a Pentium / Athlon / SPARC chip. Finally, the authors do not account for the heat dissipation of the onboard clock required to drive the digital circuitry. For these reasons, it is believed that DSPs still suffer competitively versus analog spike sorting architectures.

The following year, Rogers and Harris teamed up with Principe and Sanchez and published: *An Analog VLSI Implementation of A Multi-Scale Spike Detection Algorithm for Extracellular Neural Recordings* [152]. This circuit shares many of the same advantages (and deficits) of its predecessor, but incorporates a novel detection architecture that implements an analog wavelet transform. The circuit operates as a cascade of low pass filters (LPF) that are biased by sequential voltage taps taken from the resistive line. By tuning V_{high}, all of the other bias voltages are set automatically, and when operating the LPFs in the subthreshold regime, the linear drop across the resistors corresponds with an exponential decrease in bias current and hence filter

cutoff frequency. As a result, the filters exhibit a constant Q (center frequency / spectrum width = constant) and the circuit performs an effective wavelet decomposition of the incoming signal.

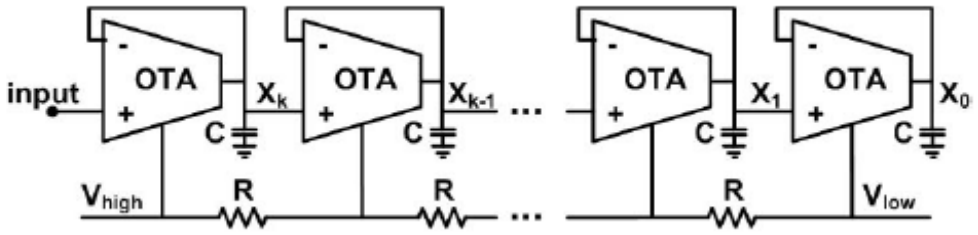


Figure 4.9: Gm-C filter bank for implementing analog wavelet transform. [152]

The transfer function for each stage is given by $H_k(s) = \frac{1}{a^k \tau \cdot s + 1}$, where a^k is a preset factor. In this scheme, spikes are detected by thresholding the difference at neighboring taps. As with their previous designs, these thresholds are set manually.

This chip demonstrates the remarkable efficiency that analog computing can bring to bear, but also serves to illustrate the difficulty in matching digital precision with existing analog architectures. More specifically, positive results for the legacy onset detector are admittedly sub-optimal: although they showed 99% detection accuracy when verified using an artificial signal, the authors demonstrated that their competing multi-scale detection scheme performed almost 8 times better at eliminating false positives when tested in MATLAB simulations using real neural data. Moreover, with respect to the analog wavelet multi-scale design, it is telling that although the authors tested the new architecture on real neural data, they never reported those results. Although it is only possible to speculate as to why, the natural suspicion is that the circuit could not reliably reproduce the theoretical results – this is

often the case with analog filter banks; one possible culprit in this case is mismatch and process variation leading to imprecise cutoff frequencies and thresholds. Using floating gate architectures to compensate for mismatch and adaptive biasing schemes to set thresholds permits considerably greater precision and robustness against noise.

During that same year, Alex Zviagintsev, Yevgeny Perelman, and Ran Ginosar, published a study, *Low-Power Architectures for Spike Sorting*, in [143]. They tested three different classification algorithms on real neural data, observing the computational complexity and error rate for each. Among the three, only PCA performed error-free but the integral transform (IT) performed with only 2.2% errors, and offered a simple hardware implementation. In, [83] Zachary S. Zumsteg and his colleagues perform a much more robust argument for the, *Power Feasibility of Implantable Digital Spike Sorting Circuits for Neural Prosthetic Systems*. Resurrecting their previous arguments, they provide a rebuttal to the primary deficiency in their earlier analysis: namely, they account for the power consumed by A/D conversion, pegging it approximately 100uW for 100 channels of 8 bit 30kHz conversion. Notwithstanding the fact that this estimate is an order of magnitude lower the design they cite to substantiate it, the authors reveal in this paper another flaw suggested by their earlier analysis. That is, for the resolution they propose, the suggested architectures cannot reliably detect or classify neural spikes below an SNR of approximately 7. Simply speaking, the quantization error introduced by digitizing the data loses the signal below a certain threshold; a threshold which precision analog implementations do not share.

In, [105], Perelman and Ginosar presented, *An Integrated System for Multichannel Neuronal Recording With Spike/LPF Separation, Integrated A/D Conversion and Threshold Detection*. This mixed-signal architecture may hail the onset of a new generation of processors, but at present it is not characterized sufficiently to be able to critique its performance. Power consumption is identified as 12 mA [sic] for an approximately 4 mm x 4 mm chip, suggesting with a 1.5 V supply, 18 mW / 16 mm^2 are consumed – far above the 80mW/cm^2 that fries brain cells. Nor are ROC statistics presented. Overall, this chip reflects the deficiencies that digital systems remain unable to overcome. They cost too much in terms of power, and they sacrifice too much resolution to be able to approach the efficiencies of analog design, hence they are less accurate.

For example, two teams of west coast researchers, led by Shenoy and Meng at Stanford, and Fetz at the University of Washington have recently developed, in parallel, digital "implantable" microchips for neural recording and stimulus [5], [6]. The Stanford authors have developed and reported an implantable digital architecture for long-term (several days) continuous recording of neural signals, called HermesB. Dr. Fetz's team has developed a similar circuit, the Neurochip, for chronic neural recording and stimulation. While these architectures both represent remarkable advances in the field, their power consumption renders them unsuitable for simultaneous recording across an array of electrodes. In particular, HermesB draws hundreds of milliwatts of power while sampling two channels, and the Neurochip dissipates a comparable 40-120 mW when recording from a single channel. Likewise, the frequency bandwidth of these systems is limited even at maximum power, so that

typical problems of electrode drift, ambient noise and superposed spikes that significantly attenuate the SNR of recorded signals pose greater problems for these low-resolution systems than for higher precision offline or continuous analog designs. In addition, although these systems are both situated supra-cranially, it is likely that chronic exposure to the relatively high power dissipation of these prototype architectures will result in tissue damage [129].

Likewise, although reported analog and mixed-signal systems are capable of remarkable feats of computational efficiency, they often lack the precision to perform unsupervised detection or classification with confidence. So while it may be possible to: measure the signal energy of a local field potential using mere nW of power, [81]; place as many as one hundred recording channels on a chip without climbing above the power ceiling, [80]; or threshold spikes with 5 bit precision over 32 channels while dissipating less 200 uW per channel, [79]; each of these systems operates according to fixed user-defined thresholds. And none of these architectures purports to classify spikes. Indeed, computational imprecision resulting from device mismatch and systems noise can cripple the performance of such low-power detection and classification circuits. Adaptive architectures for thresholding and event detection have been reported as noted above, but the circuits that perform the adaptation are themselves susceptible to mismatch and imprecision.

Finally, efforts reported in [93], [113], demonstrate that it is possible to algorithmically discriminate offline between classes of firing blowfly neurons (H1 and HS) based on data generated by real-time spike peak and trough feature extraction circuits; but no classifier architectures have been reported. In this work, I

have attempted to overcome these difficulties by using adaptive integrated circuit architectures to compensate for mismatch and adaptive biasing schemes to set thresholds in order to realize considerably greater precision and robustness against noise.

Table 4.1 provides a list of the spike sorting circuits I have reviewed along with relevant power dissipation information. Power density was computed by normalizing reported power dissipation for a given footprint to an area of one cm^2. Red shading indicates reported power density above maximum safe limits. Orange shading indicates reported power density approaching danger zone for chronic implantation (>10 mW/cm^2). Green shading indicates reported power density in acceptable range for implantation.

In sum, there are several reasons why the spike sorting research I performed matters. First, many state-of-the-art architectures either exceed or approach maximum permissible chronic exposure levels; they do not meet power-density requirements for implantable prosthetics, particularly the digital systems. Second, systems that operate on an acceptable power budget often suffer from circuit mismatch, computational imprecision and fixed, user-defined thresholds to identify spikes. Third, no reported system attempts to classify neural spikes on-chip.[*,**]

* In [155], the authors report that their circuit performs clustering of extracted features, but no clustering circuits are shown – only feature extracting circuits.

** In [157], the authors report a remarkably power efficient PCA circuit, but only simulation results are provided.

Table 4.1 Spike Sorting Circuits

Reference	Type	Detection Algorithm	Classification Algorithm	Power Density
Harrison, 2003, [102].	Analog	Adaptive peak thresholding	N/A	64 mW/cm^2
Harrison, Shenoy, 2004, [81].	Analog	Low pass filter / squaring circuit	N/A	11 µW/cm^2
Horiuchi, Abshire, 2004, 2007, [93], [113].	Analog	Peak / trough thresholding	Not reported	1.1 mW/cm^2
Rogers, Harris, 2004, [91], Principe, Sanchez, 2005, [152].	Analog	Frequency decomposition / wavelet circuit	N/A	1.8 mW/cm^2
Diorio, Fetz, 2005, [4].	Mixed-signal	Analog filters / A-D conversion / thresholding	N/A	> 40 mW/channel
Perelman, Ginosar, 2007, [105].	Mixed-signal	Low pass filter / A-D conversion / thresholding	N/A	113 mW/cm^2
Harrison, 2007, [80].	Mixed-signal	A-D conversion / thresholding / 100 channels	N/A	49 mW/cm^2
Wise, Najafi, 2007, [79].	Mixed-signal	Peak / trough thresholding / 32-64 channels	N/A	197 µW/channel
Shenoy, Meng 2007, [5].	Mixed-signal	Analog filters / A-D conversion / thresholding	N/A	~100 mW/channel
Harris, et al., 2008, [154].	Mixed-signal	Pulse-based feature extraction	N/A	~100 uW/channel
Borghi, et al., 2008 [155].	Mixed-signal	Peak / trough / thresholding	N/A*	30 mW/cm^2
Chae, et al., 2008, [156].	Digital	Max-min feature extractors	N/A	600 mW/cm^2
Chen, et al., 2008, [157].	Digital	90 nm PCA	N/A**	210 mW/cm^2

4.3 Floating Gate Template Matching

In order to develop novel, high accuracy and precision circuits that are both power and size compatible with biological systems for implantation and biosensing, I have investigated mixed-signal VLSI architectures for spike sorting across a broad feature space. Following the suggestion in the literature that template matching algorithms provided superior detection and classification accuracy [106], I designed a novel programmable template matching system for implantable spike sorting. The circuit components I have implemented operate at very low power, are programmable and adaptive, and include new stochastic signal processing architectures and trimming circuits to improve accuracy.

In following subsections, we expound upon the theoretical foundations for the designs, including functional rationales, and show simulated and measured performance data for the spike sorting circuits. In so doing, we expose the express and implicit tradeoffs made in balancing among the following: dynamic range and SNR, accuracy and precision, power efficiency and computational complexity, speed, and size or footprint of these architectures. We also further discuss the novel on-chip real-time statistical signal processing architectures that I have developed for improving the feedback control and hence performance of the sensing and classification architectures.

A system level schematic is shown in Fig. 4.10. Without loss of generality, it has three main components: (1) a template matching filter bank; (2) a variance estimation circuit; and (3) a simple classification block.

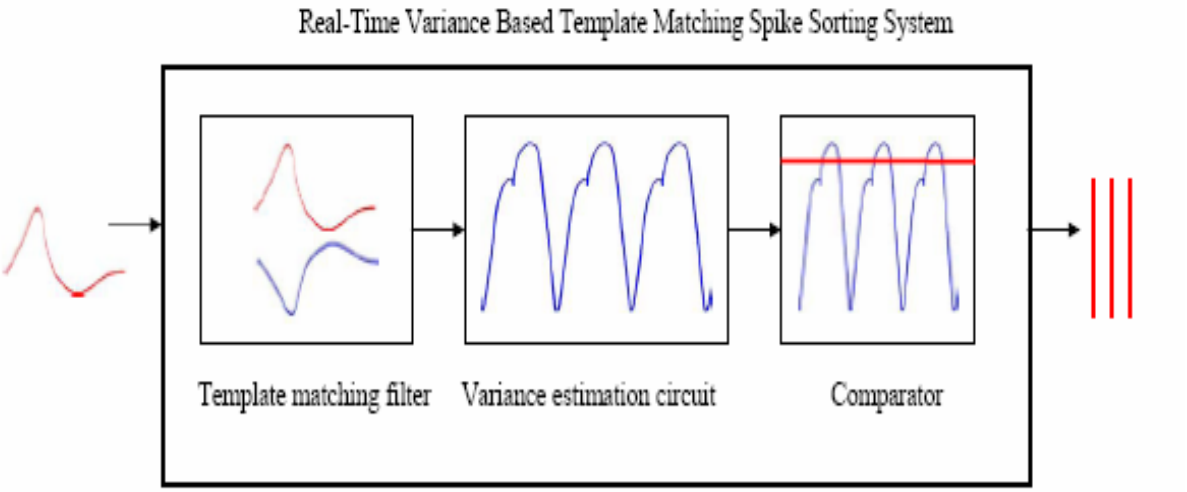
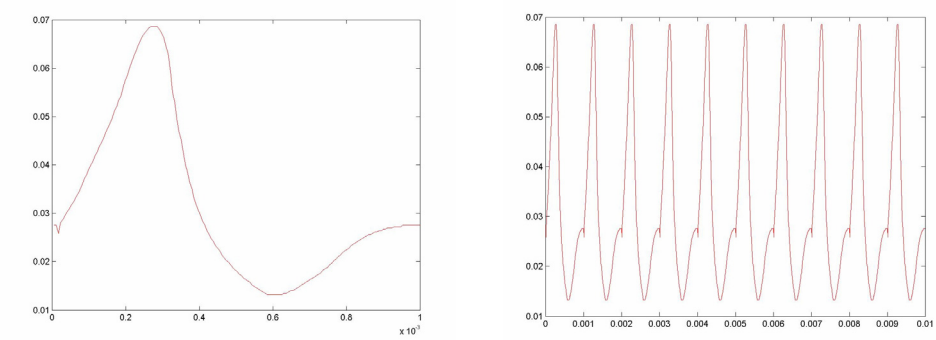


Figure 4.10: System-level diagram of the floating gate template matching spike sorting system.

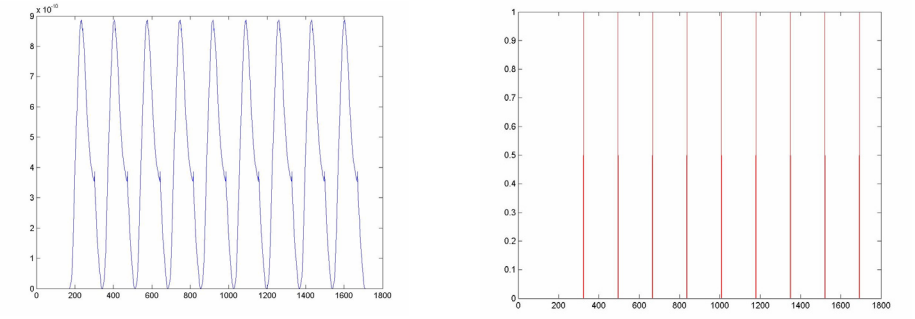
MATLAB simulations performed under ideal conditions illustrate the operation of the system in principle. An artificial spike waveform was digitally extracted from an image, and a spike train was generated:



(a) (b)

Figure 4.11: (a) left: Schematic drawing of neural action potential; (b) right: spike train consisting of cascade of simulated action potentials.

When the artificially generated spike train above (right) is virtually passed through the template matching architecture, we see the following (a) variance; and (b) negative thresholded, or spike detection waveforms:



(a) (b)

Figure 4.12: (a) left: MATLAB simulated output of the floating gate template matching variance estimator; (b) right: pulse-train where each event corresponds with a template match.

To enable the system to perform unsupervised learning of new templates, a pre-sorting component, the spike detector, has been developed.

4.3.1 Detecting Neural Action Potentials

The recorded neural data processed here was obtained from Dr. Ping-Bo Yin and Professor Shihab Shamma of the Neural Systems Lab at the University of Maryland, College Park. A schematic of the experimental setup, coupled with a segment of recorded neural activity from one channel of an electrode array implanted into the auditory cortex of a ferret are shown in Fig. 4.13. This data reflects amplified, filtered single-channel recordings over a period of 30s in response to auditory stimulation.

As Yang and Shamma observed in [108], "the overriding goal of the spike detection algorithm to be used with multielectrode arrays is *not* so much to detect the smallest spikes in the midst of noisy traces, but rather to isolate the most reliable spikes with no or minimal human intervention."

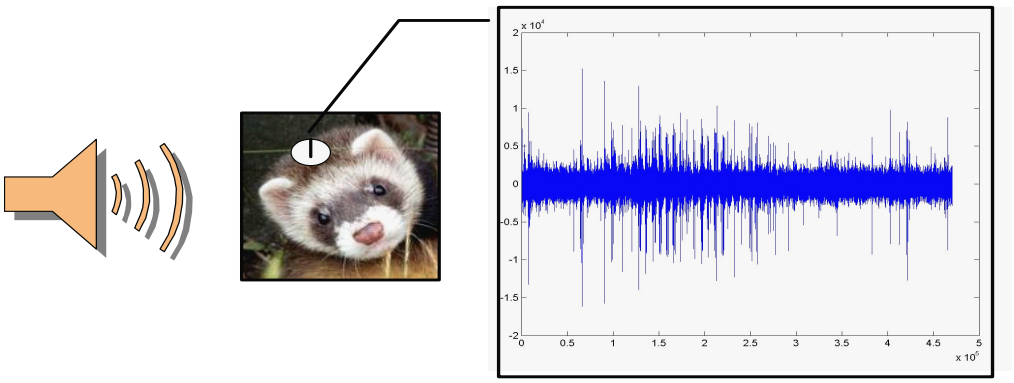


Figure 4.13: Neural signals obtained from the NSL and schematic representation of recording apparatus.

