An Asynchronous, Low-Power Architecture for Interleaved Neural Stimulation, using Envelope and Phase Information

by

Ji-Jon Sit

Bachelor of Science in Electrical Engineering
Bachelor of Arts in Computer Science
Yale University, 2000

Master of Science in Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2002

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology

May 2007
[June 2007]

Signature of Author.....
Department of Electrical Engineering and Computer Science
May 25, 2007

Certified by.....
Rahul Sarpeshkar
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by.....
Arthur C. Smith
Chairman, Committee on Graduate Students
Department of Electrical Engineering and Computer Science

An Asynchronous, Low-Power Architecture for Interleaved Neural Stimulation, using Envelope and Phase Information

by

Ji-Jon Sit

Submitted to the Department of Electrical Engineering and Computer Science
On May 25, 2007 in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

ABSTRACT

This thesis describes a low-power cochlear-implant processor chip and a charge-balanced stimulation chip that together form a complete processing-and-stimulation cochlear-implant system. The processor chip uses a novel Asynchronous Interleaved Stimulation (AIS) algorithm that preserves phase and amplitude cues in its spectral input while simultaneously minimizing electrode interactions and lowering average stimulation power per electrode. The stimulator chip obviates the need for large D.C. blocking capacitors in neural implants to achieve highly precise charge-balanced stimulation, thus lowering the size and cost of the implant. Thus, this thesis suggests that significant performance, power and cost improvements in the current generation of cochlear implants may be simultaneously possible.

The 16-channel ~90 square mm AIS processor chip was built in a 1.5μm VLSI process and consumed 107μW of power over and above that of its analog spectral processing front end, which consumed 250μW and which has been previously described. The AIS processor was found to faithfully mimic MATLAB implementations of the AIS algorithm. Two perceptual tests of the AIS algorithm with normal-hearing listeners verified that AIS signal reconstructions enabled better melody and speech recognition in noise than traditional envelope-only vocoder simulations of cochlear-implant processing. The average firing rate of the AIS processor was found to be significantly lower than in traditional synchronous stimulators, suggesting that the AIS algorithm and processor can potentially save power and improve hearing performance in cochlear-implant users.

The stimulator chip was built in a 0.7μm high-voltage VLSI process and performed dynamic current balancing followed by a shorting phase. It achieved <6nA of average DC current error, well below the targeted safety limit of 25nA for cochlear-implant patients. On +6 and -9V rails, the power consumption of a single channel of this chip was 47μW when biasing power is shared by 16 channels. It puts out a charge-balanced stimulation pulse whenever it receives an asynchronous input signal from an AIS processor encoding phase information and 7-bit amplitude information, thus making the AIS processor chip and stimulator chip fully compatible in the cochlear-implant system.

The AIS algorithm and charge-balancing circuits described in this work may be useful in other nerve-stimulation prosthetics where good fidelity in input-information encoding, minimization of electrode interactions, low-power strategies for stimulation, and compact charge-balanced stimulation are also important.

Thesis Supervisor: Rahul Sarpeshkar
Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgements

Issac Newton famously wrote to Robert Hooke in 1676, "If I have seen a little further it is by standing on ye shoulders of Giants." Recently I was thinking to myself, I have rather finished my Ph.D. only because I have been carried upon a Giant number of shoulders. Sort of like the paralytic, who was carried by his friends up on a roof, and through the tiles down to Jesus for healing (in Luke 5). I would like to give thanks to these shoulders here.

My friends in my fellowship at MIT and at Park Street Church have all played a crucial part in my "great deliverance," by wrestling with God for me in prayer. I believe God was delighted to encourage all of them by answering their common request for me to finish my thesis and pass my defense with His grace!

I also have much to be thankful for because my research group is without exception, made up of not just gifted, but more importantly, humble and upright individuals. I think it is hard to find this combination anywhere, period. I have learnt a lot from them and count myself blessed to have worked with all of them, past and present – in rough order of when I met them over my 7 years – Heemin and Maziar, Micah and Chris and George, Alex and Baker, Serhii and Lorenzo and Tim, Soumyajit and Keng Hoong, Scott and Pok and Ben. In the same breath I cannot fail to mention our administrative assistants over the years: Travis, Mary, Rose and Gretchen have also a spot in my heart for so much behind-the-scenes support.

Without the guidance and assistance from Andrew Oxenham and Andrea Simonson, I would have been at a loss in how to undertake the research pertaining to psychoacoustics, and the opportunity to work with them at MIT was really fortuitous. Mike Perrott's advice on writing my thesis around the figures, collecting only the absolutely essential data, and improving the clarity of my technical descriptions made a big difference. But what was even more valuable to me is the support he communicated along with all the details. I would probably not have come to MIT if not for the financial support from Advanced Bionics, through Mike Faltys and Van Harrison, and I'm also grateful that through Mike I've had a very smooth road to getting a job there even before graduation.

Almost needless to say, I am very thankful to my thesis supervisor, Rahul Sarpeshkar,

for taking me in, tirelessly searching for funds to support me, and training me up with invaluable research skills. But I appreciate even more his unwavering pursuit of truth and real insight, never never never giving up, and how he sees the lab bench as "talking with God" because he's seen so clearly that all the moments of truth that we can learn from God's other book are revealed not so much at the computer but in the lab.

My family, Mum, Pop and Mae, have never failed to support me with their love and concern for my well-being and future. They have always wished the best for me, and I have not the slightest doubt in my mind about that fact – which has formed the bedrock of my ability to enjoy independence as an adult, and to still pursue child-like dreams wherever they may take me. This firm, unshakable foundation is priceless. My wife's family, Papa, Mummy, and Jeffrey have also been recently grafted to me but are no less precious – their love and support is as if it comes from much further back in time – and in fact I believe it does. This is also the gift of God.

But the biggest gift of God in my time here is my beautiful wife Felicity – her unconditional love no matter how grumpy and unlovable I make myself at times is the most clear and present reflection of God's love for me. I believe that she loves me more than she loves herself – and she has shown it in countless sacrifices of time and energy to support me in everything she's seen me need. I love you, sweetiebear!

God deserves the highest thanks and honor for being the provider of all these blessings and more. He revealed Himself to me one year before grad school, perhaps knowing that it was just when I needed Him most. He is the One who gifted me to be able to attend MIT, granted me success in all I undertook, but above all else, granted me peace in my heart from the forgiveness of my sin in Jesus Christ. Without that peace, I am sure my mind and soul would have been too clouded and dark for me to do anything good. He is the source of all my inspiration, recorded in these pages, so I want all the glory to go to Him!

Does he who implanted the ear not hear?

Psalm 94:9 (NIV)

TABLE OF CONTENTS

1. INTRODUCTION

Many biomedical implants that seek to restore lost sensory function to disabled patients currently convey information to the nerves by pulsatile current stimulation, delivered via an implanted microelectrode array. The most prevalent method of pulsatile stimulation in current-generation implants is:

a) **Synchronous** (at a fixed stimulation rate, set by a fixed carrier frequency)
b) **Amplitude-Modulated** (a slowly-varying envelope of pulses reflects the sensory input intensity)
c) **Time-Interleaved** (no more than 1 or 2 electrodes fire a pulse at any one time)
d) **Biphasic** (charge-balanced with equal positive and negative phases per pulse)

However, several problems have been identified with this current paradigm of neural stimulation. Of primary concern to us is that the fidelity of sensory input information that is conveyed to the nerves within this paradigm may be inherently limited by both (a) and (b). The reasons for this will be outlined in the Background section, and a new stimulation architecture which has the potential to improve the fidelity of information transmission will be presented in section 3.

A major discovery made in the early 90s was that (c), or interleaved stimulation, as part of a strategy called Continuous Interleaved Sampling (CIS), could greatly mitigate undesirable electrode interactions which would take place due to current spreading from nearby electrodes (Wilson, Finley et al. 1991). For the same reason, the proposed stimulation architecture in this thesis also enforces interleaved stimulation as a key requirement.

Finally, this thesis will also address a practical issue related to (d). In order to prevent electrolysis and potential rejection of the implant by the immune system, the positive and negative phases of each current pulse are designed to be precisely charge balanced so as to deliver negligible net charge into the body. Traditionally, a large series blocking capacitor on the order of 100nF was sufficient to perform this function on every electrode channel, but in an effort to decrease the size and cost of the implanted electronics, it has become desirable to remove these DC-blocking capacitors. This thesis includes a current-

drive system which will balance the charge pulses electronically, down to a current error which is considered physiologically safe.

This study focuses particularly on delivering auditory information to the cochlear nerve in cochlear implants (CIs), but could impact any multi-channel stimulation paradigm. In conclusion, this thesis presents a novel architecture for neural stimulation which seeks to improve the fidelity of sensory information when encoded and conveyed to the nervous system, from the point of receiving timing-rich sensory information to the point of driving current pulses to the implant electrodes.

2. BACKGROUND[1]

2.1. Problems with synchronous stimulation

It has been shown that acoustic stimulation of the ear produces much less across-fiber synchrony and much more within-fiber jitter than electrical stimulation (Kiang and Moxon 1972), due to stochastic properties of the inner-hair-cell synapse that are not reproduced in electrical stimulation.

In particular, there is unnecessary synchronization of the neural response to the fixed-rate electrical carrier, despite the carrier containing ***no information***. Physiological studies have shown that many neural artifacts are produced, which are never seen under normal acoustic stimulation. For example, at low-rate (<500-800 Hz) stimulation, there is deterministic entrainment to the carrier (a neural discharge once per stimulus cycle), while at high-rate (>500-800 Hz) stimulation, there are severe distortions in the temporal discharge patterns caused by neural refractoriness (Litvak, Delgutte et al. 2003; Litvak, Smith et al. 2003).