This applies equally to implantable neural prosthetics – we want to extract the minimum relevant information from the vast array of incident neural data; this means identifying reliable spikes without operator supervision. Extracting reliable spikes, in turn, means reliable classification, and thus more robust implantable architectures. Furthermore, because there is greater confidence in assigning meaning to a population code when the population is large, we seek to increase recording and stimulation channel density to the extent practically possible.

For my own design, I was guided by these principles and inspired by Kim and Kim's, algorithm for the detection and classification of extracellular spikes using a non-linear energy operator ("NEO") to exaggerate spikes against the background noise [82]. They used an NEO proportional to the instantaneous product of frequency and amplitude of an incoming signal to demonstrate theoretical detection and classification at better than 90% at < 1 dB SNR using this system. To accomplish analogous results, I first tried the fabricated variance estimation circuit as a spike detector – since the variance estimator output is the difference of squares,

$$\sigma^2 = \mathrm{E}\,[(\mathrm{X} - \mathrm{E}\,[\mathrm{X}])^2] = \mathrm{E}\,[\mathrm{X}^2] - \mathrm{E}^2[\mathrm{X}] \tag{4.25}$$

it too serves as an NEO that tends to resolve spikes from background noise. Fig 4.14 (a) represents a noisy idealized spike train generated as follows: forty-element spike templates were generated by extracting the average waveforms of two previously identified classes of spikes found in a 20s ferret cortical recording; the two waveforms were normalized to a 150 mV peak-to-peak scale,[1] and then inverted about the voltage axis to serve as templates; by concatenating a sequence of one or both template classes together, at normally distributed inter-spike intervals and adding Gaussian noise using the *randn* function, the composite trains were completed. The SNR was computed as follows:

$$SNR = 20\log\left(\frac{A_{signal}}{A_{noise}}\right) \tag{4.26}$$

where A_{signal} is the computed root-mean-square amplitude of all of the elements of a given spike train, and A_{noise} is the standard deviation of the normally distributed noise amplitude. Fig. 4.14 (b) shows the outputs of the first generation variance estimation circuit with physical inputs generated by a MATLAB controlled DAQ card; the physical inputs represent the convolution of a virtual template matching filter with a temporally rescaled version of the noisy spike train. Detection of neural events is accomplished in one of two ways, either: (a) by a template match, corresponding with

[1] Extracellular recordings from real neural signals vary from tens of microvolts to tens of millivolts depending on the nature of the recording electrode and its proximity to the neurons, as well as surrounding potentials, so a 150 mV signal has been pre-amplified. Microwatt preamplifiers are well known and characterized. [57]

a local variance minimum, as an incoming neural signal is convolved with one or more matched filters in parallel; or (b) by employing the variance estimator as an energy operator whose maximum output corresponds with a neural spike. Local minima and maxima may be identified using a low-power current comparator or peak detector. In Fig 4.14 (b), both user-defined template matching (solid line) and instant energy (dashed line) thresholds are marked.

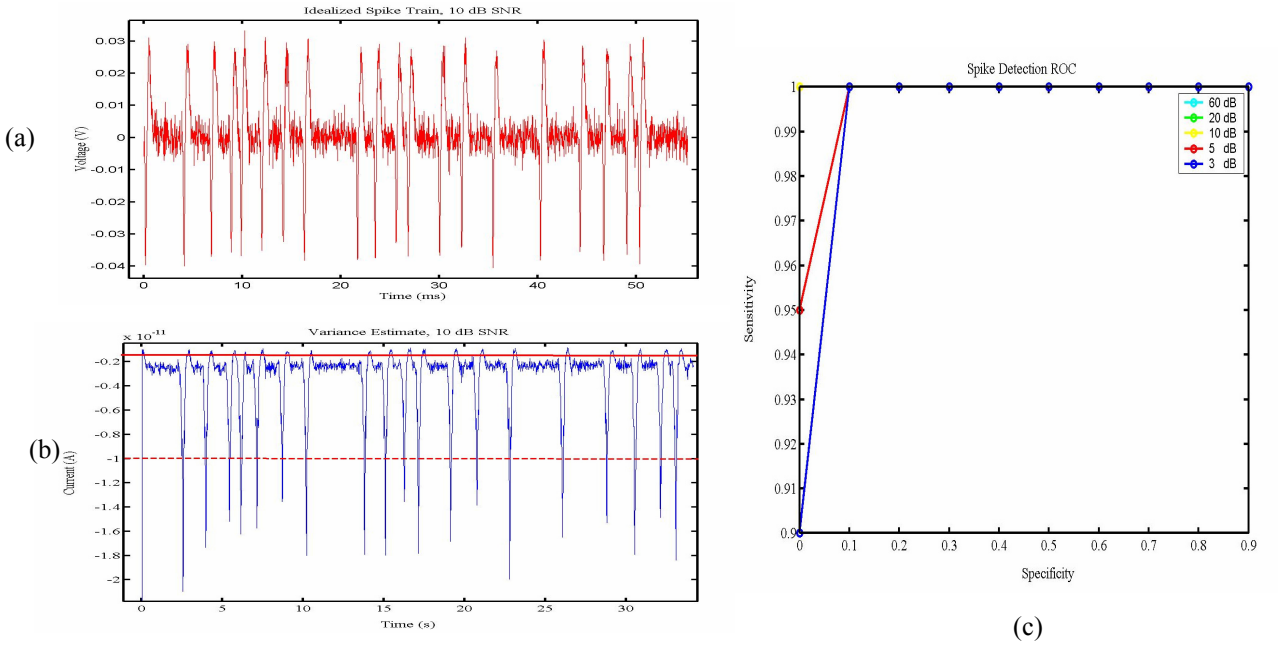


Figure 4.14: (a) simulated spike train with 10 dB SNR; (b) variance circuit estimate with user defined thresholds; (c) spike detection ROC using template matching threshold.

Detection accuracy was assessed on a series of these simulated single-unit spike trains with SNRs ranging from 60 dB down to 0 dB. Measured detection accuracy of these simulated spikes using either of the thresholding schemes described (template match or NEO) is 100% at SNR down to 10 dB, and only drops to 90% at 3 dB when using a template matching threshold. Figure 4.14 (c) shows the receiver

operating characteristic ("ROC") for single-unit spike detection as a function of SNR, using the template matching method.

Although the performance of the template matching detection scheme is satisfactory, the observed performance of the NEO element (squaring circuit) of the variance estimator was significantly better. As a result, I decided to leverage this feature by implementing an NEO detector to square the incoming signal and compares the output against a user-defined threshold. Fig. 4.15, shows the layout for the first generation detector circuit. In this figure, a compact MITE squaring circuit (left) is paired with a current mirror (middle-right) and a current comparator (right). This circuit has been fabricated and is preliminary characterization shows it operates as intended, although at low currents it is relatively slow.

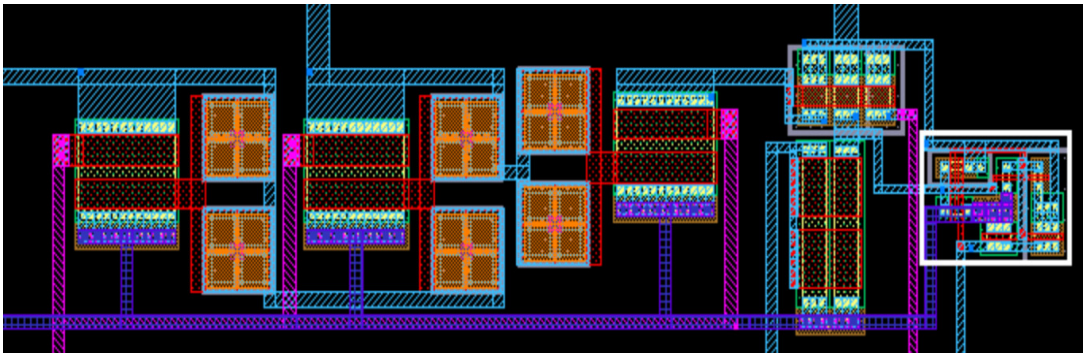


Figure 4.15: First generation spike detector layout.

As with the variance estimation circuit, the squaring element of the NEO operates in subthreshold for lower power consumption and in current mode for ease of computation. It is possible to accomplish similar result using conventional voltage mode circuits, but as always, such circuitry adds to the complexity of signal processing operations. I have subsequently updated the compact squaring circuit

(and variance estimator) to enhance operating speed and reduce device footprint; along with one of several current comparators under test, simulations and measured data indicate that it is possible to reliably resolve action potentials against background noise at near 0 dB SNR. Using either one of these action potential detection schemes, it is possible to trigger unsupervised learning in the floating gate template matching filter banks.

4.3.2 Sorting Neural Spikes

A. Floating Gate Template Matching Filter Bank

In order to sort neural spikes it is first necessary to adopt a metric for distinguishing between APs from different neurons. As we briefly discussed in the chapter on neural recording, although spikes are often considered digital all-or-nothing events, the actual AP is an analog waveform that varies significantly from neuron to neuron and under different stimulus history [44]. Likewise, proximity to or distance from a recording electrode can enhance, attenuate or distort a train of APs from one neuron versus its neighbors. Although simple feature extraction algorithms such as peak-trough detectors are capable of distinguishing between distinct classes of neurons under ideal conditions [79], and other proposed methods such as pulse coding enjoy similar success [154], template matching is believed to offer superior performance in discriminating spikes under noisy real world conditions [106].

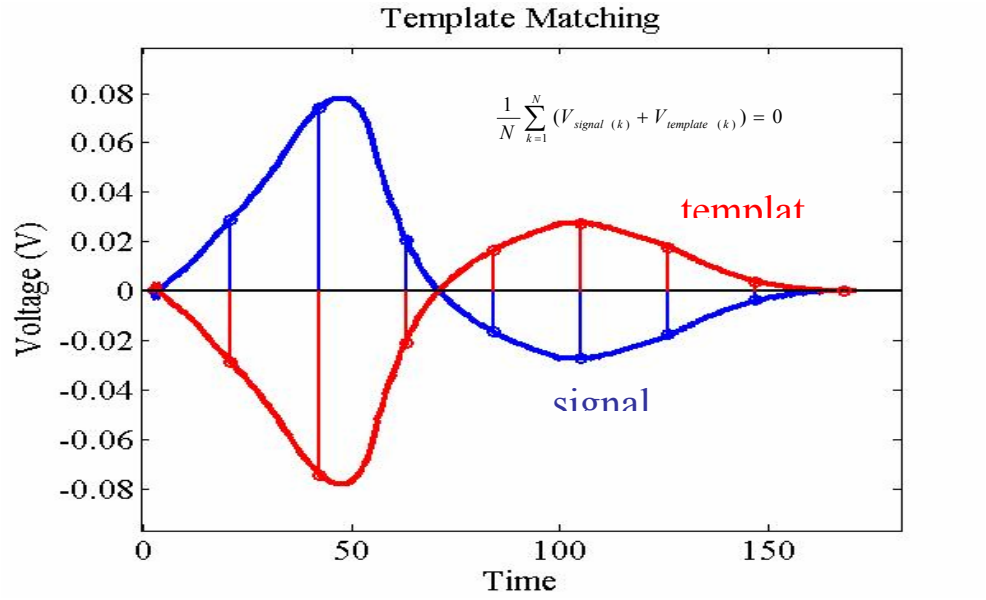
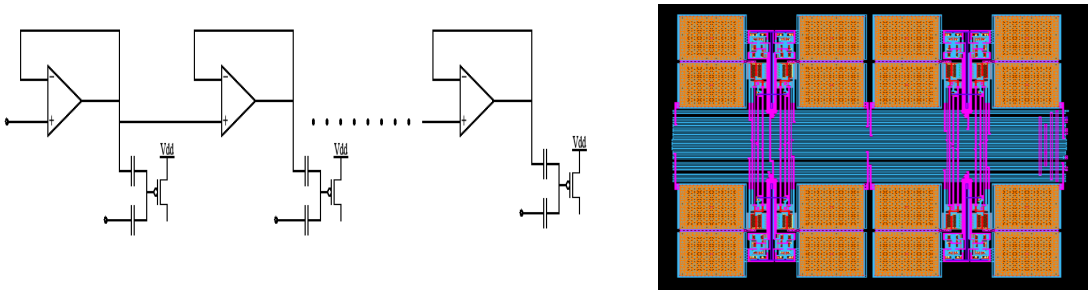


Figure 4.16: Schematic of the N (=8) point template matching method.

Thus, in order to classify spikes from one neuron from another, I have implemented an N (=8) point template matching filter. Templates for this filter can be inverted, normalized, N-point version of an idealized or measured action potential, or any other desired waveform. In hardware, the template matching circuit is implemented as a matched filter whose tap weights correspond with elements of the template. The filter decomposes the incoming signal using an analog delay line and generates eight continuous-time outputs which represent the point-wise distance of the neural waveform from the template. Classification is accomplished using the variance estimator to compute the distance between successive N-point segments of the neural signal and one or more distinct templates (which optimally form an orthornomal basis that span the signal space). Depending on the desired resolution, the filter and variance estimation circuit may be scaled to higher order, and it is also

possible to implement windowing and interpolation functions on chip. Fig. 4.16 gives a schematic representation of the template matching method.

Fig. 4.17 (a) is a schematic of the floating gate template matching filter bank; Fig. 4.17 (b) shows the layout of a fabricated filter bank. It employs a cascade of OTAs to uniformly decompose an incoming signal into eight delayed, copies of the original signal. The total delay for the line corresponds with the approximate duration of the typical biological action potential; thus at any given instant in time, the filter taps represent a spatial decomposition of an incoming neural spike. A 2-input PMOS floating gate transistor converts each tap voltage into a current; templates are stored on non-volatile, programmable floating gate memories at each tap, and control gates inputs may also be used to modulate tap currents.

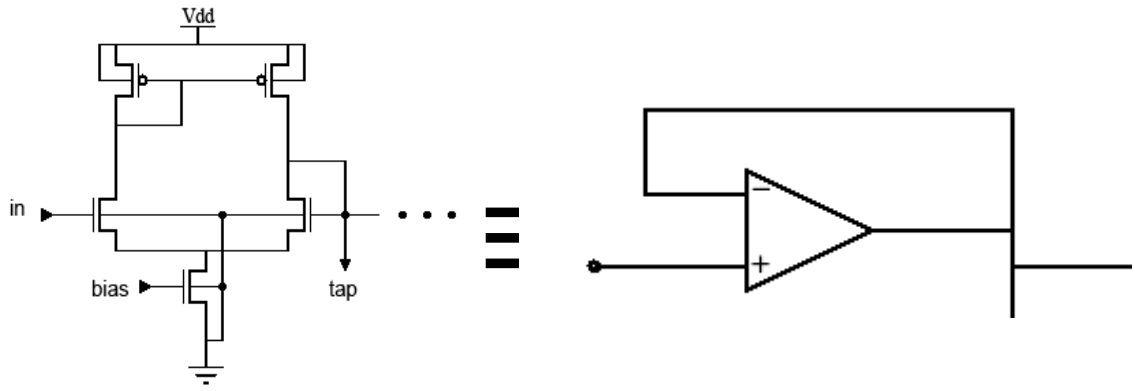


(a) (b)

Figure 4.17: (a) schematic of floating gate filter bank for neural signal decomposition; (b) layout of filter.

The RC delay for each stage is determined by a $1/g_m$ resistive component and the capacitive network between the tap node and ground; for small-signal purposes, this network primarily comprises the tap-input gate capacitance in series with the parallel combination of the second input capacitance and the parasitic capacitances of

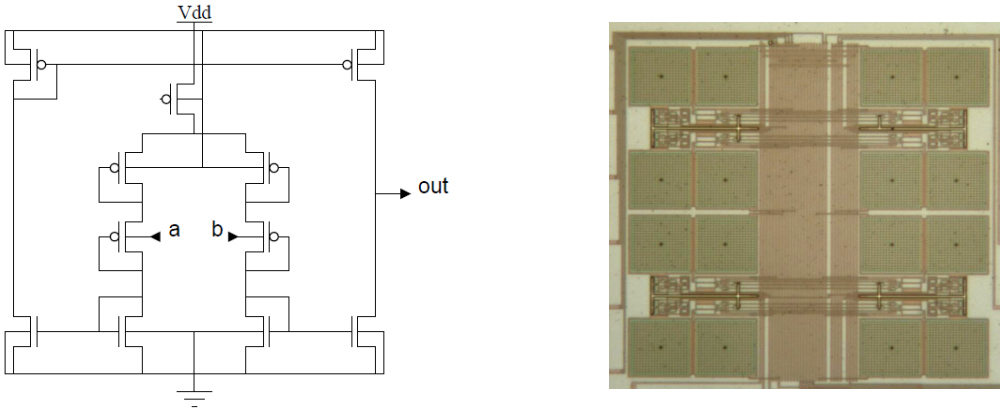
the PMOS current source and the regulating transistor (not shown) following the drain. Assuming gate oxide and poly/poly2 capacitances of about 2.5 fF/μm^2 and 1 fF/μm^2 respectively, the expected capacitance is on the order of 10 pF for each stage, whose schematic is in Fig. 4.18 below:



(a) (b)

Figure 4.18: (a) schematic of conventional OTA in unity gain configuration; (b) symbolic view.

Another implementation of this filter bank incorporates a wide linear range ("WLR") OTA designed according to the principles enumerated in [159]; Fig. 4.19 (a) shows a schematic of this amplifier; 4.19 (b) is a photomicrograph of the fabricated filter using this amplifier.



(a) (b)

Figure 4.19: (a) schematic of wide-linear range OTA with low g_m [159]; (b) fabricated filter bank incorporating filters.

In order to achieve the widest possible linear range, I followed the principles outlined in [159] and lowered g_m by source degeneration and by using well contacts as inputs. This yields superior linearity at the cost of drive, so that capacitive coupling is required – often this is a detriment, but here a feature. Bump linearization as suggested in [159] was not required for this implementation; the reduced distortion across a wider range of input signals yields theoretically superior signal decomposition with this amplifier, although attenuation must nonetheless be corrected. Fig. 4.20 (a), below shows simulated transient data for a 1 kHz sinusoidal input asserted at the input of the original OTA filter bank. Fig. 4.20 (b), shows the propagation of an idealized spike across the eight taps of a filter bank including the wide line OTA.

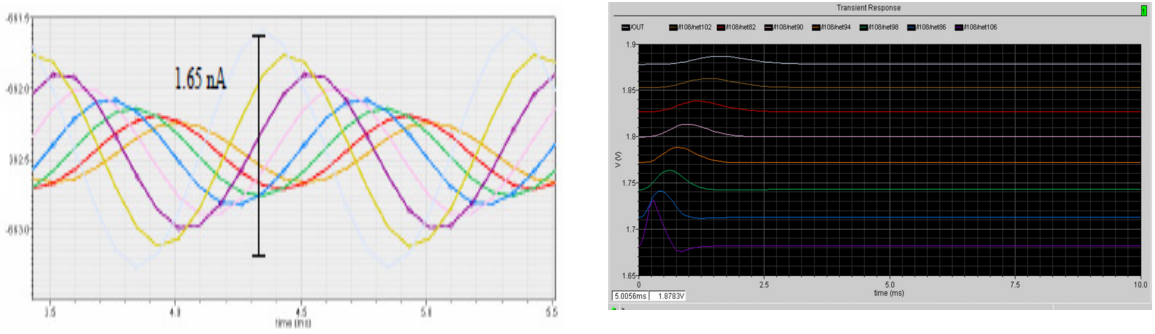
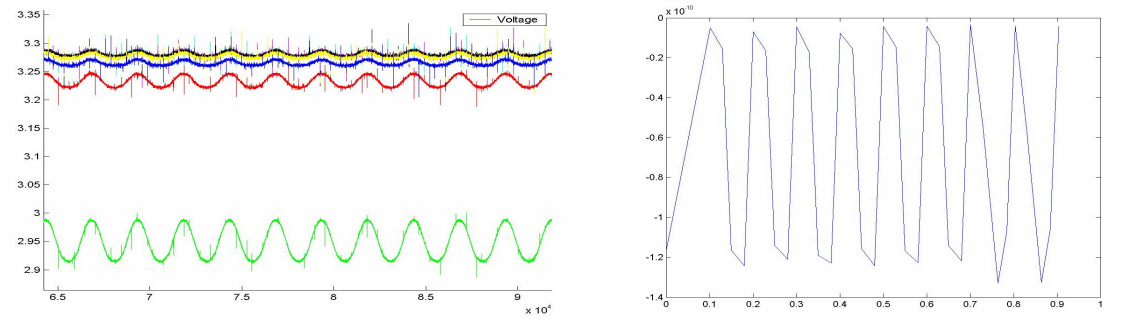


Figure 4.20: (a) simulated transient data for 1 kHz sinusoidal input asserted onto the first generation filter bank; (b) reflects the propagation of a transient spike across the taps of the WLR OTA.

Measured data from the first generation of the floating gate filter bank generated eight pico- to nano-ampere current outputs as a function of the input and bias voltages provided. Figure 4.21 shows representative measurements of (a) voltage signals generated by passing the eight outputs through a 10M+ DAQ input impedance – this serves to illustrate the temporal decomposition of an incoming

signal according to tap number and tail current bias; and also serves to illustrate fabrication mismatch variability; and (b) measured current response from a single slowly varying tap, showing the amplitude response of the filter components.



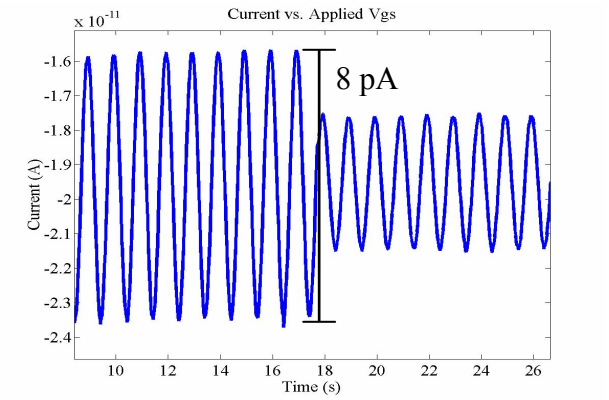
(a) (b)

Figure 4.21: (a) Direct measurement of voltage signals from floating gate taps; (b) current amplitude response of single slowly varying tap. In both cases, inferred current outputs are on the order of nA.

Two things immediately become apparent when viewing this data: (1) a higher speed, high-fidelity recording method to assess these tiny time-varying local currents is required to characterize the filter when it operates at pA-nA levels – suitable current amplifiers are will be implemented in subsequent work; (2) the tap-to-tap offsets and attenuation are significant and must be eliminated. Although it is theoretically possible to design multistage filters with considerably improved fidelity across the taps, for my proof-of-principle circuits, we can compensate attenuation across the taps more simply in one of two ways.

First, it is possible to buttress successive taps by applying a proportional DC bias onto their control gates, if the approximate offset is know in advance. Since the tap current is proportional to a weighted combination of the voltages applied at the

input and control gates, we can modulate each tap signal by simply adjusting the biases as shown in Fig. 4.22.



$$i_{tap} \approx e^{\kappa(V_{tap}+V_{control})/2V_T}$$

Figure 4.22: Modulating the current output of a fabricated filter bank tap by adjusting the capacitively coupled DC control bias.

Second, we can program arbitrary offsets onto the floating gates of each tap using using hot electron injection and FN tunneling – this accomplishes a similar result.

The template is stored onto the filter using analog non-volatile floating gate memories at each tap. These weights may be (a) programmed directly by the user with an on-chip an analog multiplexer; (b) sequentially generated by a function mapped onto silicon (silicon neuron or other periodic function); or (c) learned adaptively from training data.

Initially I had proposed to correct mismatch and program transistors using the mechanism outlined in Fig. 4.23. Floating gate tap weights were to be: (1) initially updated during a mismatch correction phase by charge injection driven by the difference of adjacent tap currents for the identical signal bias. This difference itself was to be measured by a precision, FG matched, current steering circuit, whose

output would be amplified and converted to a voltage in order to drive the injection / tunneling depending on the direction of the current flow

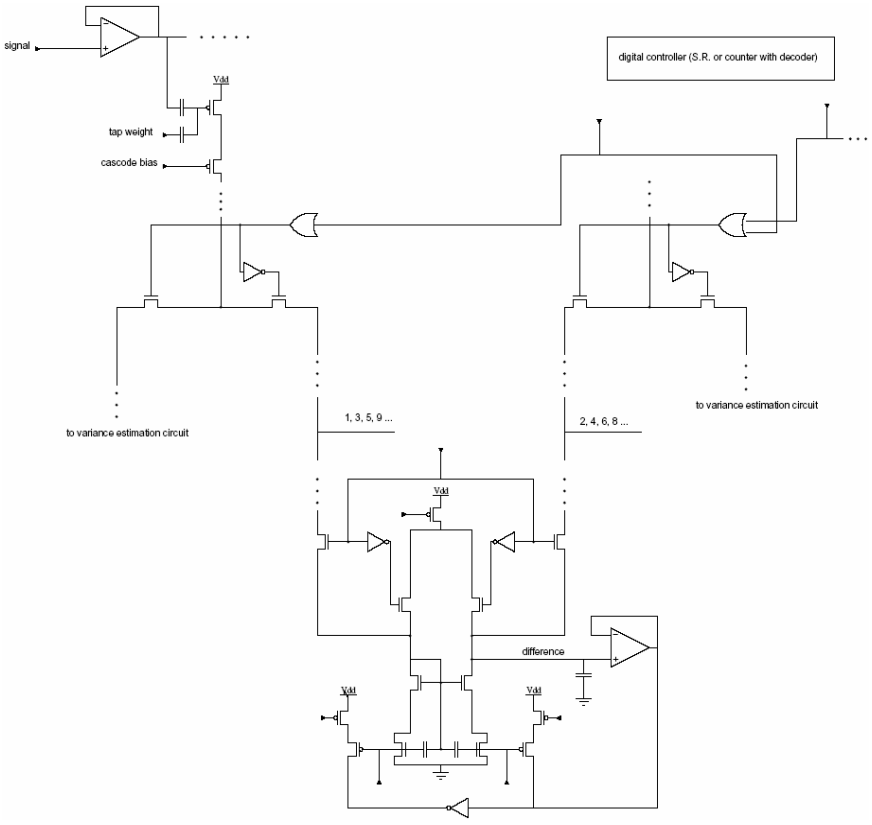


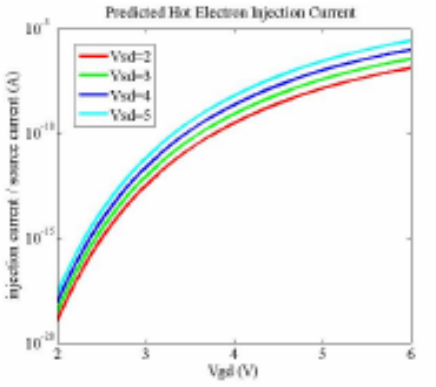
Figure 4.23: Schematic illustrating proposed method of mismatch correction.

(2) During a second, template learning, cycle tap weights would have been updated adaptively by injection / tunneling synchronized with the onset of the template waveform. The update at each tap would be keyed to a fixed delay from the time that a non-linear energy operator (NEO), (threshold amplitude)*(frequency), as proposed in [114] indicates a template waveform is detected. The NEO was to be accomplished by taking the product of a peak detector voltage and frequency which was to be evaluated as an inverse function of the difference in amplitude between

successive tap signals. Simulations performed using the equation-based Rahimi model in TSpice indicated that the proposed system would achieve convergence over time, but the overhead associated with tuning the continuous-time feedback system properly proved too high. In repeated simulations, the floating gate was charged in response to a ramped voltage signal and this modulated the current associated difference between input and output. Although negative feedback prevented the charge injection from becoming unstable, overshoot was a significant issue in these simulations.

To avoid this issue, and attain bi-directional updates approximating a rudimentary stochastic gradient descent, I redesigned the programming mechanism to compare the two currents and trigger small discrete weight updates for each time the mechanism was triggered. For hot electron injection, Eric Wong had previously characterized the Rahimi parameters for injection currents – computed IV curves are shown in Fig. 4.24, below. The mechanism for hot electron injection is schematically illustrated in Fig. 4.25; tap currents are evaluated against either a fixed bias or sampled signal current and the difference between the two triggers a voltage controlled oscillator ("VCO") that turns on a negative charge pump ("NCP") which drops the voltage at the drain of a programming transistor sufficiently to induce hot electron injection across the transistor.

Figs. 4.26 (a) and (b) show programming of a particular tap using hot electron injection. Fig. 4.26(a) illustrates the programming of offsets onto an isolated tap, which are mathematically subtracted in the associated inlay.



$$I_{inj} = 1.076\times10^{-5}\cdot I_s \exp\left(\frac{-318.125}{(V_{gd}+1.053)^2}+V_{sd}\right)$$

Figure 4.24: Theoretical hot electron injection currents based on Rahimi model and with experimentally fitted parameters.

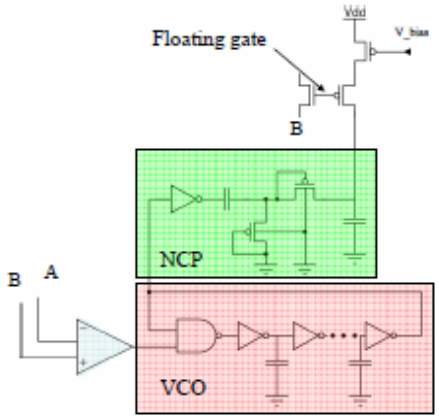
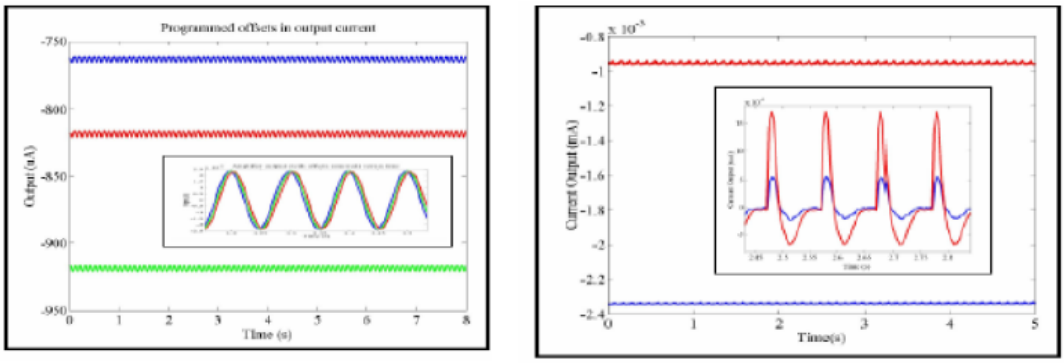


Figure 4.25: High-level schematic of programming mechanism. Comparison between desired and measured current drives injection or tunneling.

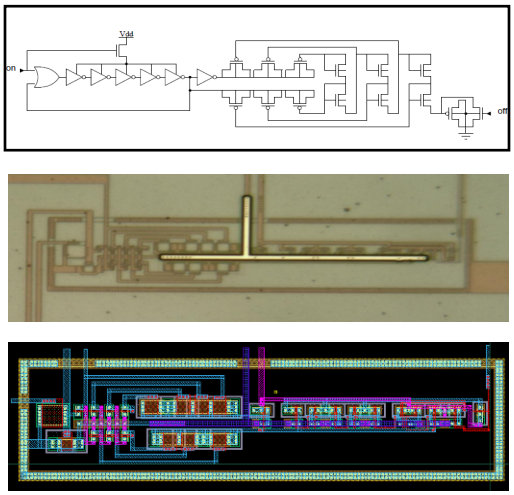


(a) (b)

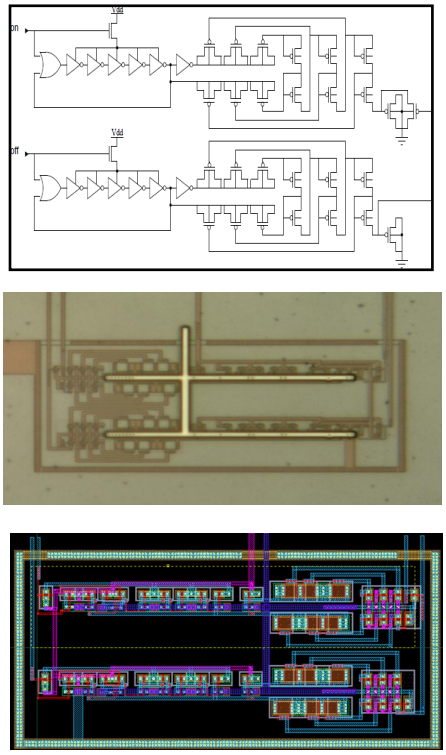
Figure 4.26: Programming arbitrary offsets onto floating gates in order to shift I-V curves and tune gain. Within each box, input signals are identical. Inlays show signals with DC offsets subtracted.

Fig. 4.26 (b) shows programming of a single tap, resulting in a modulation of both offset and gain. The subtraction may be accomplished on chip using FN tunneling as an erase mechanism. Programming may be unidirectional, or bi-directional. That is two separate comparators of opposite polarity have been designed to trigger respective negative and positive charge pumps that will pulse a fixed unit of charge onto the floating node, thereby decreasing or increasing the voltage on the tap.

Schematics and layouts for the two Dickson charge pumps, including VCO's are shown in Figs. 4.27 (a) & 4.27 (b), below.



(a)



(b)

Figure 4.27: Schematic, layout and photomicrograph of: (a) fabricated high voltage positive charge pump; (b) fabricated low voltage negative charge pump.

In order to achieve precise control of the NCP pulse width, a second charge pump was implemented to drive the negative voltage required to rapidly turn open the control switch and shunt the CP voltage to ground. As a result, pulse widths are well controlled, but inter-pulse intervals are a function of slow capacitive decay. For the positive charge pump, a rail logic high value will serve to drop the CP output to ground. The output of the positive charge pump may be asserted simultaneously during the programming period through high voltage switches at each tap and the output of a single NCP can also be multiplexed with a series of compact negative

voltage generators (mini-NCPs); for finer control, each tap may have its own pair of independently timed positive and NCPs.

Schematics and layout of some high voltage switches designed after a floating gate array reported by Hasler, and a variant on the Traff current comparator [160] used to respectively multiplex and control the programming voltages are shown in Figs. 4.28 & 4.29 below:

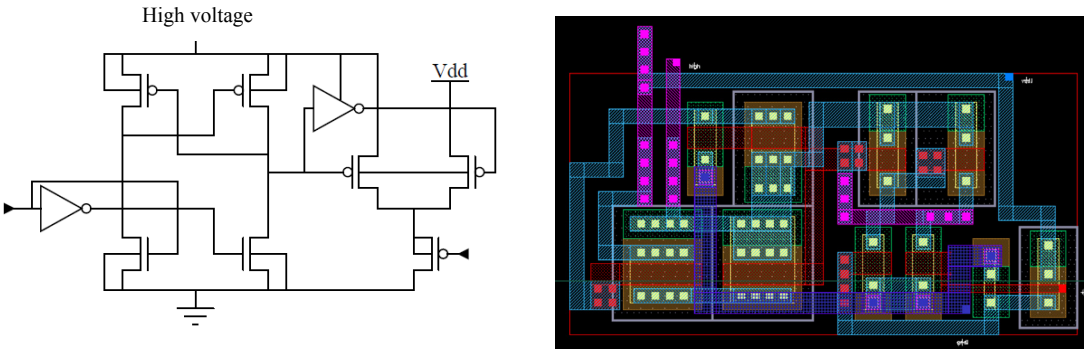


Figure 4.28: Schematic and layout for a fabricated set of high voltage switches.