Secondly but equally important, synchronous stimulation inherently excludes the possibility of delivering phase, or fine-timing information to implant users. If there is a significant amount of information which is not carried in the envelope amplitude alone, but also in the fine-time structure (or phase) of a signal, then an AM method of encoding envelope information on a fixed carrier frequency is simply unable to carry that

---

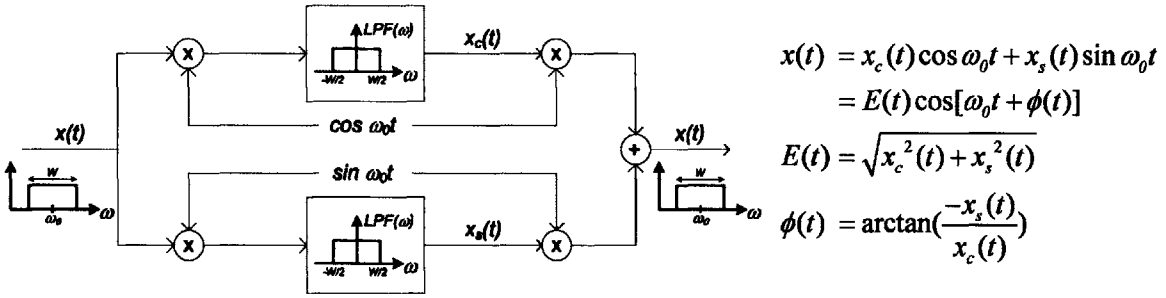
[1] Sections 2.1-2.4, 2.7, all of Chapter 3 and the first paragraph of 4.1 are © 2007 IEEE. Reprinted, with permission, from IEEE Transactions on Biomedical Engineering (Sit, Simonson et al. 2007).

information. A mathematically more rigorous description of the envelope and phase components to a signal is described further in the next section.

All the problems above can be avoided if fixed-rate stimulation is replaced by some form of asynchronous stimulation that also takes the limits of neurobiology into account. For example, the stimulation should refrain from driving the nerves heavily into refractory-rate limited firing, where cross-fiber synchrony becomes pronounced.

2.2. What is phase information?

It has been shown that any bandpass signal x(t) with center frequency ω_0 and bandwidth W can be decomposed into quadrature components $x_c(t)\cdot\cos\omega_0(t) + x_s(t)\cdot\sin\omega_0(t)$ (Siebert 1986) and hence also into envelope and phase components E(t) and $\phi(t)$ respectively, where $x(t) = E(t)\cdot\cos[\omega_0 t + \phi(t)]$ as shown in (Figure 1).



$$x(t) = x_c(t)\cos\omega_0 t + x_s(t)\sin\omega_0 t$$
$$= E(t)\cos[\omega_0 t + \phi(t)]$$
$$E(t) = \sqrt{x_c^2(t) + x_s^2(t)}$$
$$\phi(t) = \arctan\left(\frac{-x_s(t)}{x_c(t)}\right)$$

Figure 1. Quadrature decomposition of a bandpass signal x(t)

As W can be as large as $2\omega_0$, any bandlimited signal can be expressed in the general form of $E(t)\cdot\cos\theta(t)$ where the angle component $\theta(t) = \omega_0 t + \phi(t)$ when ω_0 is specified. The angle component $\theta(t)$ or, equivalently (when ω_0 is defined), the phase component $\phi(t)$ is what we will refer to as the *phase information* in a signal. In the psychoacoustic literature, phase information is commonly referred to as "fine time structure" or "FM (frequency modulation) information," as it can be thought of as how rapidly the instantaneous frequency is being modulated, as opposed to "AM (amplitude modulation) information" in E(t).

Defining $x_H(t)$ as the Hilbert transform (Figure 2) of x(t),

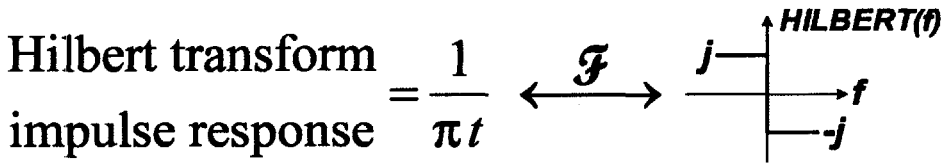


Figure 2. The Hilbert Transform in time and frequency domain representation

$$
\begin{aligned}
x(t) &= x_c(t)\cos\omega_0 t + x_s(t)\sin\omega_0 t \\
x_H(t) &\triangleq x(t) * \frac{1}{\pi t} = \mathcal{F}^{-1}\left[X(f)\cdot HILBERT(f)\right] \\
&= x_c(t)\sin\omega_0 t - x_s(t)\cos\omega_0 t \quad \text{(since the Hilbert transform serves to exchange sine and cosine components)}
\end{aligned}
\tag{1}
$$

If we then define the analytic signal $s(t) = x(t) + j\cdot x_H(t)$,

$$
\begin{aligned}
s(t) &\triangleq x(t) + jx_H(t) \\
&= x_c(t)\left[\cos\omega_0 t + j\sin\omega_0 t\right] - jx_s(t)\left[\cos\omega_0 t + j\sin\omega_0 t\right] \\
&= \left[x_c(t) - jx_s(t)\right]e^{j\omega_0 t}
\end{aligned}
\tag{2}
$$

We should note E(t) and θ(t) can then be obtained *identically* from s(t) as |s(t)| and ∠s(t) respectively:

$$
\begin{aligned}
|s(t)| &= \left|x_c(t) - jx_s(t)\right|\cdot\left|e^{j\omega_0 t}\right| \\
&= \sqrt{x_c^{\,2}(t) + x_s^{\,2}(t)} \qquad = E(t) \\
\angle s(t) &= \angle\left[x_c(t) - jx_s(t)\right] + \angle e^{j\omega_0 t} \\
&= \arctan\left(\frac{-x_s(t)}{x_c(t)}\right) + \omega_0 t \quad = \phi(t) + \omega_0 t \quad = \theta(t)
\end{aligned}
\tag{3}
$$

When analyzing a signal, it is useful to use the Hilbert transform to look at the E(t) and θ(t) components separately, especially when comparing two signals, as they may be similar in one component but not the other.

2.3. How important is phase information for CI users?

Experiments by Smith et al. have shown that the melody in music will be carried primarily in the fine structure, or phase component φ(t), when the number of bandpass analysis channels is smaller than 32 (Smith, Delgutte et al. 2002). The current generation of cochlear implants is limited to 16-24 analysis channels and delivers only E(t) at a fixed rate, discarding φ(t) altogether (Wilson, Finley et al. 1991). Thus it is no surprise that CI users have trouble perceiving tones, much less the melody in a piece of music.

In an attempt to restore tonal perception, which is important in many Asian languages like Mandarin Chinese and Cantonese, researchers have attempted to modulate the

bandpass center frequencies ω_0 by the fundamental frequency F0 in each channel, with significant success in normal-hearing listeners (Lan, Nie et al. 2004). Lan et al.'s results can be interpreted as modulating $\phi(t)$ in each bandpass channel, to better approximate the true phase of the signal.

Furthermore, recent work by Nie et al. using acoustic simulations on normal-hearing listeners shows that speech recognition in noise is significantly improved by the addition of FM information, which is essentially what we refer to as $\phi(t)$, bandlimited to 400 Hz. They hypothesize that additional FM cues allow better segregation of a target sound into a perceptual stream that is more distinct from competing noise (Nie, Stickney et al. 2005).

This may be particularly important under conditions of low spectral resolution, such as those experienced by CI users (Fu, Shannon et al. 1998). An average CI user appears to receive at most 10 usable channels of spectral information, regardless of the number of implanted electrodes (Dorman, Loizou et al. 1998; Friesen, Shannon et al. 2001). So compared to normal-hearing listeners, who potentially have as many as 35,000 independent auditory nerve fibers that can convey spectral information, this is a severe limitation.

If CI users are able to utilize FM cues similar to normal-hearing listeners ***in addition*** to however many spectral channels they currently receive, then their ability to perceive speech and music should improve. Overall, these studies and others (Litvak, Delgutte et al. 2003; Litvak, Delgutte et al. 2003) suggest that the introduction of phase information, if delivered in ways that were usable, could provide substantial benefit for CI users.

Consequently, at least two published strategies have attempted to incorporate this additional information into new methods of neural stimulation for cochlear implants, by stimulating at the peaks (Vandali, Sucher et al. 2005) or zero-crossings (Grayden, Burkitt et al. 2004) of CI filter-band outputs. Unfortunately, these strategies have yet to show significant improvement in hearing performance for CI users.

2.4. How can phase information be delivered?

While there are many reasons to believe that phase information is important for higher fidelity listening, it is not at all clear how phase adjustments or frequency modulation

should be conveyed to a CI user. The variation of $\phi(t)$ is usually very wide and rapid, which means it cannot be delivered directly to neurons, which have an absolute refractory period and hence a bandwidth usually no larger than 1 kHz. With regard to CI users in particular, perceptual data show that they cannot detect FM rate changes at base rates of more than a few hundred Hz (Zeng 2002; Chen and Zeng 2004). Hence any scheme proposing to convey phase information to CI users must practically impose both a rate limitation and a bandwidth limitation on the delivery of $\phi(t)$ (Nie, Stickney et al. 2005).

One method is to utilize a zero-crossing detector: we know $\theta(t)$ is exactly 0 or π at the zero-crossings, and so delivering this information (e.g. by stimulating at the zero-crossing times) should provide the brain with fairly rich knowledge of what the phase should be, even with the bandlimiting requirement. However, to stimulate at zero-crossing times, we need some quick and intelligent way of choosing a zero-crossing in one channel over another, because zero-crossings may arrive simultaneously on different channels, and also very rapidly in the high-frequency channels.