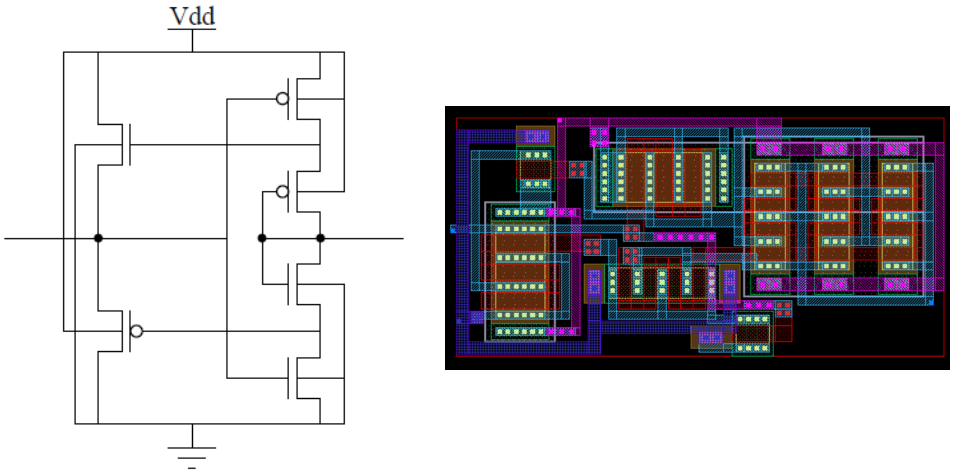


Figure 4.29: Schematic and layout for a fabricated Traff current comparator, [160].

Together, these components comprise a floating gate template matching filter bank capable of unsupervised learning and signal decomposition and filtering. To

assess the instant quality of a match, we turn next to my fabricated variance estimation circuits.

B. Variance Estimation Circuit

I considered several different metrics for evaluating the proximity of an incident neural signal to the programmed template stored in the filter bank. However, simple Euclidean distance metrics neglected spike shapes and focused solely on magnitudes of peaks and troughs – thus neither attenuated nor offset spikes would register properly. By contrast more complex theoretical algorithms often required A/D conversion and memory to properly represent and encode incident neural data. The variance estimator was chosen, therefore, to represent a low power, real-time analog architecture with no memory or clock overhead, that would nonetheless register spikes based on their shape and not on the amplitude of their features. Fig. 4.30 illustrates the idea – the N-point variance of an offset signal with identical shape to the template is 15 orders of magnitude smaller than the variance of a distorted signal with equivalent Euclidean distance from the template, but entirely different shape.

Variance is the mean square error of an unbiased estimator, and as such provides a statistically valid metric for measuring proximity of signal to template. In view of these advantages, and several others, I chose to implement a variance estimation circuit to compute variance, defined as $\sigma^2 = \mathrm{E}\ [(\mathrm{X} - \mathrm{E}\ [\mathrm{X}])^2] = \mathrm{E}\ [\mathrm{X}^2] - \mathrm{E}^2[\mathrm{X}]$, where $\mathrm{E}\ [\mathrm{X}] = \frac{1}{N}\sum_{k=1}^{N} x_k$ is the expected

value, or mean, of a random variable, X, for the input signals from the template matching filter.

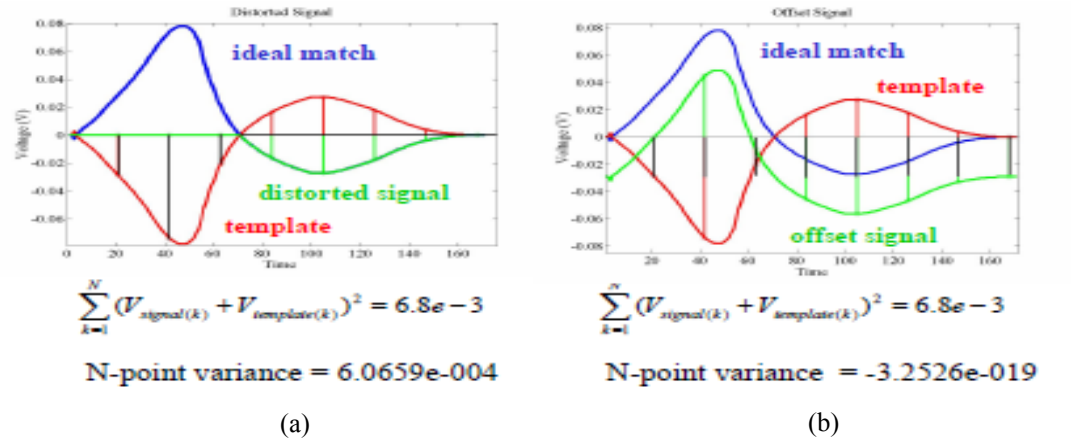


Figure 4.30: Computed variance estimates for the distance between (a) a distorted signal and the stored template; and (b) an offset signal and the stored template.

A high-level representation of the variance circuit is shown in Fig. 4.31, below:

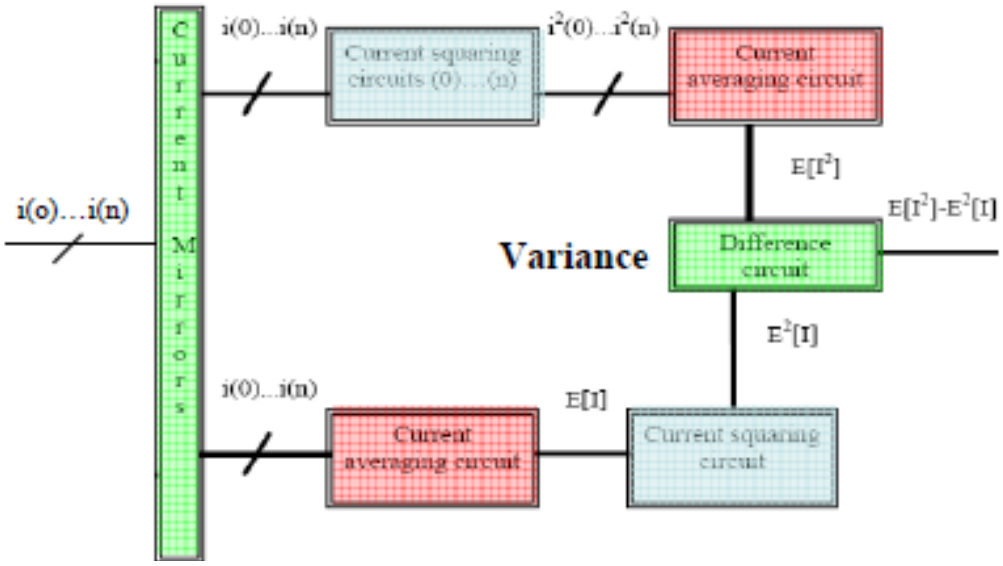


Figure 4.31: Block diagram of variance estimation circuit.

This circuit computes $\sigma^2(I)$ by: (a) copying eight input currents, $i(0)....i(7)$; (b) individually squaring and then averaging one set of currents to generate $E[I^2]$; (c) averaging and then squaring the average of the second set of currents, to compute $E^2[I]$; and then (d) subtracting the second result from the first. Precision current copies in the first generation circuit were created using stacked cascode current mirrors that provided good linearity over a wide range of input currents, from approximately 10 pA to 10 nA. Average currents are computed by providing copies of the input currents to the common gate and drain of an equal number of identical diode-connected PMOS transistors [151]. Multiple-input translinear elements ("MITE") comprised of floating gate NMOS transistors operating in subthreshold were employed as squaring circuits [158]. For matched devices operating in saturation, i.e., V_{ds} > 4kT/q ≈ 100 mV, the translinear circuits produce an output current, $i\_out \approx \frac{i^2(n)}{i\_bias}$. The variance estimation circuit thus uses a unique combination of high-precision current mirrors, simple current-averaging components, and subthreshold translinear squaring circuits to compute a true analog variance estimate without incurring the costs of sampling, quantizing, storing and manipulating digital data.

We have previously performed the squaring circuit analysis, and subtracting currents is a matter of mirrors. For n-current inputs to the averaging circuit (a 2d generation of which is shown in Fig, 4.33) we can mathematically derive the averaging function as follows [151]:

$$I_{total} = I_1 + I_2 + ... + I_n = \lambda \cdot I_0(e^{\kappa V_{GS1}/V_T}) + \lambda \cdot I_0(e^{\kappa V_{GS2}/V_T}) + ... + \lambda \cdot I_0(e^{\kappa V_{GSn}/V_T}) \quad (4.27)$$

$$= \lambda \cdot I_0 (e^{\kappa V_{GS1}/V_T} + e^{\kappa V_{GS2}/V_T} + \ldots + e^{\kappa V_{GSn}/V_T}) = n\lambda \cdot I_0 (e^{\kappa V_{GS1}/V_T}) \quad (4.28)$$

$$I_{out} = \lambda \cdot I_0 (e^{\kappa (\ln(I_{total}/n\lambda I_0)(V_T/\kappa)/V_T}) = I_{total}/n \quad (4.29)$$

In the first generation variance estimation circuit I used MITE squaring circuits [158], and oversized mirrors to enhance circuit matching and precision. For the second generation, in order to improve speed, I used MOSFETs biased into the sub-threshold regime to implement the mathematical functions without the delays associated with explicit capacitors at the inputs. For enhanced precision, I employed ultra-low-power current mirrors in all of my circuits [161], as shown in Fig. 4.32. By biasing the gate of the input transistor below its drain, as reported in [161], it is possible to extend the operating range of the mirror into the sub-pico ampere range. This permits even lower power precision current mode computation. A schematic of the new averaging circuit is shown in Fig. 4.33.

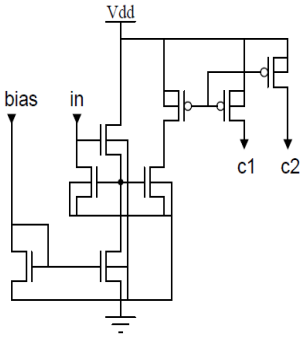


Figure 4.32: Ultra-low-current current mirror

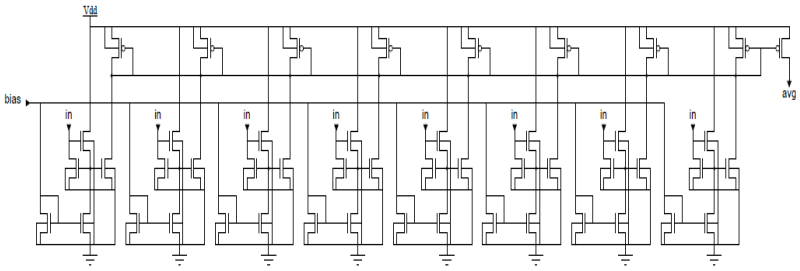


Figure 4.33: Second generation current averaging circuit.

Figure 4.34(a) shows the subthreshold current squaring circuit, and the difference circuit is a variation on the low-power mirror, as is shown in Fig. 4.34 (b).

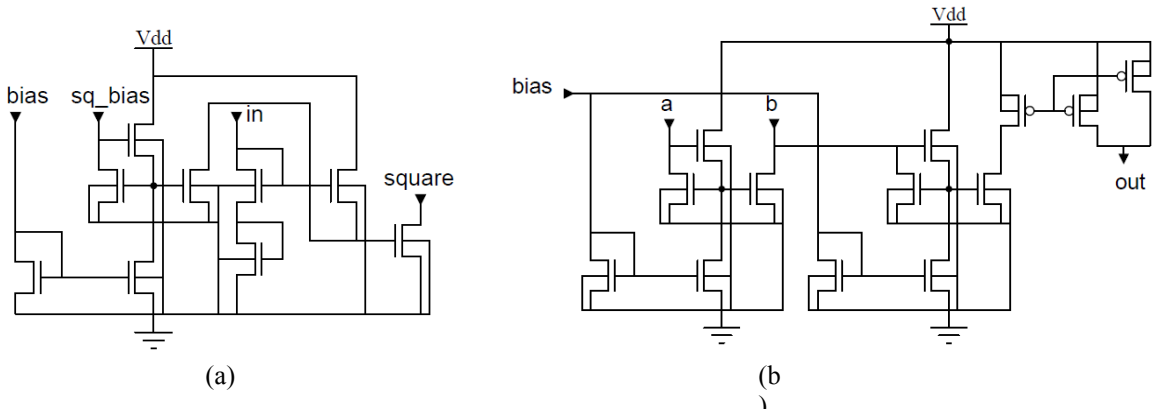


Figure 4.34: (a) second generation subthreshold squaring circuit; (b) second generation current subtraction circuit.

Integrating these elements together, I have developed an ultra-low-power current-mode circuit that generates an analog variance estimate across N current inputs in real-time. The first generation variance estimation circuit has a footprint of $< 0.15\ mm^2$ for N=8 in a commercial 0.5 µm 3-metal, 2-poly process

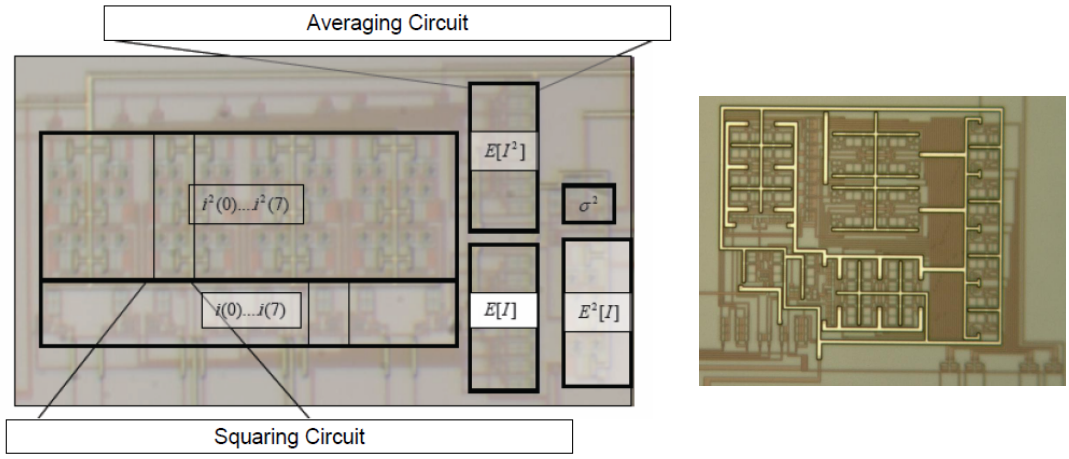


Figure 4.35: (a) Labeled photomicrograph of first generation variance estimation circuit; (b) bare photomicrograph of fabricated second generation circuit.

. A photomicrograph of the original MITE variance estimation circuit is shown in Fig. 4.35, alongside the layout of the second generation circuit, which is approximately 2/3 the size of the original and theoretically over an order of magnitude faster. As an online signal classifier, the N=8 circuit computes the instantaneous distance of a continuous-time input signal from an eight element template while dissipating less than 10 nW of power. A comparison of simulated and measured results demonstrates the suitability of the circuit for integrated mixed-signal applications. In particular, Monte Carlo mismatch and process analysis reveal a picoampere floor on current accuracy, while variance estimates as a function of DC sweeps confirm both theoretical (MATLAB) and simulated (Cadence Spectre) estimates. Fig. 4.36 shows the measured common-mode response of the variance estimation circuit (1st G) to a DC sweep of input currents.

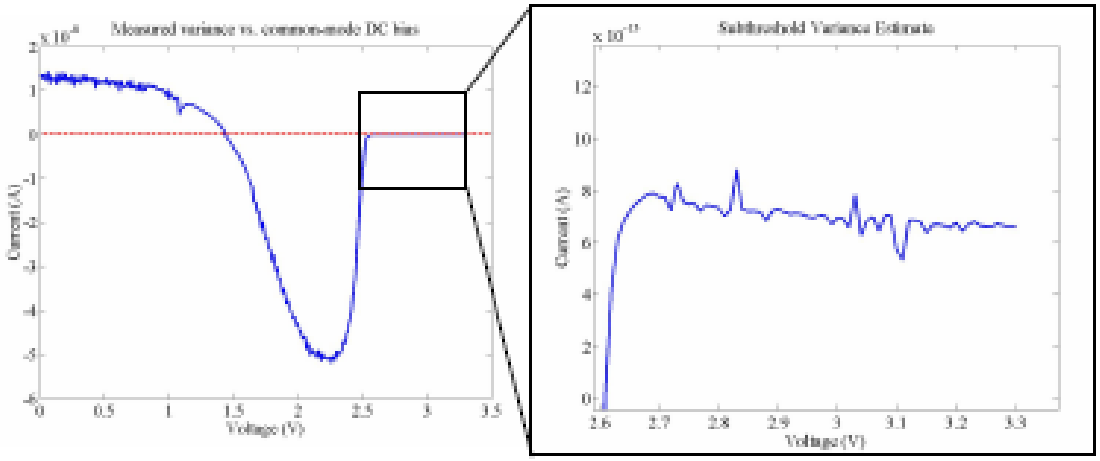
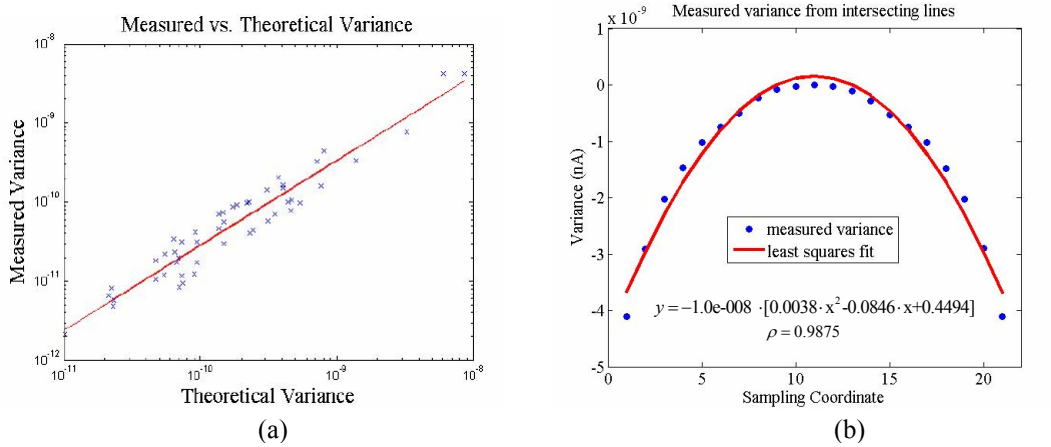


Figure 4.36: DC response of first generation variance estimation circuit.

Ideally, the CM response is zero, but systematic circuit biases and device mismatch cause a current offset. For input voltages within the operating range between 2.6 and 3.3 V, corresponding with input currents between 10 nA and 100 fA,

the magnitude of the CM offset is less than 1 pA. This represents < 0.1 % error when circuit inputs are in the nA range.

DC operating characteristics for the first generation circuit were verified using a 16 bit, digital-to-analog converter to drive the eight VCCS inputs to specified values and the SMU of an HP 4156B semiconductor parameter analyzer to measure the variance output. In one experiment, random sequences of eight voltages, normalized to produce currents within the operating range (10 pA to 10 nA), were simultaneously asserted at one second intervals and theoretical and measured variance estimates were compared, as shown in Figure 4.37 (a). The correlation coefficient between the logarithms of the average measured and computed values for this data is ρ=0.96; repeated trials show that this result is typical, although outliers can dramatically affect the correlation (e.g., for one N=100 trial, ρ=0.92). Figure 4.37(b) illustrates the measured and best-fit variances from another experiment in which input currents are computed by sampling eight intersecting lines at regular intervals along the x-axis of the plot representing these lines.



(a) (b)

Figure 4.37: Measured versus theoretically computed variance estimates. In (a) inputs are random; *p*=0.96; in (b) inputs are converging (intersecting) linear currents, red line is fit.

In all cases, the measured and predicted results show good agreement, although the theoretical estimates do not perfectly reflect actual circuit operation because they neglect real-world parasitics and (correctable) sources of circuit and device mismatch.

When employed as a dynamic distance estimator for a template matching algorithm, we observe a 1:1 correspondence between spike template match (as rendered by a MATLAB controlled 8-channel D-to-A converter) and variance estimate minimum as shown in the following figures:

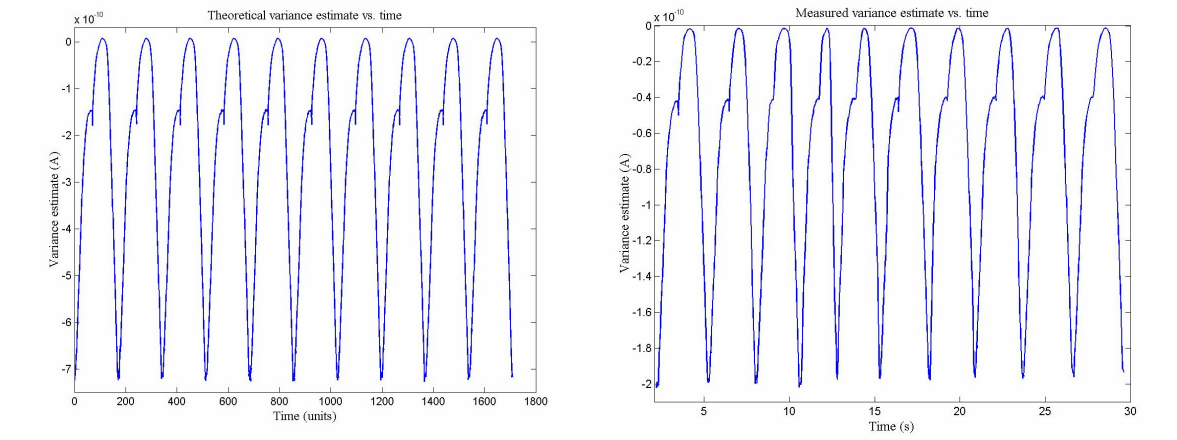


Figure 4.38: Theoretical and measured response of variance estimation circuit to MATLAB simulated output of floating gate filter bank to artificial neural spike train.

Finally, as reported above in our discussion on detection of neural action potentials, I have characterized the first generation variance estimation circuits' classification capability across two different classes of neural APs. Results indicate reliable binary classification down to 20 dB SNR, although these results do not account for uncompensated filter bank process and mismatch variations. Figure 4.14 displays representative data confirming the suitability of the variance estimation circuit as a distance estimator, albeit a slow one.

The second generation variance estimation circuit was designed to operate on a faster time-scale at lower currents, which should provide adequate bandwidth for biological signal processing and classification. It is also possible to scale the distance estimator, along with the control and filtering circuits, to processes with smaller feature sizes in order to enhance speed and performance, although such scaling is not trivial. In particular, deep submicron scaling is not linear and parasitics and mismatch variation must be considered carefully as scaled designs are drawn.

C. Classification Block

The detection / classification decision element for the circuit is a simple current comparator triggered by a negative threshold crossing, indicating a template match. The speed of the current comparator is principally a function of the time it takes to charge the implicit capacitors of the input transistors – this in turn is a function of transistor sizing and bias conditions. In practice, with appropriate DC biasing, switching times on the order of ns [160] have been reported.

4.3.3 System Performance

As noted in the subsection on detection, ferret auditory neural signals have been quantized, decomposed and fed through the variance estimator to assess the reliability of the template matching method, see Fig. 3.14. Results are promising, but depend heavily on the template generation scheme. Subsequent stages of testing may

continue to use pre-recorded neural signals as inputs, or actual cultured cells atop an integrated spike sorting chip with bioamplifiers and electrodes will be implemented. No live animal testing is anticipated, although blowfly neural recordings may be taken.

Over the course of the past two years, I have fabricated the proposed spike sorting architectures and tested them against one another and theoretical performance limits, evaluating their performance against the feature space of mixed-signal spike sorting architectures. I have developed circuits to permit the system to blindly learn its programmed tap weights from incoming neural signals. Further, I have designed a tunable silicon neuron template generation mechanism (not shown) for unsupervised training of the template matching architectures. While performing this work, I have continued my research into extracellular recording of neural signals from cultured cells and also investigated low-noise electrode, bioamplifier and optical front-ends, particularly focusing on arrays of sensor devices. I have tested and characterized fabricated circuits using real neural data and biologically realistic simulations, and evaluated the power efficiency and detection and classification of these circuits against one another and theoretical analog and reported digital performance limits.

One aim of my research has been to develop novel mixed-signal circuits for neural recording, event detection and classification that are both ultra-low-power and high-precision. Simulated and experimentally verified performance of these circuits demonstrates reasonable precision and ultra-low-power operation. There is a tradeoff between the two that can be exploited while still coming under power budget. Given the demonstrated performance of the components of the system, it is possible that

optimized versions could be integrated into the next generation of implantable neural prosthetics – enhancing performance of these prosthetics by increasing the number of recording channels simultaneously measured, while reliably detecting, classifying and encoding neural events in real-time. It is my hope that this research will thus enable more detailed investigation into local cortical function and better closed-loop feedback control of neural prosthetics, facilitating the study and treatment of those who suffer from debilitating neural injury or disease.

Chapter 5: The Two Transistor Synapse

In this Chapter, we consider biological Hebbian learning, explore a new analog circuit architecture for implementing biologically realistic learning, and conclude with an illustrative pattern recognition application. Original contributions of this thesis to the field include the development and implementation of a novel two transistor synapse that exhibits spike timing dependent plasticity and can implement adaptive pattern classification and silicon learning.

5.1 Two Transistor Synapse with STDP

5.1.1 Hebbian Learning and STDP

In biology, spike-timing-dependent-plasticity ("STDP") describes the strengthening (potentiation) or weakening (depression) of synaptic connections between neurons according to the coincidence of pre- and post-synaptic action potentials. It has been experimentally observed that when a pre-synaptic action potential is followed within some time (typically on the order of ms) by a post-synaptic action potential, the strength of the synaptic connection between the two neurons is increased in proportion to the coincidence of the firing times (or equivalently in inverse proportion to the time between firings). Conversely, it has

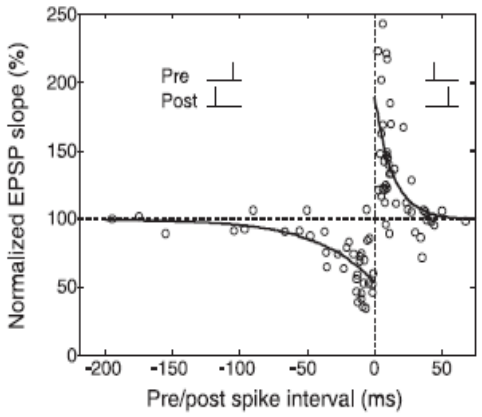


Figure 5.1: Biological spike timing dependent plasticity. [162],[163]

been shown that when a pre-synaptic action potential follows within some time (typically on the order of ms) a post-synaptic action potential, then the strength of the synaptic connection between the two neurons is decreased in proportion to the coincidence of the firing times (in inverse proportion to the time between firings). These rules which predict the potentiation and depression of synaptic weights and which together give rise to STDP are known as Hebbian learning. The concepts are graphically illustrated in a plot of biological data shown in Figure 5.1, as adapted from [162] by [163].

While the biological mechanisms of STDP have not been fully elucidated, they represent a ubiquitous and important mode of neural adaptation and learning. Therefore, in order to begin to realize electronic analogues of rudimentary cortical functions, we must develop synaptic mechanisms that incorporate STDP-like behavior. Furthermore, in order to accurately synthesize even the simplest neural architectures, STDP synapses must be realized in ultra-compact form with very high integration density. As a first step down that path, I have developed a two transistor synapse for implementing biologically realistic STDP [15].

5.1.2 Two Transistor Synapse

The two transistor synapse ("2TS") is designed to exhibit spike timing dependent plasticity ("STDP"), as with a real biological synapse. In short, temporal coincidence of synthetic pre- and post- synaptic action potentials across the 2TS induces localized floating gate injection and tunneling that result in proportional Hebbian synaptic weight updates. In the absence of correlated pre- and post- synaptic activity, no significant weight updates occur. A compact implementation of the 2TS

has been simulated, and fabricated in a commercial 0.5 μm process. Suitable synthetic neural waveforms for symmetric STDP have been derived and circuit and network operation have been modeled and tested. Simulations agree with theory and preliminary experimental results.

There are several reported single, [170], and two-transistor synapses, [165]-[169], some of which are programmable [164], [168], [169]. Further, there are reported STDP synapses that comprise many transistors [171], [172]. However, because the 2TS employs the same control signals to concurrently update synaptic weights and to pass information between pre- and post- synaptic nodes, as in biological systems, the 2TS is both simpler and smaller than any other integrated STDP realization.

The two transistor synapse ("2TS") comprises two PFET transistors with a floating gate node that is common to both. Figure 5.2 shows a circuit schematic of one 2TS configuration wherein the "pre"-synaptic signal is asserted at both PMOS sources, and the "post"-synaptic signal defines the potential of the drain of the programming transistor on the left. The PFET on the right passes current generated by "pre"-synaptic spikes to the integration node, or soma, of the post-synaptic neuron (not shown). The body of the programming transistor on the left is connected to the source, while the body of the right transistor is held at a fixed potential.

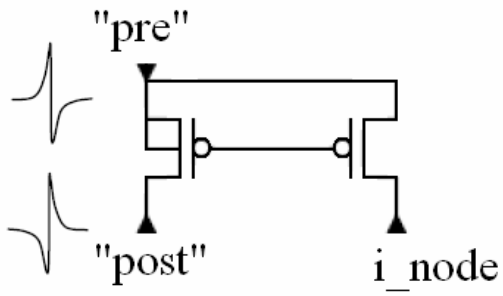


Figure 5.2: Schematic of the two-transistor synapse with illustrative "pre" and "post" waveforms.

In theory, the operation of the 2TS circuit is relatively straightforward. PRE and POST synaptic waveforms are asserted at the corresponding labeled nodes. If PRE occurs first, but POST occurs within some prescribed time of PRE, then PRE and POST will overlap resulting in a large transient difference in the source-drain voltage of the programming transistor causing hot electron injection to decrease the stored voltage on the floating gate and thereby increase the synaptic weight. On the other hand, when POST occurs first, but PRE occurs within some prescribed time of POST, then PRE and POST will overlap at a high voltage causing Fowler-Nordheim tunneling to increase the stored voltage on the floating gate and thereby decrease the synaptic weight.

In practice, circuit operation is more intricate. First, it is necessary to rely on an implicit injection threshold to avoid significant positive weight updates in the absence of PRE/POST overlap. This operating assumption follows from the characteristic exponential relationship between the injection current and programming transistor source to drain voltage. Likewise, when the drain of the programming transistor is held low, a transient PRE overvoltage that is also coupled to the body of the programming transistor will not generate significant negative weight updates. Preliminary experimental results indicate that holding the drain voltage on the programming transistor several volts below the programming threshold voltage reduces the field across the oxide sufficiently to adequately suppress tunneling.

There are many factors in the physical realization which will impact the layout of the 2TS. Among the most significant of these are the gate length of the programming transistor, which can mitigate short-channel effects, and the doping

density of the source and drain regions, which impact local field strength and carrier transport. However, while such considerations are important for optimization of the structure and function, for a proof-of-principle implementation, I fabricated a simple symmetric 2TS in a commercial 3-metal, 2-poly, 0.5 μm process, as shown in Figure 5.3. Each of the six terminals of this 2TS is connected to a separate pad for testing.

The non-optimized proof-of-principle layout has a fairly large footprint at just under 400 μm^2. However, it is estimated that a minimum-sized variant of the 2TS satisfying analog design rules in a commercial 90 nm process consumes less than 6 μm^2 of real estate; we have fabricated a 90 nm process 2TS.

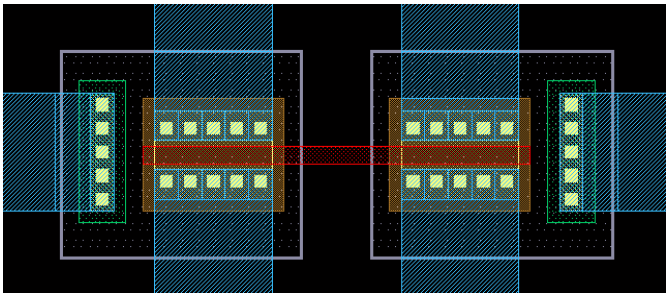


Figure 5.3: Physical layout of a 2TS in a commercial 0.5 μm process.

Allowing both PFETs to share the same well would reduce the minimum dimensions even further; our model simulations and experimental characterizations of tunneling behavior in fabricated devices suggest that it may be possible to place both transistors of the 2TS in the same well without significantly compromising operation. With modern lithographic techniques extending integrated circuit technologies deep into the nanometer regime, a sub-micrometer 2TS is technically feasible. However, it should be noted that owing to the non-linear scaling of submicron CMOS – neither threshold voltages, nor gate leakage currents scale proportionally with size in this

regime – that such scaling poses additional technical challenges for deep submicron design. Among the concerns are high gate leakage currents for ultra-thin gate oxides, potentially resulting in non-volatile memory storage on the order of minutes to days rather than years. These and other considerations must be addressed in any scaled design.

There are an infinite number of potential waveforms that will yield positive and negative Hebbian weight updates when asserted across the 2TS, see, *e.g.*, [164]. However, the subset that can accomplish bidirectional Hebbian learning according to biologically realistic STDP rules is considerably smaller. In this work, I developed and investigated two such classes of waveforms: (1) a uniphasic PRE signal and a biphasic POST signal; and (2) paired, mirror-symmetric biphasic PRE and POST signals. Integrated circuit architectures for implementing these types of signals have been previously reported, see, *e.g.* [172].

Synaptic weight updates are accomplished by balanced hot electron injection and Fowler-Nordheim tunneling. Hot electron injection in MOSFETs is a function of transistor source- and gate- drain voltages, and has been empirically shown to obey the following relationship [64]:

$$I_{inj} = \alpha \cdot I_s \cdot \exp\left(-\frac{\beta}{(V_{gd} + \delta)^2} + \lambda \cdot V_{sd} \right) \tag{5.1}$$

where α, β, and δ are experimentally derived process dependent constants, I_s is the source-drain current flowing through the transistor, and V_{gd} and V_{sd} are the voltages

across the gate-drain and source-drain regions, respectively. For a relatively constant gate voltage, the exponential dependence of the injection current on the source and drain voltages, allows us to determine a threshold voltage below which no significant (< 1% of max.) injection occurs.