Even if zero-crossing stimulation is not used, channel selection is still required to enforce interleaved stimulation, to avoid simultaneous interactions between electrodes. The channel selection process must also prevent one or two strong (high input-intensity) channels from dominating the firing and not giving other channels a chance to fire.

However, zero-crossing stimulation may not be the ideal method of delivering phase information. In the presence of noise, the reliability of zero-crossing times as an indicator of $\phi(t)$ rapidly deteriorates. Furthermore, stimulation at a deterministically precise phase in the signal is very different from what actually happens in biology, as nerves tend to fire at times which are smoothly distributed over all phases, but concentrated near the peak in signal energy (Litvak, Delgutte et al. 2003).

To introduce FM information into the stimulation paradigm, Nie et al. have proposed that either a carrier be frequency modulated with $\phi(t)$, or that the carrier itself be replaced by the band-limited $\phi(t)$ signal (Nie, Stickney et al. 2005). In this thesis, and in the spirit of Nie et al.'s latter idea, our objective is to convey $\phi(t)$ but depart entirely from the concept of either an AM or FM carrier.

Finally, in the search to uncover the neural coding of pitch perception, a study on tones transposed by high frequency modulation (to activate more basal regions of the

cochlea) showed that fine time structure should be mapped tonotopically, or to the correct place along the cochlea, if the pitch is to be perceived (Oxenham, Bernstein et al. 2004). This must be taken into account when incorporating phase information into the stimulation.

2.5. Increased membrane noise in the relative refractory period

In (Matsuoka, Rubinstein et al. 2001), membrane noise properties during the relative refractory period were characterized by electrically stimulating chemically deafened cats according to the paradigm shown in (Figure 3).

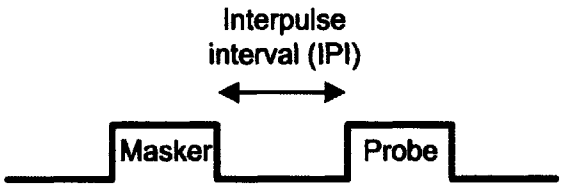


Figure 3. Stimulus paradigm to study growth of the evoked action potential (EAP) during the relative refractory period. A masker pulse was applied, followed by a probe after a fixed interpulse interval (IPI).

Various interpulse intervals were used and the refractory growth function was measured, defined as the amplitude of the EAP resulting from the probe as a function of the stimulus current. The refractory growth function was fit using an integrated Gaussian, and a significant peak in the relative spread (RS, defined as the standard deviation divided by the mean of the Gaussian) of the refractory growth function was measured when the IPI was varied from 0.7 to 4ms. The peak in RS occurred consistently across the 5 implanted cats between an IPI of 0.9ms and 1.1ms, and is re-plot from the paper in (Figure 4).

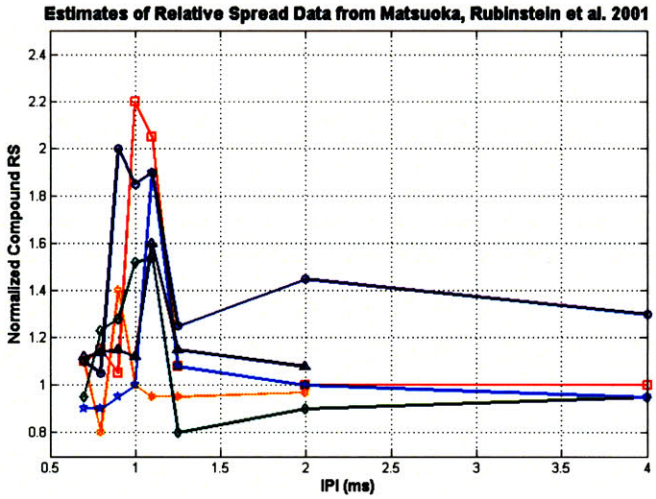


Figure 4. Normalized RS responses for 5 chemically deafened cats. A strong peak shows that the slope in the I/O function is flattest around 1ms. One animal has responses recorded for both cathodal and anodal stimulation.

The increase in RS at the "specific IPI" of ~1ms is due to the smaller number of noisy sodium channels available (due to deactivation) during the relative refractory period. A large number of sodium channels would increase the averaging effect that causes the stochastic response to converge upon a mean, deterministic response. Correspondingly, the gradual reduction in active and noisy sodium channels during the relative refractory period would cause the variance of the response, and hence the RS, to rise.

To summarize, the basic result is that there exists a "specific IPI" which increases membrane noise in a nerve. Thus, if electrical stimulation was configured to prefer this IPI, cross-fiber synchrony may be reduced by activating only a stochastically-varying fraction of the neuron population possessing responsive sodium channels. This potential possibility will have to be verified in a clinical setting.

2.6. Gamma distributions in the Interspike Interval Histogram (ISIH)

(1) It has been claimed that ideally, interspike intervals should be exponentially distributed, as this corresponds to an underlying Poisson process (Rubinstein, Wilson et al. 1999).

(2) More precisely, as we understand neuronal spiking not to be a Poisson process in itself but rather a result of integrating Poisson distributed stimuli up to a fixed firing threshold, the inter-spike interval should have a Gamma

(or Erlang random variable) distribution (Stein 1965). Two sample distributions are shown in (Figure 5).

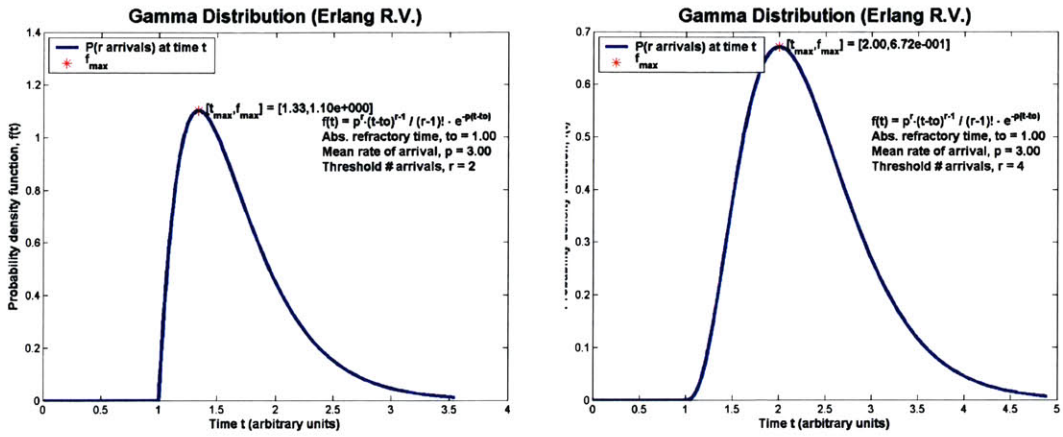


Figure 5. Gamma distribution with parameters to=1.0, p=3.0, r=2 and r=4. Note that the value of r controls the slope of the rising edge, and r=1 reduces the gamma distribution to an exponential distribution with an abrupt (infinite) slope at to.

A bio-realistic ISIH should therefore reflect a more gently rising slope, which peaks before an exponential "decay," where the steepness of the rising slope is determined by the voltage threshold for spiking: a high threshold will have a gentler slope, while a low threshold will have a steeper slope.

2.7. Rate limitation in the auditory response

A final piece of research which impacts our architecture is that of the limitation on rate discrimination, a.k.a. rate limitation in the auditory response. It has been shown that the discrimination of rate changes above 300 pulses per second (pps) is very poor in cochlear implant patients (Shannon 1983). Normal hearing listeners can discriminate changes in rate (i.e. changes in pitch, or F0) of less than 1% for pure tones at low to moderate frequencies, while implant patients have F0 difference limens (F0DLs) on the order of 10% if even measurable, around 300 Hz (Tong and Clark 1985).

This could be a constraint that limits the usefulness of frequency-modulated information (i.e. phase information) in cochlear implants. Furthermore, current CI users may experience a perceptual dissonance due to the lack of consistency of phase cues between neighboring channels that would be present in a normal traveling-wave architecture (Loeb 2005).

Nevertheless, the ability of a CI user to discriminate changes in pulse rate varies between individuals (Townshend, Cotter et al. 1987), and some CI users may yet benefit from rate-coding of pitch and tonal information.

3. Asynchronous Interleaved Stimulation (AIS)

3.1. The "race-to-spike" algorithm

If we do not use zero-crossings, how might we generate good asynchronous firing times? In this section, we propose a bio-inspired way to provide stimulation pulses at asynchronous times which still have a definite correlation with $\phi(t)$. However, if we make the firing times a sole function of $\phi(t)$, this decouples the local decision to fire from global knowledge of activity in other channels, which may be important. In addition, we also do not want to decouple the decision to fire from the envelope strength, as it should be probabilistically more important to provide correct FM information when the AM signal is strong. Therefore, our strategy attempts to incorporate global information across all channels about E(t), and favors the firing of a pulse when E(t) in that channel is larger than others. The resulting times will naturally be asynchronous, and should even appear pseudorandom for natural, arbitrary sounds, given complex interactions between channels during the channel-selection process. Further biological justification for this technique is presented in section 0.

The system is comprised of coupled electronic neurons that incorporate information about E(t) across all channels, and in competition with each other, generate the asynchronous firing times. The strategy is also termed a "race-to-spike" algorithm, as the neurons are set up in a race where the winning channel gets to deliver a stimulation spike on its electrode. The algorithm is described in the following steps:

(1) The system receives as inputs half-wave rectified currents from a bank of bandpass analysis filters, which could be actual currents like those generated by an analog processor (Sarpeshkar, Baker et al. 2005; Sarpeshkar, Salthouse et al. 2005), or a digital version as produced by a DSP.

(2) There is one integrate-and-fire neuron (Dayan and Abbott 2001) associated with each channel, receiving the above-mentioned current input from that

channel, to charge up its neuronal capacitance from the ground state. This begins the "race-to-spike."