Similarly, Fowler-Nordheim tunneling in MOSFETs exhibits an exponential dependence on the voltage across the oxide barrier that is given by [64]:

$$I_{tun} = -I_{tun0} \cdot WL \cdot \exp\left(-\frac{V_f}{V_{ox}}\right) \tag{5.2}$$

where I_{tun0} is a pre-exponential current, V_f is a process-dependent constant, and V_{ox} is the voltage across the oxide barrier. For semi-empirical modeling, V_{ox} was computed to first order as representing a weighted average of source, body and drain voltages. This simplification discounts localized potential differences and barrier reduction to some degree, although experimental data suggests that it is a reasonable approximation in test devices.

From the analyses and investigation performed, the most suitable waveforms for implementing realistic STDP appeared to be mirror-symmetric, biphasic PRE and POST signals. Since tunneling and injection are both exponential functions, a variant of the exponential STDP curve itself was used as a starting point for the PRE signal. Then, in order to produce the desired Hebbian updates – that is, to balance the positive and negative increments for correlated PRE and POST signals, strengthening synaptic

connections when PRE precedes POST and weakening them when PRE follows POST – the POST signal was taken as the mirror-symmetric, biphasic version of the PRE signal. The PRE and POST signals shown in Figure 5.4 can be mathematically weighted, with appropriate thresholds, to yield the biologically realistic STDP curve shown in Figure 5.5.

In Figure 5.5, each point in the STDP curve represents the integration of the injection and tunneling contributions at a single instant in time as the PRE and POST waveforms are convolved past one another. For this simulation, both mechanisms were assumed to contribute currents and corresponding weight updates that increase exponentially beyond the relevant threshold voltage (source-drain voltage for injection and oxide voltage for tunneling). For simplicity, the exponential coefficients were taken to be the same for injection and tunneling although they differ in actual circuits. Likewise, tunneling and injection thresholds for these simulations were selected to balance the positive and negative weight updates and represent theoretical, rather than experimentally derived, estimates. More realistic empirical models for injection and tunneling have been developed and are being used to inform the development of a next generation neural network based on these circuits.

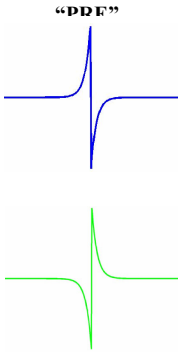


Figure 5.4: Ideal discrete PRE and POST synaptic spike waveforms.

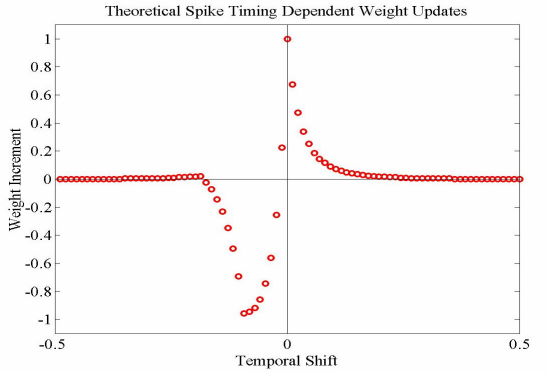


Figure 5.5: Simulated STDP as a function of biphasic mirror-image input waveforms. Computed weight update is shown in red.

Using the circuit shown in Figure 5.6, we first experimentally characterized the performance of the 2TS using floating gate test structures that contained equivalent, but differently sized, transistors to the 2TS. Leaving all unused connections on these test structures (one poly control capacitor, one MOSCAP, and one NMOS follower) floating we asserted the biphasic waveforms as shown in Fig. 5.7 (a) repeatedly at 100 Hz with PRE occurring at a small, fixed (~20 degrees) phase shift ahead of POST. Over 20 s, with an applied peak-to-peak voltage of approximately 10V, we observed the spiking output of the signal at the integration node shown in Figure 5.7 (b), illustrating a small, but significant weight increment over time; we have highlighted the increment with a red trend line that follows the increase. Thus we see positive Hebbian learning using the biphasic PRE and POST inputs.

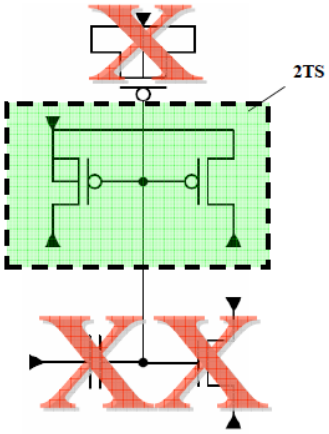
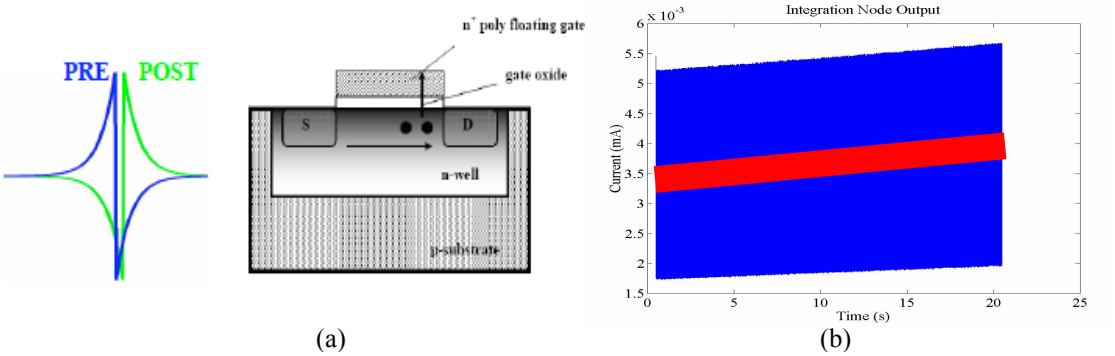


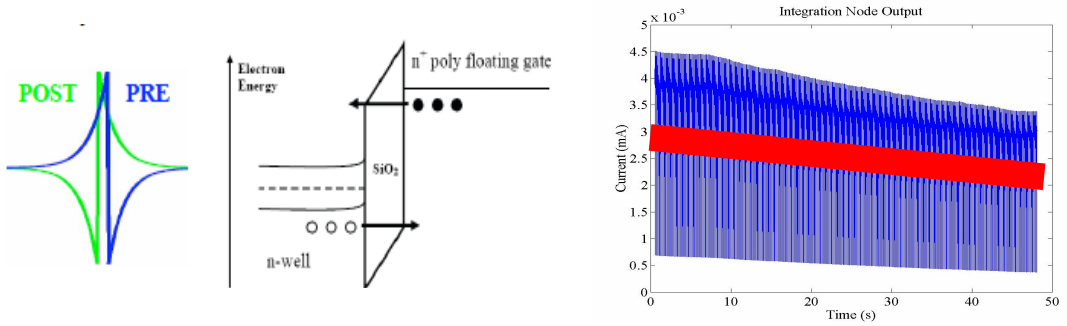
Figure 5.6 Equivalent, but differently-sized, 2TS test structure.



(a) (b)

Figure 5.7: (a) Cartoon of "pre"-"post" overlap for potentiation and hot electron injection weight update; (b) Measured output of circuit integration node as a function of successive positive weight updates. Trend line shown in red.

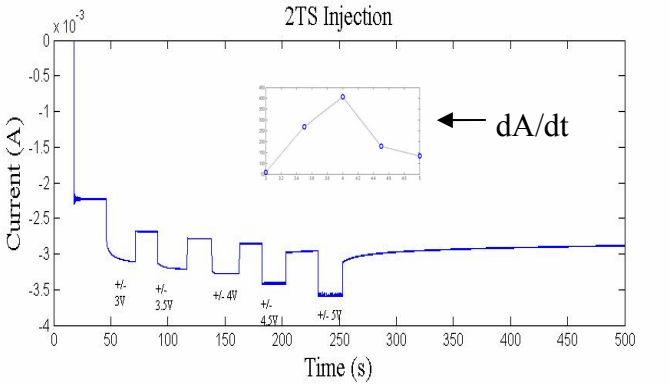
When my colleague, Timir Datta, amplified the PRE and POST signals slightly, to approximately 13 volts peak-to-peak, and asserted them across the 2TS at 10 Hz, with POST preceding PRE by a similar phase shift over 50s, he measured the spiking output of the signal at the integration node shown in Figure 5.8. This data reflects a small, but significant weight decrement over time, consistent with negative Hebbian learning. Thus we have shown proof-of-principle Hebbian learning using the derived biphasic waveforms.



(a) (b)

Figure 5.8: (a) Cartoon of "pre"-"post" overlap for depression and FN tunneling weight update; (b) Measured output of circuit integration node as a function of successive negative weight updates. Trend line shown in red.

I am presently characterizing the full performance of the fabricated 0.5 µm 2TS structures, by repeatedly tunneling and injecting charge carriers off of and onto the floating node of the 2TS, to attain accurate dynamic programming of the circuit that approximates the better theoretical data from new models. Figure 5.9 represents the results of one such test – the drain current at "i_node" for the 2TS is shown as a function of repeated hot electron injection cycles at drain voltages from -3 to -5 V.



5.9: 2TS current increases owing to successive hot electron injection weight updates. Inlay shows dA/dt for each pulsed drain voltage asserted.

The two transistor synapse exhibits theoretical and measured spike timing dependent plasticity in response to biphasic neural waveforms. I intend to continue this work by fully characterizing the fabricated 2TS circuits, and by further investigating the performance of the 2TS in neural network applications for unsupervised learning that may one day be incorporated into implantable neural prosthetics for closed-loop control, e.g. of sensorimotor integration and feedback.

5.2 Neural Network Implementation

One particularly interesting engineering application that leverages this technology is unsupervised pattern recognition. Figure 5.10 illustrates the architecture of a simple Hebbian learning neural network incorporating the 2TS.

To demonstrate the potential of such a network, I developed a MATLAB learning and pattern recognition simulation whose architecture corresponds roughly with fabricated circuit components. First, input vector, $p_1 \ldots p_n$, represents pre-synaptic signals, which may be neural spikes, or as here, simple vectors representing

alphanumeric code. Second, weight matrix, $\underline{W}$, could correspond with an array of 2TS synapses. Finally, distance estimators programmed with distinct template classes, T_1 .. T_m, may be used to generate post-synaptic "spikes" when computed template matches occur.

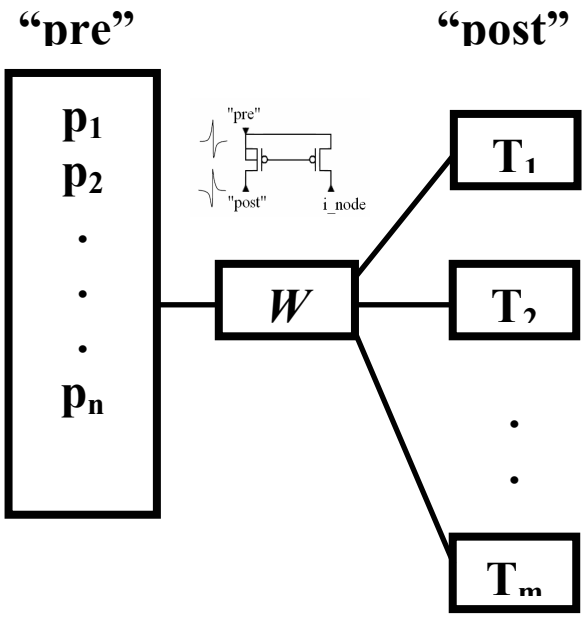


Figure 5.10: Block diagram of Hebbian learning system based on 2TS as synapse.

For the particular simulation shown here, unbiased templates were programmed using noisy versions of ideal templates shown in Fig. 5.11 (a). In circuits, this corresponds with floating gate template programming using hot electron injection and FN tunneling pulses. By iteratively asserting even noisy versions of the signal to update the template vector and employing some thresholded distance metric (such as variance estimation) to end the updates, it is possible to theoretically program to nearly arbitrary precision. Mismatch and process variation, plus circuit and signal noise impose real world limits on the precision of such circuits, as

discussed in greater detail in Chapter 4. Furthermore, unsupervised template programming is also possible, using the methods described in the preceding Chapter.

In any event, once the templates are programmed, pre-synaptic input signals are weighted and correlated with each of the templates to assess whether there is a match. If so, the corresponding template sends a post-synaptic signal indicating a match that simultaneously reinforce and attenuates synapses in proportion to their contribution to the template match. In such a fashion, the weight matrix, or 2TS array, is proportionally strengthened and weakened according to the coincidence of pre- and post- synaptic activity. Moreover, after some number of iterations of programming, the weights themselves correspond with templates so that merely convolving incoming signals with the synaptic array should result in proper classification.

In this case, I used a simple supervised three template classification in order to illustrate the process. The ideal templates represent the letters U, M and D, as shown in Fig. 5.11(a) and each template comprises a vector of 1's and 0's corresponding with dark and light pixels, respectively. Noisy variants of these ideal templates are programmed into analog memory by repeatedly asserting them, comparing them against stored values, and updating until an arbitrary minimum distance is reached; this results in the non-ideal stored templates shown in Fig. 5.11 (b).

Once the templates have been programmed, noisy variants of each of the ideal templates are asserted and weights are updated when template matches occur in proportion to their contribution to the match. This represents competitive Hebbian learning.

(a)

(b)

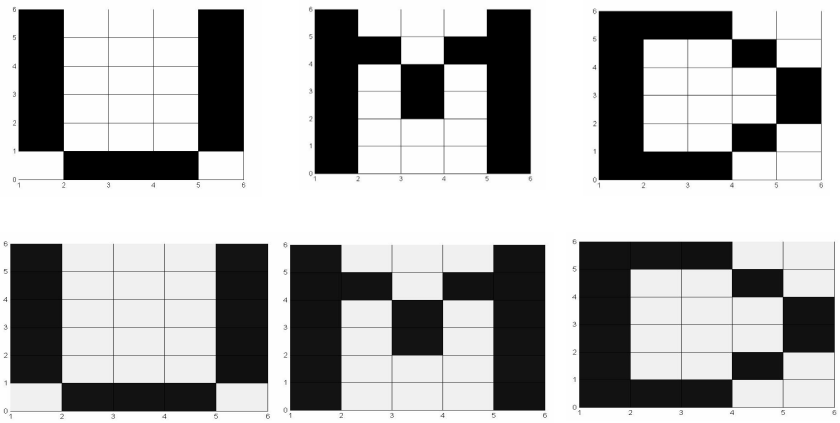


Figure 5.11: (a) graphical plot of ideal template; (b) plot of noisy programmed template.

As in biology, if a post-synaptic spike fires, the contributing synapses are proportionally strengthened and non-contributing synapses are weakened. Finally, once the weight matrix has been trained, we see robust pattern recognition demonstrated over repeated trials, as shown in Fig. 5.12 where the two patterns on the left are mapped onto the correct identifications on the right using the trained weight matrix.

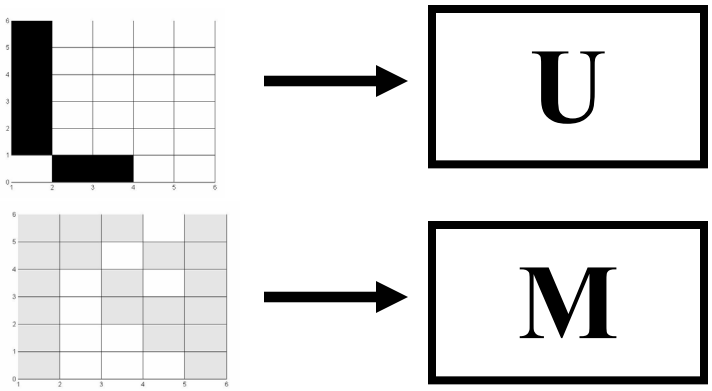


Figure 5.12: Two examples of pattern recognition using the trained NN. In the first, a partial letter is correctly identified. In the second, a noisy and attenuated letter is also correctly identified by the trained network.

In addition to the specific example presented above, it is possible to employ a network of 2TS to register and learn correlations between extracted features of, *e.g.*, neural spikes, in real-time. Previously reported architectures for extracting salient features of neural action potentials and mapping them to biphasic spike trains exist [173]. Building upon this work and classical neural network theory I have begun to evaluate the performance of 2TS networks for unsupervised spike sorting.

Chapter 6: Conclusions

6.1 Summary

We have identified and explored the principal original contributions of this dissertation: (1) programmable electrode arrays for enhanced electrophysiological recording and for directing nerve cell growth; (2) integrated image sensors for the unsupervised detection of significant biological events; (3) ultra-low power, programmable floating gate template matching circuits for the detection and classification of neural action potentials; and (4) a two transistor synapse for the compact hardware implementation of silicon learning. We have further illustrated how these contributions fit into the context of and advance the field. It is believed that these contributions represent enabling technologies for integration with the next generation of implantable neural prosthetics. As such, it is hoped that the work presented here will be used to aid in the restoration of lost sensory and motor function among those individuals suffering from debilitating neural injury and disease.

6.2 Future Directions

In each of the principal research directions we have explored, I anticipate developing the relevant circuits and systems further. An arbitrated AER contact imaging array for neural sensing has been submitted for fabrication; when it returns I plan to characterize its performance using real biological cells. Likewise, in conjunction with colleagues in the bioengineering department and at the National Institutes of Health, I have also planned a number of *in vitro* experiments for further testing of the programmable electrode arrays and to demonstrate on-chip

galvanotropism. In addition, the spike sorting architectures I designed are a mere first step along the path to ultra-low-power mixed signal architectures for implantable neural prosthetics, with integrated spike train decoding architectures. Finally, I am developing hardware networks of the 2TS circuit for unsupervised on-chip learning and pattern recognition.

This is the beginning, not the end.