(3) The first neuron to reach a fixed voltage threshold "wins" and resets all capacitors back to zero. This ensures that the interleaved stimulation requirement is satisfied, as there can be only one winner.

(4) The winning neuron then fires a current spike (which is an asynchronous timing event) on its electrode that is scaled by the channel envelope's energy.

(5) Once a neuron wins, its input current is inhibited (i.e. weakened) for a period determined by a relaxation time constant, to prevent it from winning repeatedly.

(6) After the winning neuron has fired its spike, we start the neuronal "race-to-spike" (step 2) again.

Natural variations of this algorithm could be implemented, and we will list just a few. The inputs to the integrate-and-fire neurons could be generated by any signal analysis front-end, for example a modulated filter bank or a cochlear cascade of low-pass filters. The voltage threshold could be different for different channels, to create pre-emphasis or to accommodate different sensitivity to stimulation. The level of the electrode current spike could be an arbitrarily complex function of its input, such as an average past envelope stored dynamically on a capacitor in that channel. The level of inhibition following a spike, and also its duration could be arbitrarily complex functions of the past. For the purposes of this thesis, we will implement in simulation only a simple version of this strategy, to be elaborated on below.

We wanted to build on existing work on an ultra-low-power analog bionic ear processor (ABEP) to eventually implement a very low power strategy in hardware (Sarpeshkar, Baker et al. 2005; Sarpeshkar, Salthouse et al. 2005). Thus, to examine the effectiveness of this strategy in conjunction with low-power analog signal-processing components already built, we simulated as our front-end the bandpass filters and envelope detectors exactly as they were implemented in the ABEP: The bandpass filters were conventional 2-stage 4th-order filters with Q=4 and 12 dB/octave highpass and lowpass rolloffs; 16 channels were implemented, with center frequencies scaled logarithmically from 116 Hz to 5024 Hz. The envelope detectors were implemented with

asymmetric attack and release 1st-order low-pass time constants, with τA=1ms and τR=3ms respectively, acting on full-wave rectified filter outputs in order to compute an approximation for E(t). The envelope detector may therefore be thought of as a low-pass filter with cutoff frequencies of 1 kHz and 333 Hz for rising and falling transients respectively. To simulate the finite sampling rate in digitizing E(t), E(t) was passed through a sample and hold running at a rate of 1.8 kHz. Half-wave rectified versions of the bandpass filter outputs were used as inputs to the integrate-and-fire neurons.

Neuronal capacitances are simply modeled as capacitive state variables, and their voltage threshold was set to be the same for all channels, where a low value of around 35mV (in simulation units) turned out to give good results. Using this threshold and input speech tokens at conversational sound pressure level, all channels reach their threshold and fire within ~400 μs of receiving an input pulse, with a time-to-spike of 50-200 μs being most common. A histogram of typical times-to-spike is shown in (Figure 6).

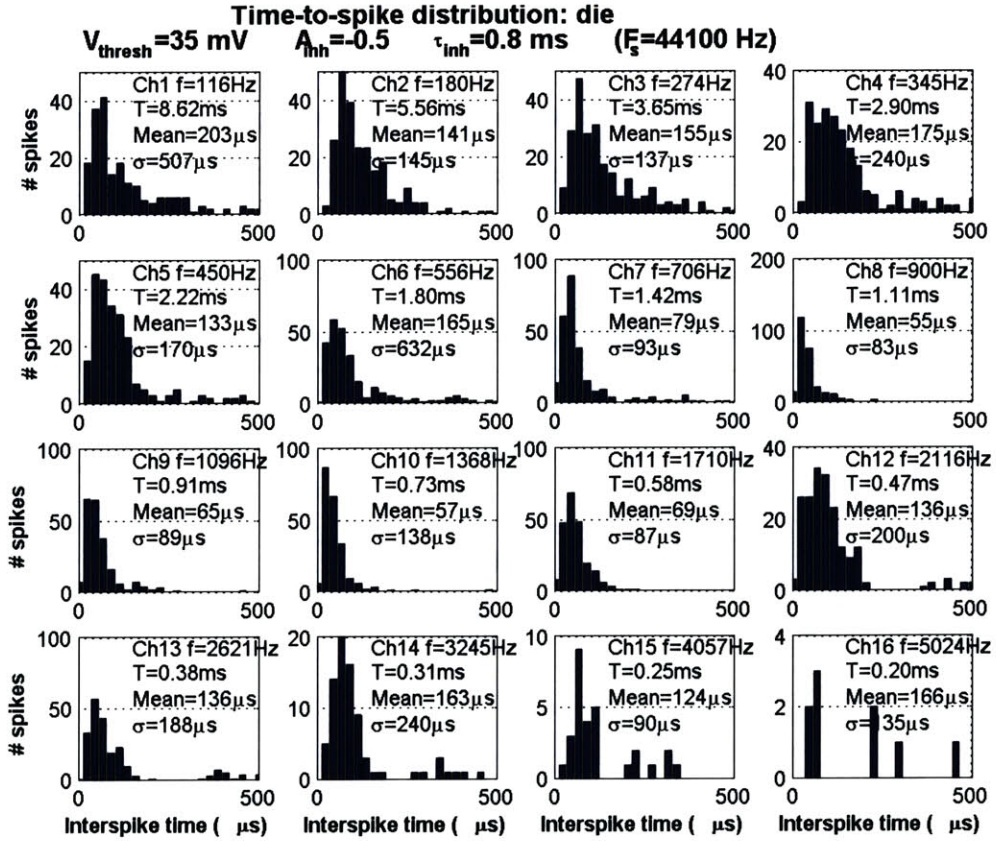


Figure 6. Histogram of times-to-spike, with Vthresh=35mV (in simulation units) and conversational SPL for the speech token "die."

The size of the spike fired was simply set to the sampled-and-held value of E(t), which could be easily implemented in hardware by D/A converters.

The last important detail is the time course of the inhibition described in step (5) above. We wanted to ensure that firing was absolutely prohibited for a minimum amount of time that is determined by the absolute refractory period of biological neurons. This prohibition avoids wasting stimulation power when biological neurons are unable to respond. However, after the absolute refractory period, we would like to softly turn off the inhibition current, thus enabling a very strong input to overcome the imposed inhibition.

To accomplish these objectives, we designed the time course of the inhibition current to be modeled by a Fermi-Dirac exponential roll-off, given by the equation $h_{inh}(t) = \frac{1}{1+e^{k(t-\tau_{inh})}}$ where τ_{inh} sets the time where the inhibition falls to half its maximum value, fixed at 0.8 ms in our simulations and k sets the steepness of the rolloff, fixed at 10^4. The value of τinh was chosen to be 0.8 ms to enforce a minimum inter-spike interval near the refractory period of auditory neurons (Miller, Abbas et al. 2001). The shape of this time course is shown in (Figure 7).

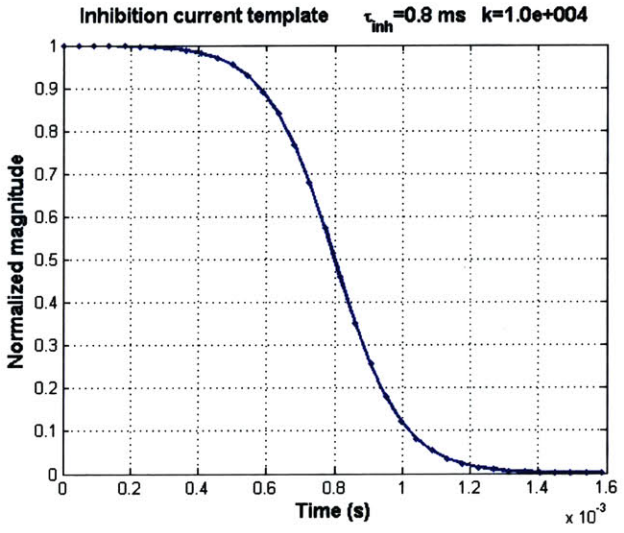


Figure 7. Time profile of inhibition current, with exponential rolloff modeled by the Fermi-Dirac equation. Note the current falls to half its value when $t = \tau_{inh}$.

It also happens to match closely with the decrease in current output of a subthreshold current source if the gate voltage on a pass transistor is linearly decreased, which can be easily implemented in electronics. The inhibition current $I_{inh}(t)$ is then defined as $A_{inh} \cdot h_{inh}(t)$, where A_{inh} sets the maximum inhibition current, fixed at -0.5 in our simulation units.

The algorithm is described in the following pseudocode:

Initialize: capacitor voltages $V_{cap}(t) = 0$ $\forall$ channels

Initialize: time-of-last-spike $t_{lastspike} = -\infty$ $\forall$ channels

Initialize: spiking output $spike(t) = 0$ $\forall$ channels

At each timestep t, $\forall$ channels,

Compute: $I_{inh}(t) = A_{inh} \cdot h_{inh}(t - t_{lastspike})$ ... (step 5)

Increment: $V_{cap}(t)$ += max$[HWR(t) - I_{inh}(t), 0]$ ... (step 1,2)

where $HWR(t)$ is the half-wave-rectified bandpass filter output for that channel

If $\max_{\forall channels} [V_{cap}(t)] > (V_{thresh} = 35\text{mV})$

Find: max_ch = channel with $\max_{\forall channels} [V_{cap}(t)]$

Set: $V_{cap}(t) = 0$ $\forall$ channels ... (step 3)

Set: $spike(t)$ = E(t) in channel max_ch ... (step 4)

Set: $t_{lastspike} = t$ in channel max_ch ... (step 5)

3.2. Results from MATLAB simulations

Various sound files in the .wav format were loaded into MATLAB and input to the simulation. We tested speech utterances of the words "die" and "buy," a Handel chorus containing a word vocalization of "hallelujah," and a piece of music (jazz) that does not contain words.