Bibliography

[1] F. Shen, et al., "Adeno-Associated Viral Vector-Mediated Hypoxia-Inducible Vascular Endothelial Growth Factor Gene Expression Attenuates Ischemic Brain Injury After Focal Cerebral Ischemia in Mice," *Stroke*, vol. 37, Oct. 2006, pp. 1-6.

[2] M Laplaca, C. Simon, G. Prado, D. Cullen, "CNS injury biomechanics and experimental models," *Prog Brain Res.* 2007, vol. 161, pp. 13-26.

[3] C. Lobsiger, S. Boillée, D. Cleveland, "Toxicity from different SOD1 mutants dysregulates the complement system and the neuronal regenerative response in ALS motor neurons," *Proc Natl Acad Sci* USA, May 2007, vol. 104, no. 18, pp. 7319-326.

[4] N. Fitzsimmons, W. Drake, T. Hanson, M. Lebedev, and M. Nicolelis, "Primate reaching cued by multichannel spatiotemporal cortical microstimulation," *J Neurosci.* May 2007, vol. 27, no. 21, pp. 5593-5602.

[5] G. Santhanam, M. Linderman, V. Gilja, A. Afshar, T. Meng, K. Shenoy, "HermesB: A Continuous Neural Recording System for Freely Behaving Primates," *IEEE Transactions on Biomedical Engineering,* vol. 54, no. 11, Nov. 2007, pp. 2037-2050.

[6] J. Mavoori, A. Jackson, C. Diorio, E. Fetz, "An autonomous implantable computer for neural recording and stimulation in unrestrained primates," *J. Neurosci. Methods*, vol. 148, 2005, pp. 71-77.

[7] http://www.ninds.nih.gov/funding/research/npp/index.htm.

[8] Honghao Ji, D. Sander, A. Haas, P. Abshire, "Contact Imaging: Simulation and Experiment," *TCAS-I*, vol. 54, no. 8, Aug. 2007, pp. 1698-1710.

[9] H. Ji, D. Sander, A. Haas, P. Abshire, "A CMOS contact imager for locating individual cells**,"** *ISCAS* 2006, pp. 3357-3360.

[10] A. Haas., S. Williams, M. Cohen, P. Abshire, "Dark address event representation imager," *MWCAS***,** 2005, pp. 388-391.

[11] A.Haas, "Adaptive Image Sensor for Optical Spike Detection," *IEEE/NIH Life Science Systems & Applications Workshop*, April 2009, 4 pages.

[12] S. Prakash, N. Nelson, M. Urdaneta, A. Haas, V. Jeng, E. Smela, and P. Abshire, "BioLabs-On-A-Chip: Monitoring Cells Using CMOS Biosensors," *IEEE/NLM Life Science Systems & Applications Workshop*, Bethesda, MD, 2006, 2 pages.

[13] A.Haas, "Programmable High Density CMOS Microelectrode Array," *IEEE SENSORS Conference*, 2008, pp. 890-893.

[14] A.Haas, M.Cohen, P.Abshire, "Real-Time Variance Based Template Matching Spike Sorting System," *IEEE/NIH BISTI Life Science Systems & Applications Workshop*, 2007, pp. 104-107.

[15] A. Haas, T. Datta, et al., "Two Transistor Synapse with Spike Timing Dependent Plasticity," http://www.lib.umd.edu/drum/handle/1903/8650.

[16] W. Croft, *Under the microscope: a brief history of microscopy*, World Scientific, 2006, pp. 6-7.

[17] C. Murphy, et al. "Gold Nanoparticles in Biology: Beyond Toxicity to Cellular Imaging" *Acc. Chem. Res.*, Aug. 2008, vol. 41, no. 12, pp. 1721-1730.

[18] C. Darby, "Uniquely insidious: Yersinia pestis biofilms," *Trends Microbiol.* Apr. 2008 Apr, vol. 16, no. 4, pp.158-64.

[19] G. Aston-Jones, S. Chen, Y. Zhu, M. Oshinsky, "A neural circuit for circadian regulation of arousal" *Nat Neurosci.*, Jul. 2001, vol. 4, no. 7, pp. 732-38.

[20] http://www.olympusamerica.com/seg_section/product.asp?product=1023.

[21] G. Panigrahi, "Charge-Coupled Memories for Computer Systems," *Computer*, vol. 9, no. 4, Apr. 1976, pp. 33-42, citing original work by Smith, Boyle and Thompsett.

[22]http://www.usa.canon.com/consumer/controller?act=ModelInfoAct&fcategoryid=145&modelid=15262.

[23] http://www.apple.com/iphone.

[24] E. Fossum, A. Krymski, "High speed CMOS imaging," *Digest of the LEOS*, Summer Topical Meetings, July 2000, pp. I3-I4.

[25] http://www.tau.ac.il/~phchlab/experiments/Sucrose/Photodiode.jpg.

[26] http://www.astro.virginia.edu/class/whittle/astr1230/im/em-op-spectrum.gif.

[27] T. Hendel, M. Mank, B. Schnell, O. Griesbeck, A. Borst, D. Reiff, "Fluorescence changes of genetic calcium indicators and OGB-1 correlated with neural activity and calcium in vivo and in vitro," *J. Neurosci*., Jul. 2008, vol. 28, no. 29, pp. 7399-7411.

[28] http://members.misty.com/don/sipdresp.gif.

[29] Y. Shang, W. Zhang, Y. Guan, X. Tan, "Research on a method to extend dynamic range of CMOS APS," Int'l Workshop Imaging Systems and Techniques, Sept. 2008, pp. 212 -216.

[30] D. Ng, T. Tokuda, S. Shiosaka, Y. Tano, K. Ohta, "Implantable Microimagers Sensors, 2008, vol. 8, pp. 3183-3204.

[31] K. Boahen, "Point-to-Point Connectivity Between Neuromorphic Chips using Address-Events," *IEEE TCAS II, Analog and Digital Signal Processing*, vol. 47, no. 5, May 2000, pp. 416-434.

[32] W-L Zhou, et al., "Intracellular Long Wavelength Voltage-Sensitive Dyes for Studying the Dynamics of Action Potentials in Axons and Thin Dendrites," *J. Neurosci Methods*, Aug. 2007, vol. 164, no. 2, pp. 225–239.

[33] M. Dandin, P. Abshire, E. Smela, "Optical filtering technologies for integrated fluorescence sensors," *Lab Chip*, 2007, vol. 7, pp. 955-977.

[34] http://www.olympusmicro.com/primer/java/jablonski/jabintro/index.html.

[35] H. Ji, M. Dandin, E. Smela, P. Abshire, "Integrated Fluorescence Sensing for Lab-on-a-chip Devices," IEEE*/NLM Life Science Systems and Applications Workshop* , Bethesda, Maryland, 2006, pp. 1-2.

[36] E. Kandel, J. Schwartz, T. Jessell, *Principles of Neural Science*, 4th ed., The McGraw-Hill Companies, Inc., 2000.

[37] W. Cowan, T. Sudhof, C. Stevens, *Synapses,* The Johns Hopkins University Press, 2001.

[38] I. Brouk, A. Nemirovsky, Y. Nemirovsky, "Analysis of noise in CMOS image sensor," *COMCAS*, May 2008, pp. 1-8.

[39] O. Yadid-Pecht, E. Fossum, "Wide intrascene dynamic range CMOS APS using dual sampling," *IEEE Transactions on Electronic Devices*, vol. 44, no. 10, Oct. 1997, pp. 1721-23.

[40] C-B. Chien, J. Pine, "Voltage-sensitive dye recording of action potentials and synaptic potentials from sympathetic microcultures," *Biophys. J.*, vol 60, Sept. 1991, pp. 697-711.

[41] J. Burrone, Z. Li, V. Murthy, "Studying vesicle cycling in presynaptic terminals using the genetically encoded probe synaptopHlourin," *Nature Protocols*, vol. 1, no. 6, 2006, pp. 2970-2978.

[42] G. Indiveri, "A low-power adaptive integrate-and-fire neuron circuit,"

ISCAS, May 2003, pp. 25-28.

[43] F. Rieke, D. Warland, R. de Ruyter van Steveninck, W. Bialek, *Spikes: Exploring the Neural Code*, MIT Press, 1999.

[44] G. de Polavieja, A. Harsch, I. Kleppe, H. Robinson,M. Juusola, "Stimulus History Reliably Shapes Action Potential Waveforms of Cortical Neurons," *The Journal of Neuroscience*, June 2005, vol. 25, no. 23, pp. 5657–5665.

[45] C. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley, 1989.

[46] http://www.odec.ca/projects/2004/syed4s0/public_html/action.jpg.

[47] R. Plonsey, R. Barr, *Bioelectricity: A Quantitative Approach*, Plenum Press, 1988.

[48] M. Mahowald, R. Douglas, "A Silicon Neuron," *Nature*, vol. 354, Dec. 1991.

[49] E. Farquhar, P. Hasler, "A Bio-Physically Inspired Silicon Neuron," *TCAS-I*, vol. 52, no. 3, Mar. 2005, pp. 477-488.

[50] S. Saïghi, J. Tomas, Y. Bornat, S. Renaud, "A Conductance-Based Silicon Neuron with Dynamically Tunable Model Parameters," *EMBS*, Mar. 2005, pp. v-viii.

[51] Sakmann, B., Neher, E*., Single-Channel Recording*, 2d ed., Plenum Press, 1995.

[52] N. Reeves, et al., "Integrated MEMS Structures and CMOS Circuits for Biolelectronic Interface with Single Cells," *ISCAS 2004*, pp. 673-676.

[53] U. Frey, C. Sanchez-Bustamante, T. Ugniwenko, F. Heer, J. Sedivy, S. Hafizovic, B. Roscic, M. Fussenegger, A. Blau, U. Egert, A. Hierlemann, "Cell Recordings with a CMOS High-density Microelectrode Array," *IEEE EMBS*, Aug. 2007, pp. 167-170.

[54] B. Eversmann, et al., "A 128x128 CMOS Biosensor Array for Extracellular Recording of Neural Activity," *Solid State*, vol. 38, no. 12, Dec. 2003, pp. 2306-2317.

[55] R. Kelly, et al., "Comparison of Recordings from Microelectrode Arrays and Single Electrodes in the Visual Cortex," *The Journal of Neuroscience*, Jan. 2007, vol. 27, no. 2, pp. 261–264.

[56] D. Kipke, D. Pellinen, R. Vetter, "Advanced neural implants using thin-film polymers," *ISCAS 2002*, pp. 173-176.

[57] R. Harrison, C. Charles, "A Low-Power Low-Noise CMOS Amplifier for Neural Recording Applications," *IEEE JSSC*, 2003, vol. 38, pp. 958-965.

[58] L.A. Geddes, *Electrodes and the Measurement of Bioelectric Events*, Wiley-Interscience, 1972.

[59] S. Ingebrandt, et al., "Investigation of extracellular signal shapes recorded by planar metal microelectrodes and field-effect transistors, *IEEE Sensors*, 2005.

[60] D. Borkholder, *Cell Based Biosensors Using Microelectrodes*, PhD Dissertation, Stanford University, November, 1998.

[61]http://www.cvphysiology.com/Arrhythmias/Ventricular%20action%20potential.gif.

[62] http://www.fysiologie.be/images/pictures_onderzoek/Roeland/Roelan1.gif.

[63] http://iac-usnc.org/Methods/ wholecell/equipment.html.

[64] K. Rahimi, C. Diorio, C. Hernandez, M. Brockhausen, "A Simulation Model for Floating-Gate MOS Synapse Transistors," *ISCAS 2002*, pp. 532-35.

[65] L. Hinkle, C. McCaig, K. Robinson, "The Direction of Growth of Differentiating Neurones and Myoblasts From Frog Embryos in an Applied Electric Field," *J. Physiol.*, 1981, vol. 314 pp. 121-135.

[66] N. Patel, M-M Poo, "Orientation of Neurite Growth by Extracellular Electric Fields," *J. Neuroscience*, 1982, vol. 2, no. 4, pp. 483-96.

[67] C.D. McCaig, "On the Mechanism of Nerve Galvanotropism," *Biol. Bull*. 1989, vol. 176, no. 5, pp. 136-139.

[68] C.D. McCaig, "Nerve branching is induced and oriented by a small applied electric field," *J. Cell Science*, 1990, vol. 95, pp. 605-615.

[69] C.D. McCaig, "Nerve growth in a small applied electric field and the effects of pharmacological agents on rate and orientation," *J. Cell Science*, 1990, vol. 95, pp. 617-622.

[70] C. Schmidt, V. Shistrit, E. Furnish, R. Langer, "Electrical Stimulation of Neurite Outgrowth and Nerve Regeneration," *l7th Southern Biomedical Engineering Conference*, 1998, pp. 117.

[71] A Rajnickek, K. Robinson, C.D. McCaig, "The Direction of Neurite Growth in a Weak DC Electric Field Depends on the Substratum: Contributions of Adhesivity and Net Surface Charge," *Dev. Biology*, 1998, vol. 203, pp. 412–423.

[72] C.D. McCaig, A. Rajnicek, B. Song, M. Zhao, "Controlling Cell Behavior Electrically: Current Views and Future Potential," *Physiol. Rev*., 2005, vol. 85, pp. 943-978.

[73] A. Rajnicek, L. Foubister, C.D. McCaig, "Temporally and spatially coordinated roles for Rho, Rac, Cdc42 and their effectors in growth cone guidance by a physiological electric field / Growth cone steering by a physiological electric field requires dynamic microtubules, microfilaments and Rac-mediated filopodial asymmetry," *J. Cell Science*, 2006, vol. 119, pp. 1723-35, 1736-1745.

[74] D. Bohnert, S. Purvines, S. Shapiro, R. Borgens, "Simultaneous application of two neurotrophic factors after spinal cord injury," *J. Neurotrauma*, 2007, vol. 24, no. 5, pp. 846-863.

[75] S. Prasad, et al., "Electric Field Assisted Patterning of Neuronal Networks for the Study of Brain Functions," *Biomedical Microdevices*, vol. 5, no. 2, June 2003, pp. 125-137.

[76] http://nmrc.bu.edu/tutorials/motor_units/decomp.html.

[77] J. Carmena, M. Lebedev, R.Crist, J. O'Doherty, D.Santucci, D. Dimitrov, P. Patil, C. Henriquez, and M. Nicolelis, "Learning to control a brain-machine interface for reaching and grasping by primates," *PLoS Biology*, Nov. 2003, vol. 1, no. 2, pp. 193-208.

[78] http://www.cyberkineticsinc.com/content/medicalproducts/braingate.jsp.

[79] A. Sodagar, K. Wise, K. Najafi, "A Fully Integrated Mixed-Signal Neural Processor for Implantable Multichannel Cortical Recording," *IEEE TBME*, vol. 54, no. 6, June 2007, pp. 1075-1088.

[80] R. Harrison et al., "A Low-Power Integrated Circuit for a Wireless 100-Electrode Neural Recording System," *IEEE JSSC*, vol. 42, no. 1, Jan. 2007, pp. 123-133.

[81] R. Harrison, G. Santhanam, K. Shenoy, "Local Field Potential Measurement with Low-Power Analog Integrated Circuit," *IEEE EMBS*, Sept. 2004, pp. 4067-70.

[82] Kyung H. Kim and Sung J. Kim, "Neural Spike Sorting Under Nearly 0-dB Signal-to-Noise Ratio Using Nonlinear Energy Operator and Artificial Neural-Network Classifier," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 10, October 2000.

[83] Z. Zumsteg, C. Kemere, S. O'Driscoll, G. Santhanam, R. Ahmed, K. Shenoy, T. Meng, "Power Feasibility of Implantable Digital Spike Sorting Circuits for Neural

Prosthetic Systems," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Sept. 2005, vol. 13, no. 3.

[84] Z. Zumsteg, R. Ahmed, K. Shenoy, T. Meng, "Power Feasibility of Implantable Digital Spike-Sorting Circuits for Neural Prosthetic Systems," *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, San Francisco, CA, Sept. 1-5, 2004.

[86] P. Zhang, J. Wu, Yi Zhou, P. Liang, J. Yuan, "Spike sorting based on automatic template reconstruction with a partial solution to the overlapping problem," *Journal of Neuroscience Methods*, 2004, vol. 135, pp. 55-65.

[87] J. Letelier, P. Weber, "Spike sorting based on discrete wavelet transform coefficients," *Journal of Neuroscience Methods*, 2000, vol. 101, pp. 93-106.

[88] Kyung H. Kim, "A Fully-Automated Neural Spike Sorting Based on Projection Pursuit and Gaussian Mixture Model," *Proceedings of the 2nd International IEEE EMBS Conference on Neural Engineering*, Arlington, VA, March 16-19, 2005.

[89] T. Hermle, C. Schwarz, M, Bogdan, "Employing ICA and SOM for spike sorting of multielectrode recordings from CNS," *Journal of Physiology*, Paris 98 (2004), pp. 349-356.

[90] G. Wang, P. Liang, "Method for Robust Spike Sorting with Overlap Decomposition," *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, Shanghai, China, Sept. 1-4, 2005.

[91] C. Rogers, J. Harris, "A Low-Power Analog Spike Detector for Extracellular Neural Recordings," *ICECS*, Dec. 2004, pp. 290-293.

[92] R. J. Vogelstein, et al., "Spike Sorting with Support Vector Machines," *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, San Francisco, CA, Sept. 1-5, 2004.

[93] T. Horiuchi, T. Swindell, D. Sander, P. Abshire, "A Low-Power CMOS Neural Amplifier with Amplitude Measurements for Spike Sorting," *ISCAS 2004*.

[94] P. Zhang, Jin Wu, Yi Zhou, Pei-Ji Liang, Jing-Qi Yuan, "Spike Sorting in Multi-Channel Extracellular Recordings of Retinas, *IEEE International Conference Neural Networks & Signal Processing*," Nanjing China, Dec. 14-17, 2003.