Being a digital simulation, we had to choose some level of timing resolution with which to discretize the time steps, thus a "sampling rate" of Fs=44100 Hz was used, indicating a timing resolution of about 23 μs. Preliminary results (not shown) suggest that a timing resolution much worse than that degrades the accuracy of our experiments. We present results from "die," which is one of the two speech utterances tested, to illustrate the performance of the system.

3.2.1. Capacitor Voltage and Inhibition Current waveforms

(Figure 8) is a zoomed-in figure of typical capacitor voltage and inhibition current waveforms. We note that spikes are fired when the capacitor voltage reaches a threshold, turning on a negative inhibition current that has the profile as described in (Figure 7). Immediately after a spike, all capacitors are then reset to zero as given by the algorithm. As the level of inhibition current A_{inh} at -0.5 was set to be higher than the highest input level, no channel fires again until its inhibition current has almost returned to baseline.

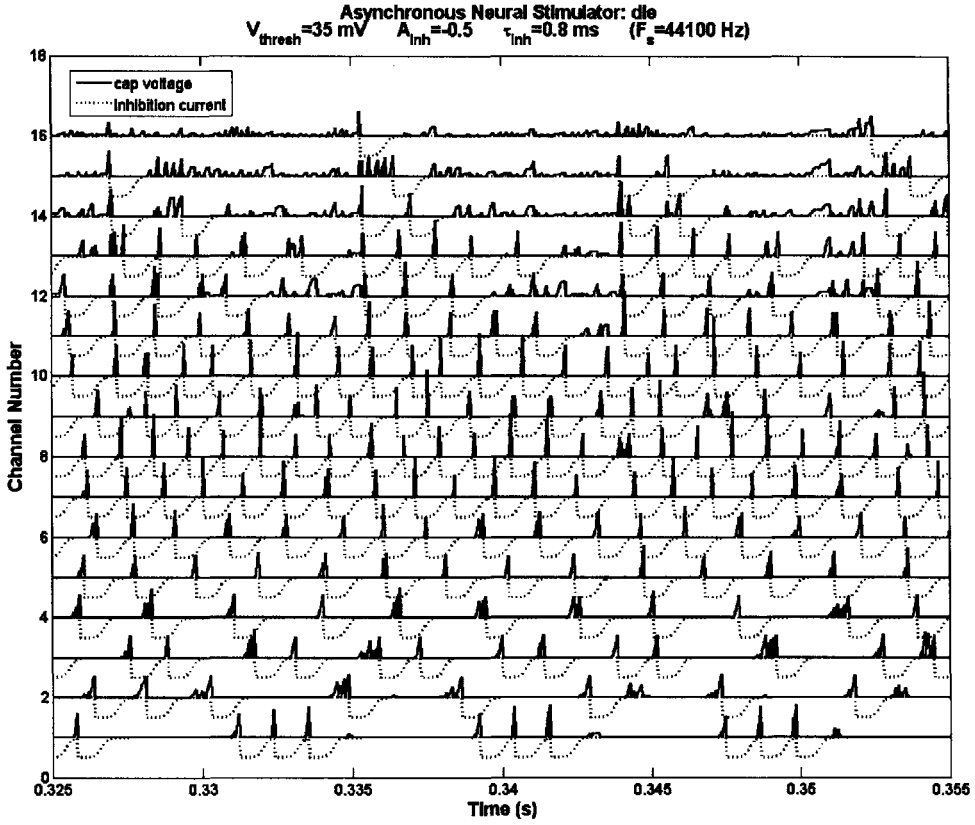


Figure 8. Capacitor voltage (solid lines) and inhibition current (dotted lines) waveforms from race-to-spike simulation of the speech utterance "die."

3.2.2. Half-wave rectified inputs against spiking outputs

(Figure 9) shows the half-wave rectified outputs of the bandpass filters, used as inputs to the asynchronous race-to-spike system, and plotted against its spiking outputs for the same time window as in (Figure 8). We should note that spikes tend to fire near the beginning of each positive excursion in the filter output, but the phase at which a spike is

fired is not deterministic. Instead, the time of firing exhibits some pseudo-stochastic variation due to the competition between channels. We should point out that our strategy does not explicitly model stochastic firing that arises normally in the auditory system, but may, however, introduce stochastic responses that are similar to those encountered normally. By firing only when enough charge has accumulated within a fraction of a half-wave rectified cycle, and when the intensity of a channel is high enough to be the first to spike, a channel will generate spike times that are correlated with but not precisely determined by important features in the phase of the signal $\phi(t)$. The resulting spike trains, as we shall show later, are sufficient for high-fidelity signal reconstruction, and encode phase information with a high degree of correlation.

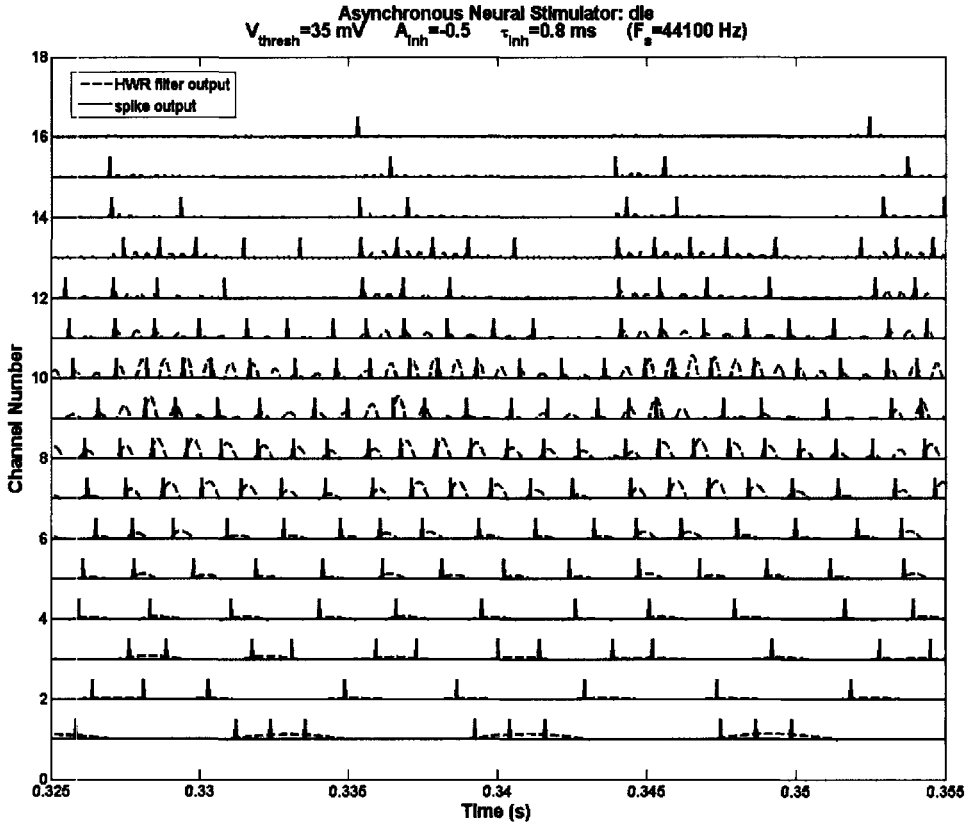


Figure 9. Half-wave-rectified bandpass filter outputs (dashed lines) used as inputs to the race-to-spike simulation, plot against the spiking outputs (solid lines). The spikes in this figure are shown before scaling by E(t), for clarity.

Of note are also the low-frequency channels, where we note multiple spikes may be

fired over a single pulse of energy. This allows pulses of long duration to be well represented, which would not occur in zero-crossing based stimulation.

3.2.3. Interspike-Interval Histograms

The interspike-interval histogram of each channel is presented in (Figure 10). An "absolute refractory period" of about 1 ms is shown to be enforced, keeping the instantaneous firing rate in each channel below a maximum of 1 kHz. The distributions also look fairly natural, where a perfectly natural distribution due to spontaneous firing would be a smooth gamma distribution indicating Poisson arrival of inputs to a neuron. The histograms do not show exact gamma distributions, but the distribution is nevertheless more natural than any distribution produced by synchronous firing.

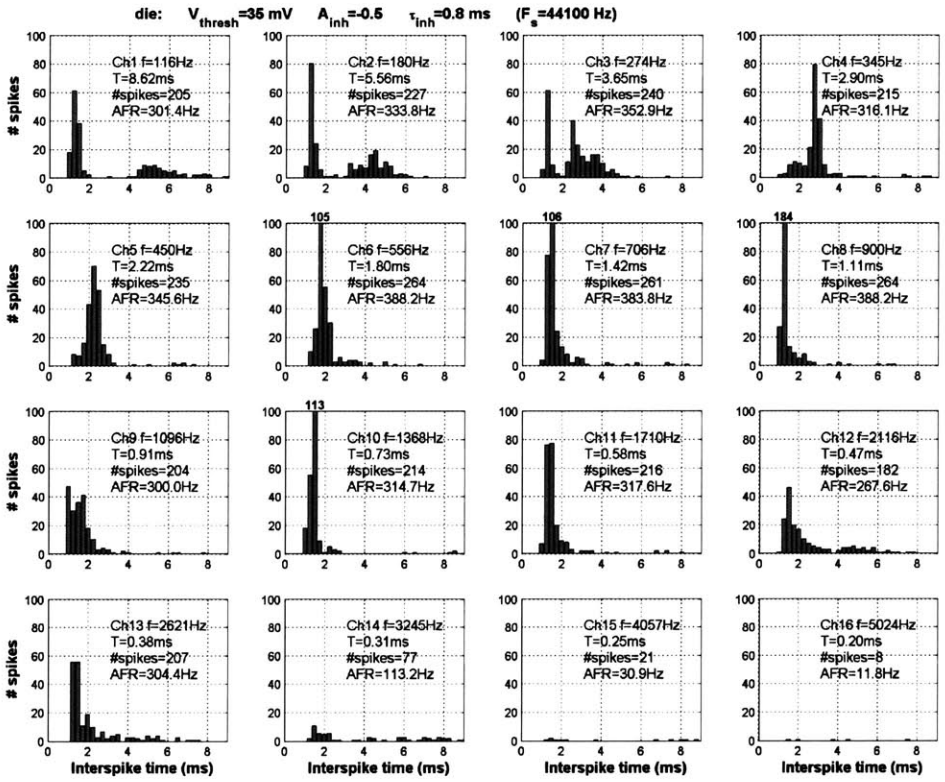


Figure 10. Interspike-interval histograms for the 16 stimulation channels. Also shown for each channel are f, the center frequency of each channel, T = 1/f, the number of spikes in that channel, and the AFR. Spike counts > 100 are clipped (but reported numerically above the figure), and spike intervals > 9ms are not shown in the histogram.

3.3. *Power Savings*

A major advantage of asynchronous stimulation is that the average firing rate (AFR) is able to fall significantly below the worst case maximum firing rate. In contrast, synchronous paradigms need a clock that is fast enough to handle the fastest event rate. Asynchronous stimulation allows the input firing rate to be averaged across time and across channels such that no power is spent during quiet periods or on quiet channels to continuously sample the signal. Sensory codes that adapt to signal statistics are efficient (Smith and Lewicki 2006) and asynchronous stimulation allows us to adapt our sampling rate in time and spectral space to the needs of the signal.

We should note in our simulation above that the AFR in each channel is about 300-388 Hz for most channels, and when averaged over all 16 channels comes down to only 279 Hz per channel. This AFR is lower than the firing rate in conventional synchronous stimulation where firing rates are not adapted to the input stimuli. (Table 1) shows the power savings possible with our technique by stating the worst-case channel's AFR and also contrasting it with the mean AFR of all the channels. We see that averaging in time reduces the worst-case AFR below 1 kHz, thus saving power, and that averaging across channels reduces the mean AFR, also saving power.

Table 1. Maximum and mean (across channels) of each channel's average firing rate (AFR)

Sound file	Maximum AFR (Hz)	Channel with max AFR	Mean AFR (Hz)
"die"	388.2	6	279.4
"buy"	466.5	7	288.9
Handel chorus	772	9	451.9
Jazz music	556.9	7	275.3

3.4. *AIS spike reconstruction and comparisons*

In this section, we present a method of comparing the AIS strategy against other acoustic simulations of cochlear implants. Many acoustic simulations reconstruct the sound input to a CI from its output of channel envelopes by using white-noise or tonal carriers for each channel (Shannon, Zeng et al. 1995; Dorman, Loizou et al. 1997), and are known as noise/tone vocoding reconstructions. To similarly reconstruct a sound from

the train of spikes generated by AIS, we use a spike-based reconstruction technique that has its foundation in prior neurophysiology work, showing that analog waveforms can be accurately reconstructed from spiking waveforms. For example, optimal low-pass filters can be designed, that when applied to the recorded spike trains from the *Eigenmannia* electric fish, produce as their output a very well-correlated reconstruction of the input voltage variations in the fish's sensed aquatic environment (Wessel, Koch et al. 1996). Such experiments show that frequency modulations in neuronal spike trains can encode an analog input so well that only conventional low-pass filters are needed for stimulus reconstruction.

Recent work has also shown that a spike-based auditory code can very efficiently encode speech, outperforming gamma-tone, wavelet and Fourier decompositions, achieving the highest SNR in representing speech at coding rates <60 kbits/s (Smith and Lewicki 2006). In this research, sound signals are also reconstructed from spikes using tuned kernel filters that have impulse responses closely resembling those of high-order resonant low-pass filters. Interestingly, these impulse responses independently matched physiological data; the dictionary of 32 kernel filters that were adapted to give good reconstructions turned out to match the tonotopic population of auditory neuron impulse responses in cats, as derived from reverse correlation analysis (deBoer and deJongh 1978). These results lend further support to the validity and usefulness of impulse response reconstructions for neuronal spiking codes.

3.4.1. Channel-by-channel spike reconstruction

Motivated by prior work on spike-based reconstruction techniques like that in the electric fish, we applied 2-stage 4th-order low-pass resonant filters (with the same center frequency and Q of 4 as the bandpass filters in each channel and with 24 dB/octave low-pass rolloffs) to the spiking outputs from each channel. The spikes effectively behave like envelope-scaled impulses and the filters sum the impulse responses from various spikes to recreate the analog information in each channel:

$$reconstruction_{ch}(t) = \frac{\int_0^\infty spike_{ch}(\tau) \cdot h_{ch}(t-\tau)\, d\tau}{\max_{0 \le t < \infty} h_{ch}(t)} \quad (4)$$

where $spike_{ch}(t)$ contains the $E(t)$-scaled spikes from channel *ch*,
$h_{ch}(t)$ is the low-pass filter impulse response for channel *ch*,
and $reconstruction_{ch}(t)$ is the reconstructed signal on channel *ch*

As the peak in the impulse response $h_{ch}(t)$ for each channel increases linearly with the

center frequency, we needed to normalize the reconstructions across channels by dividing by the peak value of $h_{ch}(t)$. The reconstructed waveforms on a channel-by-channel basis are shown in (Figure 11):

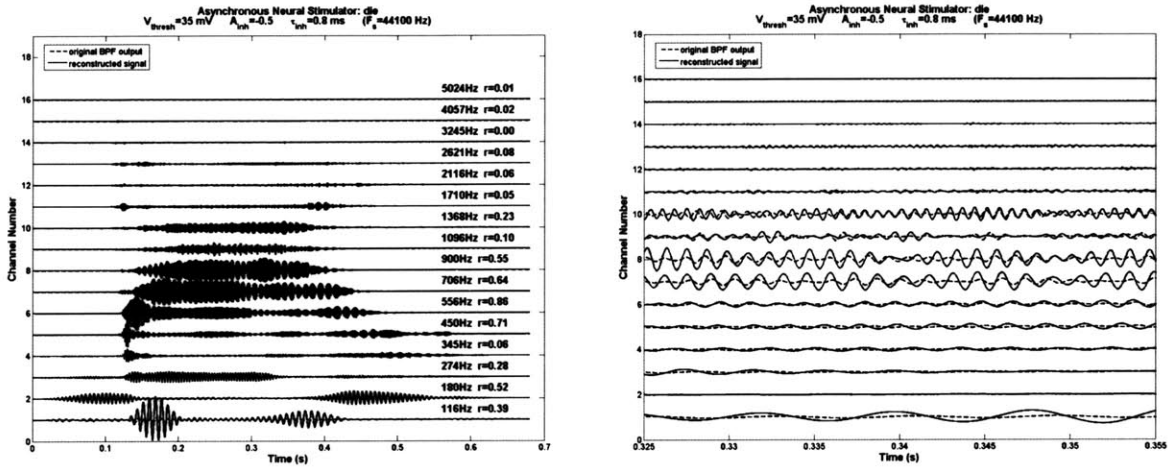


Figure 11. Sixteen-channel spiking reconstruction (solid lines) of the bandpass filter outputs (dashed lines). A magnified version of the same time window as in earlier figures is shown on the right.

Correlation coefficients (r) for each channel are computed from the following equation:

$$r = \frac{\int_0^\infty BPFoutput_{ch}(t) \cdot reconstruction_{ch}(t)\, dt}{\sqrt{\int_0^\infty BPFoutput_{ch}^{\,2}(t)\, dt \cdot \int_0^\infty reconstruction_{ch}^{\,2}(t)\, dt}} \quad (5)$$

where $BPFoutput_{ch}(t)$ is the bandpass output for channel *ch*, and both $BPFoutput_{ch}(t)$ and $reconstruction_{ch}(t)$ have had their means removed, i.e. are zero-mean signals

The correlation coefficient is on a scale from 0 to 1, with 1 indicating a perfect correlation between the bandpass filtered output and the spike reconstruction. The correlations between each channel are fairly good, with a low-frequency channel having a correlation coefficient as high as 0.86. In performing these correlations, it was also important to account for group delay introduced by the bandpass and low-pass filters, which causes the composite *reconstruction*(*t*) (as defined in the next section) to lag the original signal. Thus, to compensate for the group delay (which does not affect the sound fidelity), the cross-correlation between *reconstruction*(*t*) and the original signal was performed, and the lag corresponding to the peak in the cross-correlation was then used to time-shift *reconstruction*(*t*) and align it with the original signal. The lags for the 4 sounds ranged from 4.8 – 9.5 ms.

3.4.2. Composite signal spike reconstruction

If we sum all the reconstructed channels together, we can generate a composite reconstruction of the original signal just as in CI acoustic simulations using vocoding reconstruction, defined as follows:

$$reconstruction(t) = \sum_{ch=1}^{16} reconstruction_{ch}(t) \tag{6}$$

The summated signal is shown in (Figure 12), with a zoomed-in version again on the right. The phase relationships are clearly preserved, and the envelope is also fairly well tracked.