[95] M. Chelaru, Mandar S. Jog, "Spike source localization with tetrodes," *Journal of Neuroscience Methods*, 142 (2005) 305-315.

[96] C. Bossetti, J. Carmena, M. Nicolelis, P. Wolf, "Transmission Latencies in a Telemetry-Linked Brain-Machine Interface," *IEEE Transactions on Biomedical Engineering,* June, 2004, vol. 51, no. 6.

[97] B. Gosselin, M. Sawan, "An ultra low-power CMOS action potential detector," *ISCAS 2008*, pp. 2733-2736.

[98] Kyung H. Kim and Sung J. Kim, "Method for Unsupervised Classification of Multiunit Neural Signal Recording Under Low Signal-to-Noise Ratio," *IEEE Transactions on Biomedical Engineering*, Apr. 2003, vol. 50, no. 4.

[99] O. Chibirova, et al., "Unsupervised Spike Sorting of extracellular electrophysiological recording in subthalamic nucleus of Parkinsonian patients," *BioSystems*, 2005, vol. 79, pp.159-171.

[100] F. Wood, et al., "On the Variability of Manual Spike Sorting," *IEEE Transactions on Biomedical Engineering*, June 2004, vol. 51, no. 6.

[101] E. Hulata, R. Segev, E. Ben-Jacob, "A method for spike sorting and detection based on wavelet packets and Shannon's mutual information," *Journal of Neuroscience Methods*, 2002, vol. 117, pp. 1-12.

[102] Reid R. Harrison, "A Low-Power Integrated Circuit for Adaptive Detection of Action Potentials in Noisy Signals," *Proceedings of the 25th Annual International Conference of the IEEE EMBS*, Cancun, Mexico, Sept. 17-21, 2003, pp. 3325-28.

[103] F. Wood, S. Goldwater, M. Black, "A Non-Parametric Bayesian Approach to Spike Sorting," *Proceedings of the 28th IEEE EMBS Annual International Conference*, New York, NY, Aug. 30-Sept. 3, 2006.

[104] E. Hulata, R. Segev, Y. Shapira, M. Benveniste, and E. Ben-Jacob, "Detection and Sorting of Neural Spikes Using Wavelet Packets," *Physical Review Letters*, Nov. 2000, vol. 85, no. 21.

[105] Yevgeny Perelman and Ran Ginosar, "An Integrated System for Multichannel Neuronal Recording With Spike/LPF Separation, Integrated A/D Conversion and Threshold Detection," *IEEE Transactions on Biomedical Engineering*, Jan. 2007, vol. 54, no. 1.

[106] Michael S. Lewicki, "A review of methods for spike sorting: the detection and classification of neural action potentials," *Network: Comput. Neural Syst*., 1988, vol. 9, R53-R78.

[107] C. Pouzat, O. Mazor, G. Laurent, "Using noise signature to optimize spike-sorting and to assess neuronal classification quality," *Journal of Neuroscience Methods*, 2002, vol. 122, pp. 43-57.

[108] X. Yang, S. Shamma, "A Totally Automated System for the Detection and Classification of Neural Spikes," *IEEE Transactions on Biomedical Engineering*, Oct. 1988, vol. 35, no. 10.

[109] R. Quian Quiroga, Z. Nadasdy, Y. Ben-Shaul, "Unsupervised Spike Detection and Sorting with Wavelets and Superparamagnetic Clustering," *Neural Computation,* 2004, vol. 16, pp. 1661-1687.

[110] R. Quian Quiroga, O.W. Sakowitz, E. Basar, M. Schurmann, "Wavelet Transform in the analysis of the frequency composition of evoked potentials," *Brain Research Protocols*, 2001, vol. 8, pp. 16-24.

[111] M. Laubach, "Wavelet-based processing of neuronal spike trains prior to discriminant analysis," *Journal of Neuroscience Methods*, 2004, vol. 134, pp. 159-168.

[112] R. Brychta, et al., "Wavelet Methods for Spike Detection in Mouse Renal Sympathetic Nerve Activity," *IEEE Transactions on Biomedical Engineering*, Jan. 2007, vol. 54, no. 1.

[113] T. Horiuchi, D. Tucker, K. Boyle, and P. Abshire, "Spike Discrimination Using Amplitude Measurements with a Low-Power CMOS Neural Amplifier," *ISCAS*, May 27-30, 2007, pp. 3123 – 3126.

[114] Kyung H. Kim and Sung J. Kim, "Classification of Neural Spike Under Nearly 0 dB Signal-to-Noise Ratio," *Proceedings of the First Joint BMES/EMBS Conference*, 1999, Atlanta, GA.

[115] J. Choi, H. Jung, T. Kim, "A New Action Potential Detector Using the MTEO and Its Effects on Spike Sorting Systems at Low Signal-to-Noise Ratios," *IEEE Transactions on Biomedical Engineering*, Apr. 2006, vol. 53, no. 4.

[116] K.V. Shenoy et al., "Neural prosthetic control signals from plan activity," *NeuroReport*, 2003, vol. 14, pp. 591-596.

[117] J. Carmena, M. Lebedev, R.Crist, J. O'Doherty, D.Santucci, D. Dimitrov, P. Patil, C. Henriquez, and M. Nicolelis, "Learning to control a brain-machine interface for reaching and grasping by primates," *PLoS Biology*, Nov. 2003, vol. 1, no. 2, pp. 193-208.

[118] Y. Gao, M. Black, E. Bienenstock, S. Shoham, and J. Donoghue, "Probabilistic inference of arm motion from neural activity in motor cortex," *Advances in Neural Information Processing Systems*, 2002, vol. 14, pp. 221-228.

[119] M. Serruya, N. Hatsopoulos, L. Paninski, M. Fellows and J. Donoghue, "Instant neural control of a movement signal," *Nature*, Mar. 2002, vol. 416, pp. 141-142.

[120] D. Taylor, S. Helms-Tillery, A. Schwartz, "Direct cortical control of 3d neuroprosthetic devices," *Science*, June 2002, vol. 296, no. 3, pp. 1829-1832.

[121] J. Wessberg, C. Starnbaugh, J. Kralik, P. Beck, M. Laubach, J. Chapin, J. Kim, S. Biggs, M. Srinivasan, and M. Nicolelis, "Real-time prediction of hand trajectory by ensembles of cortical neurons in primates," *Nature*, 2000, vol. 408, pp. 361-365.

[122] I.N. Bankman and S.J. Janselewitz, "Neural waveform detector for prosthesis control," *Proc. 17th Ann. Conf. IEEE EMBS*, 1995, pp. 963-964.

[123] J.P. Stitt, R.P. Gaumond, J.L. Frazier, and F.E. Hanson, "A comparison of neural spike classification techniques," Proc. *19th international conf. IEEE/EMBS* , 1997, Chicago, IL, pp. 1092-1094.

[124] A. Owens, T. Denison, H. Versnel, M. Rebbert, M. Peckerar, S. Shamma, "Multi-electrode array for measuring evoked potentials from surface of ferret primary auditory cortex," *J Neurosci Methods.*,1995, vol. 58, pp. 209-20.

[125] N. Fitzsimmons, W. Drake, T. Hanson. M. Lebedev, M. Nicolelis, "Primate reaching cued by multichannel spatiotemporal cortical microstimulation," *J. Neurosci.*, 2007, vol. 23, pp. :5593-602.

[126] http://www.fcc.gov/oet/info/documents/bulletins/#56.

[127] H. Cho, et al., "A New Template Matching Method using Variance Estimation for Spike Sorting," *IEEE EMBS 2d Int'l. Conf. on Neural Engineering*, March, 2005, pp. 225 – 228.

[128] R. Vollgraf and K. Obermayer, "Improved Optimal Linear Filters for the Discrimination of Multichannel Waveform Templates for Spike-Sorting Applications," *IEEE Signal Processing Letters*, Mar. 2006, vol. 13, No. 3.

[129] T.M. Seese, H. Harasaki, G.M. Saidel, and C.R. Davies, "Characterization of tissue morphology, angiogenesis, and temperature in the adaptive response of muscle tissue to chronic heating," *Lab. Investigation*, 1998, vol. 78, no. 12, pp. 1553-1562.

[130] I.N. Bankman, K.O. Johnson, A. Menkes, S.D. Diamond, and D.O'Shaughnessy, "Automated analyzer for on-line recognition of neural waveforms in extracellular recordings of multiple neurons," *IEEE EMBS*, 1992, vol. 14, pp. 2852-2853.

[131] I.N. Bankman, K.O. Johnson, W. Schneider, "Optimal detection, classification, and superposition resolution in neural waveform recordings," *IEEE Trans. Biomed. Eng.*, 1993, pp. 836-841.

[132] B.C. Wheeler and W.J. Heetderks, "A comparison of techniques for classification of multiple neural signals," *IEEE Trans. Biomed. Eng*., 1982, vol. BME-29, pp. 752-759.

[133] K. Mirfakhraei and K. Horch, "Classification of action potentials in multi-unit intrafascicular recordings using neural network pattern-recognition techniques," *IEEE Trans. Biomed. Eng.*, 1994, vol. 41, pp. 89-91.

[134] K.H. Kim and S.J. Kim, "Method for action potential detection from extracellular neural signal recording with low signal-to-noise ratio," *IEEE Trans. Biomed. Eng.*, Aug. 2003, vol. 50, no. 8, pp. 999-1011.

[135] A. Culhane, M. Peckerar, C. Marrian, "A Neural Net Approach to Discrete Hartley and Fourier Transforms," *TCAS,* May 1989, vol.36, no. 5, pp. 695-703.

[136] F. Ohberg, H. Johansson, M. Bergenheim, J. Pedersen, and M. Djupsjobacka, "A neural network approach to real-time spike discrimination during simultaneous recording from several multi-unit nerve filaments," *J. Neurosci. Meth.*, 1996, vol. 64, pp. 181-187.

[137] J.S. Oghalai, W.N. Street, W.S. Rhode, "A neural network based spike discriminator," *J. Neurosci Methods*,1994; vol. 54, pp. 9-22.

[138] M.F. Sarna, P. Gochin, J. Kaltenback, M. Salganicoff, G.L. Gerstein, "Unsupervised waveform classification for multi-neuron recordings: a real-time, software based system. II. Performance comparison to other sorters," *J Neurosci Methods*, 1988, vol. 25, pp.189-96.

[139] M. Salganicoff, M. Sarna, M. Sax, G.J. Gerstein, "Unsupervised waveform classification for multi-neuron recordings: a real-time, software-based system. I. Algorithms and implementation," *J. Neurosci Meth*, 1988 vol. 25, pp. 181-187.

[140] X. Yang and S.A. Shamma, "A totally automated system for the detection and classification of neural spikes," *IEEE Trans. Biomed. Eng.*, Jan. 1994, vol. 41, pp. 89-91.

[141] G. Zouridakis, D.C. Tam, "Identification of reliable spike templates in multi-unit extracellular recordings using fuzzy clustering," *Comput Methods Programs Biomed*, 2000, vol. 61, pp. 91-8.

[142] G. Zouridakis, D.C. Tam, "Multi-unit spike discrimination using wavelet transforms," *Comput Biol Med* , 1997, vol. 27, pp. 9-18.

[143] Alex Zviagintsev, Yevgeny Perelman, Ran Ginosar, "Low-Power Architectures for Spike Sorting," *Proceedings of the 2nd International IEEE EMBS Conference on Neural Engineering*, Arlington, Virginia, March 16-19, 2005.

[144] S. Shoham, M. Fellows, R. Normann, "Robust, automatic spike sorting using mixtures of multivariate t-distributions," *Journal of Neuroscience Methods*, 2003, vol. 127, pp. 111-122.

[145] R. Harrison, R., Bragg, J.A., Hasler, P., Minch, B.A., Deweerth, Stephen P., "A CMOS Programmable Analog Memory-Cell Array Using Floating-Gate Circuits," *IEEE TCAS-II*, 2001, Vol. 48, No. 1, pp. 4-11.

[146] Rishi Chandra and Lance M. Optican, "Detection, Classification, and Superposition Resolution of Action Potentials in Multiunit Single-Channel Recordings by an On-Line Real-Time Neural Network," *IEEE Transactions on Biomedical Engineering*, Vol. 44, No. 5, May, 1997.

[147] T. Aksenova, et al., "An unsupervised method for sorting neuronal spike waveforms in awake and freely moving animals," *Methods,* 2003, vol. 30, pp. 178-187.

[148] S.A. Jackson, J.C. Killens, B.J. Blalock, "A Programmable Current Mirror for Analog Trimming Using Single-Poly Floating-Gate Devices in Standard CMOS Technology," *TCAS-II*, Jan. 2001, vol. 48, No. 1, pp. 100-102.

[149] D.W. Graham, E. Farquhar, B. Degnan, C. Gordon, P. Hasler, "Indirect Programming of Floating-Gate Transistors," *TCAS-I*, May 2007, vol. 54, No. 5, pp. 951-63.

[150] Z. Nenadic, J.W. Burdick, "Spike Detection Using the Continuous Wavelet Transform," *TBME*, Jan. 2005, vol. 52, no. 1, pp. 74-87.

[151] S. Vlassis, S. Siskos, CMOS Outlier Rejection Circuit, *TCAS-I*, July, 2001, vol. 48, no. 7, pp. 910-914.

[152] Christy L. Rogers, John G. Harris, Jose C. Principe, Justin C. Sanchez, "An Analog VLSI Implementation of A Multi-Scale Spike Detection Algorithm for Extracellular Neural Recordings," *Proceedings of the 2d International IEEE EMBS Conference on Neural Engineering*, Mar. 16-19, Arlington, VA, 2005.

[153] F. Wood, M. Fellows, J.P. Donoghue, M.J. Black, "Automatic Spike Sorting for Neural Decoding," *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, Sept. 2004, San Francisco, CA.

[154] J. Harris, J. Principe, J. Sanchez, Du Chen, C. She, "Pulse-based signal compression for implanted neural recording systems," *ISCAS 2008*, pp. 344-347.

[155] A. Bonfanti, T. Borghi, R., et al., "A low-power integrated circuit for analog spike detection and sorting in neural prosthesis systems," *BioCAS 2008*, pp. 257-260.

[156] M. Chae, et al**.,** **"**A 128-Channel 6mW Wireless Neural Recording IC with On-the-Fly Spike Sorting and UWB Tansmitter," *ISSCC 2008*, 3 pages, pp. 146-603.

[157] T-C Chen, W. Liu, L-G Chen, "VLSI architecture of leading eigenvector generation for on-chip principal component analysis spike sorting system," *EMBS*, Aug. 2008, pp. 3192-3195.

[158] Bradley A. Minch, Paul Hasler, Chris Diorio, *Multiple-Input Translinear Element Networks*, TCAS-II, Vol. 48, No. 1, Jan. 2001, pp. 20-28.

[159] R. Sarpeshkar, R. Lyon, C. Mead, "A Low-PowerWide-Linear-Range Transconductance Amplifier," *Analog Integrated Circuits and Signal Processing,* 1997, vol. 13, pp. 123–151.

[160] D. Banks, C. Tomazou, "Low-power high-speed current comparator design," Electronics Letters, vol. 44, no. 3, Jan. 2008, pp. 171-72.

[161] B. Linares-Barranco, T. Serrano-Gotarredona, "On the design and characterization of femtoampere current-mode circuits," *IEEE Journal of Solid State Circuits*, Aug. 2003, vol. 38, no. 8, pp. 1353-1363.

[162] R. Froemke, Y. Dan Y, "Spike-timing-dependent synaptic modification induced by natural spike trains," *Nature*, Mar. 2002, vol. 416, pp. 433–438.

[163] Y. Dan, M. Poo, "Spike timing-dependent plasticity: from synapse to perception," *Physiol. Rev*., July 2006, vol. 86, no. 3, pp. 1033-048, Jul. 2006.

[164] P. Hasler, J. Dugger, "Correlation learning rule in floating gate PFET synapses," *IEEE TCAS-II*, vol. 48, no. 1, Jan. 2001, pp. 65-73.

[165] A. Andreou, K. Boahen, "Synthetic neural circuits using current domain signal representations," *Neural Computation*, 1989, vol. 1, no. 4, pp. 489-501.

[166] D. Watola, D. Gembala, J. Meador, "Competitive learning in asynchronous-pulse-density integrated circuits," *ISCAS*, May 1992, pp. 2216-2219.

[167] J. Meador, A. Wu, C. Cole, N. Nintunze, P. Chintrakulchai, "Programmable impulse neural circuits," *IEEE Transactions on Neural Networks*, Jan. 1991, vol. 2., no. 1, Jan. 1991, pp. 101-109.

[168] Yariv, C. Neugebauer, and A. Agranat, "Programming synapse for neural network applications," U.S. Patent No. 5,353,382, Oct. 4, 1994.

[169] A. Kramer, "Array-based analog computation," *IEEE Micro*, vol. 16, no. 5, Oct., 1996, pp. 20-29.

[170] R. Carmona, F. Jiménez-Garrido, R. Domínguez-Castro, S. Espejo and A. Rodríguez-Vázquez, "A CMOS analog parallel array processor chip with programmable dynamics for early vision tasks," *ESSCIRC 2002*, pp. 371-374.

[171] S-C Liu, R. Mockel, "Temporally learning floating-gate VLSI synapses," *ISCAS* 2008, pp. 2154-157.

[172] G. Indiveri, E. Chicca, R. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, Jan. 2006, pp. 211-221.

[173] C. Rogers, J. Harris, J. Principe, J. Sanchez, "A pulse-based feature extractor for spike sorting neural signals," *IEEE/EMBS CNE*, 2007, pp. 490-493.