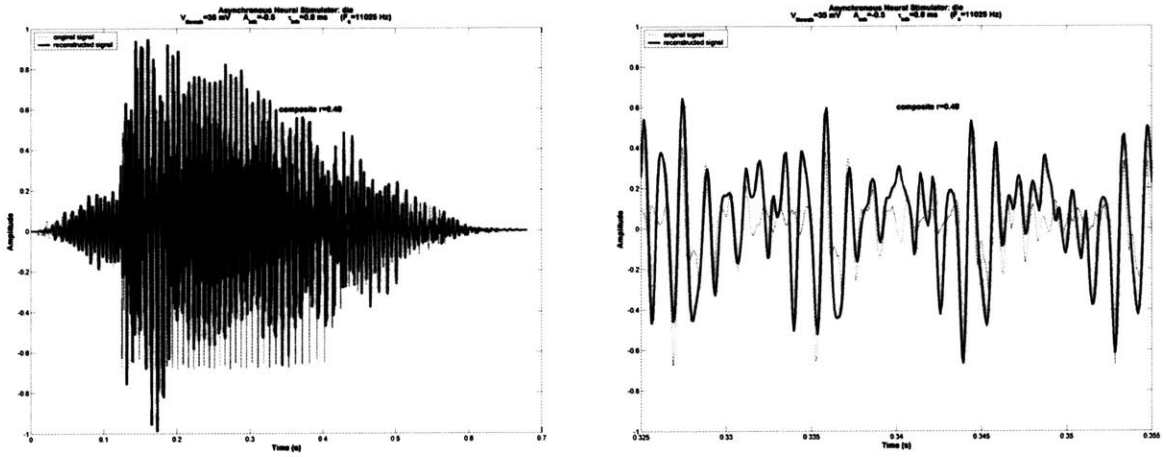


Figure 12. Composite spiking AIS reconstruction (solid line) of the waveform "die," by summing the channel reconstructions together. Note the reconstruction is downsampled to match the sampling rate of the original sound signal (dotted line). The composite correlation coefficient is 0.48.

The composite correlation coefficient is 0.48 in this example, the word *die*, and is calculated as follows:

$$r = \frac{\int_0^\infty original(t) \cdot reconstruction(t)\, dt}{\sqrt{\int_0^\infty original^2(t)\, dt \cdot \int_0^\infty reconstruction^2(t)\, dt}} \tag{7}$$

where $original(t)$ is the original sound signal, and both $original(t)$ and $reconstruction(t)$ have had their means removed, i.e. are zero-mean signals

Note that the sampling rate was reduced to match the same sampling rate of the original signal, to make a fair comparison. A flowchart of our entire reconstruction technique is shown in (Figure 13).

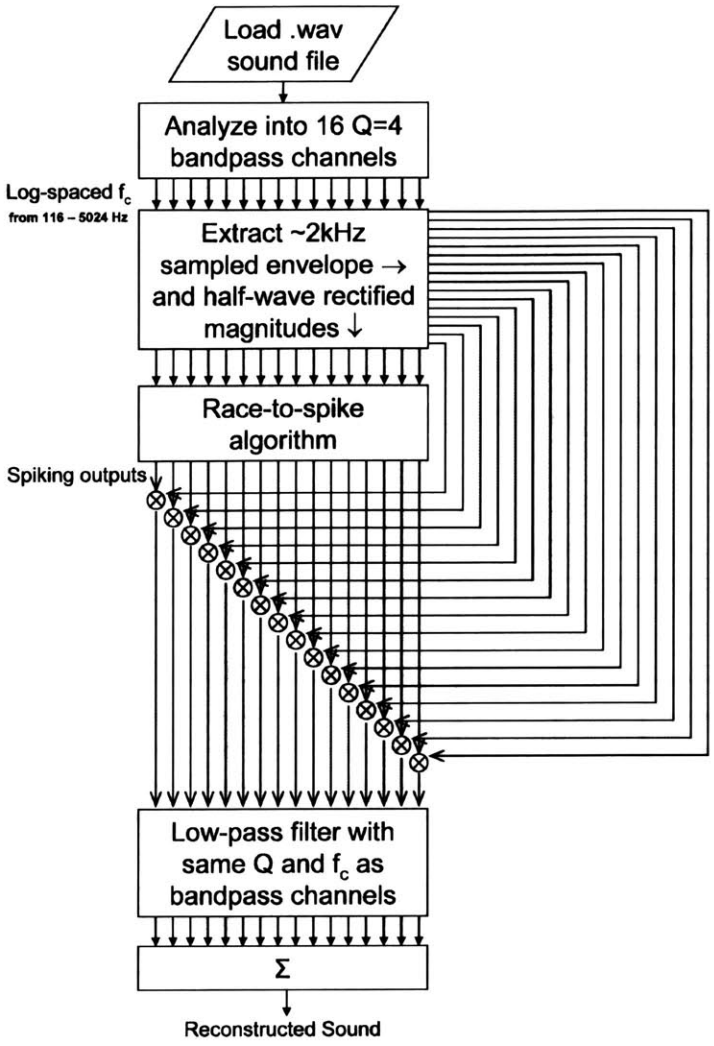


Figure 13. Flowchart of the entire AIS reconstruction process.

3.4.3. Hilbert decomposition and correlation

It is not immediately obvious how much of the phase information $\phi(t)$ is retained in the spike output from the AIS strategy. In order to better quantify the transmission of $\phi(t)$, we performed a Hilbert decomposition of the reconstructed signal into $E(t)$ and $\theta(t)$ components as described in Section 2.2. These components were then correlated separately with the original signal's envelope and phase. In order to see whether the correlations for $E(t)$ and $\theta(t)$ were significant for AIS, we then compared them against CIS noise and tone vocoding reconstructions, and also a CIS spike-based reconstruction which employs the same reconstruction filters as described in Eq. (4), except that the non-overlapping spike input is now sequentially and synchronously firing at a fixed rate

of 1.4 kHz. Other firing rates were tested as well but found not to make a significant difference to the results. The sound samples used were the words die and buy, to be representative of speech, and snippets from Handel's Hallelujah chorus and a jazz piece, Highway Blues, to be representative of music. As the results for CIS noise vocoding could vary significantly between trials due to randomness in the noise input, we conducted 100 trials and present the mean and standard deviation of those trials. Envelope correlations are shown in (Table 2), phase correlations are shown in (Table 3) and composite correlations are shown in (Table 4).

Table 2. Correlation in E(t) envelope component for different processing methods

Sound file	AIS spike-based reconstruction	CIS noise-based vocoding (100 trials)	CIS tone-based vocoding	CIS spike-based reconstruction (1.4 kHz firing rate)
"die"	0.6	0.50 ± 0.04	0.57	0.48
"buy"	0.6	0.46 ± 0.08	0.46	0.37
Handel chorus	0.54	0.27 ± 0.06	0.33	0.27
Jazz music	0.49	0.18 ± 0.02	0.29	0.52

Table 3. Correlation in θ(t) phase component for different processing methods

Sound file	AIS spike-based reconstruction	CIS noise-based vocoding (100 trials)	CIS tone-based vocoding	CIS spike-based reconstruction (1.4 kHz firing rate)
"die"	0.19	0.03 ± 0.02	0.07	0.16
"buy"	0.13	0.03 ± 0.02	0.01	0.07
Handel chorus	0.42	0.03 ± 0.01	0.02	0.01
Jazz music	0.4	0.01 ± 0.01	0.01	0.01

Table 4. Correlation in composite reconstruction for different processing methods

Sound file	AIS spike-based reconstruction	CIS noise-based vocoding (100 trials)	CIS tone-based vocoding	CIS spike-based reconstruction (1.4 kHz firing rate)
"die"	0.48	0.10 ± 0.02	0.09	0.07
"buy"	0.43	0.10 ± 0.02	0.08	0.04
Handel chorus	0.67	0.08 ± 0.02	0.04	0.01
Jazz music	0.58	0.04 ± 0.01	0.02	0.01

Correlation coefficients for θ(t) are in general lower than that for E(t) because θ(t) by nature varies much more rapidly than E(t). However, θ(t) correlation coefficients are

clearly higher for AIS than other reconstruction techniques, especially for the music pieces. As the E(t) correlation coefficients are not that much different between reconstruction strategies, the improved match in the composite reconstruction for AIS may likely be due to the improved transmission of $\phi(t)$ matching up with E(t) in the spike output. A tone vocoder reconstruction analogous to the AIS reconstruction in (Figure 12) is shown in (Figure 14), and the effect of poor phase correlation can be clearly seen.

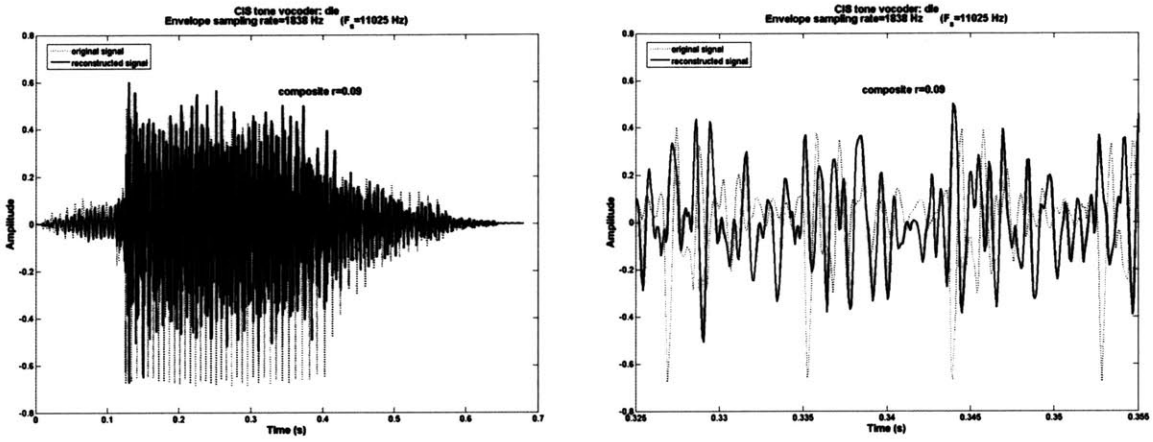


Figure 14. Composite tone vocoder reconstruction (solid line) of the original speech waveform "die" (dotted line). The composite correlation coefficient is 0.09.

3.4.4. What do they sound like?

Correlation coefficients only give us a very general indication of the signal fidelity. The added value of performing a spike reconstruction is that the level of information encoded can be assessed by listening. If the sound fidelity is poor, it does not necessarily mean that a high level of information is absent, but if the sound fidelity is good, it demonstrates that that a high level of information must be present. Some sample reconstructions can be listened to at http://avnsl.mit.edu/AIS.

The sound quality in AIS reconstructions is noticeably improved in that they sound more natural, and AIS channels should contain sufficient information for tonal languages to be correctly represented. In the case of music, while other reconstructions retain only the rhythm, a clear melody and even different musical instruments are perceptible in the AIS reconstructions.

3.5. Perceptual tests in noise

In order to determine whether AIS can provide any real advantage in cochlear implants, testing our strategy with CI users is absolutely necessary. However, such testing is costly in time and resources, and hence often prohibitive unless we are convinced of the new strategy's potential merits. One faster and less costly way of evaluating a new CI strategy is to perform perceptual tests on normal-hearing listeners, using acoustic reconstructions of cochlear implant outputs as described above. While previous results from CI simulations have been found to correlate with CI performance on some perceptual measures (Friesen, Shannon et al. 2001), it should be emphasized that they can only gauge the best possible outcomes for CI users, as many differences between acoustic and electric stimulation are not accounted for in these reconstruction techniques, such as the channel interactions and poor spatio-temporal coding in real implants. Furthermore, the tests we perform to evaluate AIS are tasks relying on the perception of fine time structure, and unlike tasks that rely only on the perception of envelope cues, there is only a small body of evidence to suggest that the results will correlate with actual CI performance (Collins, Wakefield et al. 1994). Nevertheless, these perceptual tests should provide an indication of whether a new CI strategy is worth testing on real CI users. In this section, we present two psychoacoustic experiments that were designed to verify whether AIS provides any advantage in coding speech and music, particularly in the presence of noise, as perceived by normal-hearing listeners.

3.5.1. Methods

Eight normal-hearing listeners were recruited from a local on-line bulletin board to participate in this study. Five subjects were female and three were male, ranging in age from 23-48 with a mean age of 29.5 years. Their hearing thresholds were screened before the test to be at 20 HL or better. Signals were presented at 70 dB SPL over Sennheiser 580 headphones in a sound-attenuating booth. Speech-spectrum shaped noise was used in conditions where noise was added to the stimulus. Each experiment began with practice trials, during which feedback was provided.

The target sounds and noise maskers were mixed before processing, using either AIS spike reconstruction as described in the previous section, or envelope vocoding with tonal

carriers (as in (Dorman, Loizou et al. 1997; Poissant, Whitmal et al. 2006)). In both cases, 16 contiguous frequency channels were used, with center frequencies spaced equally on a logarithmic scale, and an overall passband extending from 100 to 5000 Hz. Pilot data from CIS spike-based reconstructions and CIS noise-based vocoding both resulted in poorer performance than CIS tone-based vocoding for speech recognition in noise and were therefore not used, as we wanted to compare the best-performing CIS acoustic simulation against our AIS acoustic simulation.

In the first experiment, subjects were told that they would be listening to distorted speech sounds in a noisy background. They were told that some of the utterances would be very hard to understand, and that they should type all the words that they think they hear. In the practice trials, they were presented with 8 lists of 10 HINT sentences (Nilsson, Soli et al. 1994), counterbalanced across subjects for the 2 processing conditions tested, namely AIS spike reconstruction and CIS tone vocoding. Half of the practice trials were presented in quiet (SNR=∞) and the other half at +6 dB SNR, in alternating sequence. During this stage, subjects were given the chance to hear sentences again after typing their response, and were shown on the screen what they had just heard. Actual trials used 16 different lists of 10 HINT sentences at four different SNR's (6, 3, 0 and -3 dB). Lists were randomly selected, and SNR and processing condition were randomized for each list. This resulted in 2 lists (or 20 sentences) for each condition. Repetition of the stimuli was not allowed and no feedback was provided.

In the second experiment, subjects were presented with 34 common melodies that had all rhythmic information removed, and were synthesized from 16 equal-duration notes using samples from a grand piano. These melodies were also used in previous studies on melody recognition (Lobo, Toledos et al. 2002; Smith, Delgutte et al. 2002). Subjects were then asked to select 10 melodies they were most familiar with, which were then played back in random order for them to identify. All subjects were able to find 10 melodies that they could easily identify correctly. Actual trials presented subjects with their 10 melodies which were processed by AIS spike reconstruction and CIS tone vocoding. Melodies were presented at two SNR's (in quiet, SNR=∞, and in noise, with SNR=0 dB), counterbalanced across subjects for both SNR and processing condition. All melodies were presented twice in random order for each experimental condition. Subjects

were instructed to identify the melody, and were forced to select their response from the closed set of 10 melody names on the screen in front of them. Repetition of the stimulus was not allowed and no feedback was provided.

3.5.2. Results

(Figure 15) shows HINT sentence recognition scores as a function of SNR for the two processing conditions, AIS spike reconstruction and CIS tone vocoding. In general, subjects did no worse with tone vocoding than with AIS for the conditions of 6, 3 and 0 dB SNR. However at -3 dB SNR, subjects performed better with AIS by 17 percentage points. Analysis of variance (ANOVA) on arcsine transformed data (to normalize the compression of variance near 100% and 0%) showed a significant main effect for SNR [$F(2,13) = 145$, $p<0.001$], processing condition [$F(1,7) = 5.80$, $p<0.05$] and an interaction between SNR and processing [$F(2,15) = 7.53$, $p<0.005$]. A post-hoc analysis (including Bonferroni correction) using a paired samples t-test revealed that only the -3 dB SNR condition showed a statistically significant difference between the two processing conditions [$t(7)=5.61$, $p<0.005$]. These results confirm the experiments of Nie et al. (Nie, Stickney et al. 2005), which suggest that additional FM cues improve performance more in noise than in quiet. AIS may therefore improve the hearing of speech in noise, if additional phase information can indeed be delivered to CI users by this strategy.

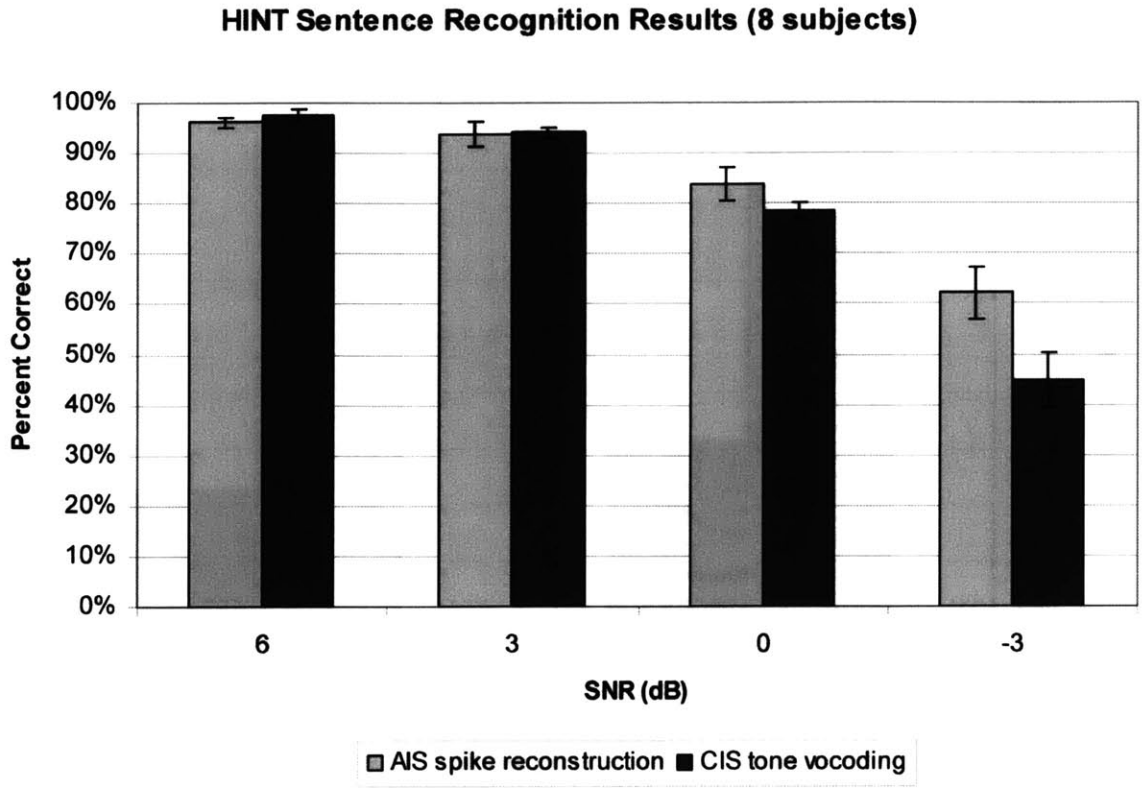


Figure 15. Sentence recognition scores for AIS vs. CIS tone vocoding reconstructions in noise. Error bars show one standard error.

(Figure 16) shows melody recognition scores for the two SNR conditions and two processing conditions tested. Subjects performed better with AIS by 55 percentage points in quiet (∞ SNR), and by 61 percentage points in noise (0 dB SNR). A repeated-measures ANOVA on arcsine transformed data revealed a significant main effect for processing condition [$F(1,7) = 107$, $p<0.001$]. In general, subjects were clearly more able to recognize melodies correctly when listening to AIS spike reconstructions. Thus, tonal perception in CI users may also be improved, if additional phase information is in fact transmitted by AIS. There was no significant main effect for SNR or any interaction. Thus, the addition of noise at an SNR of 0 dB had no significant effect on subjects' ability to recognize melodies in either processing scheme.

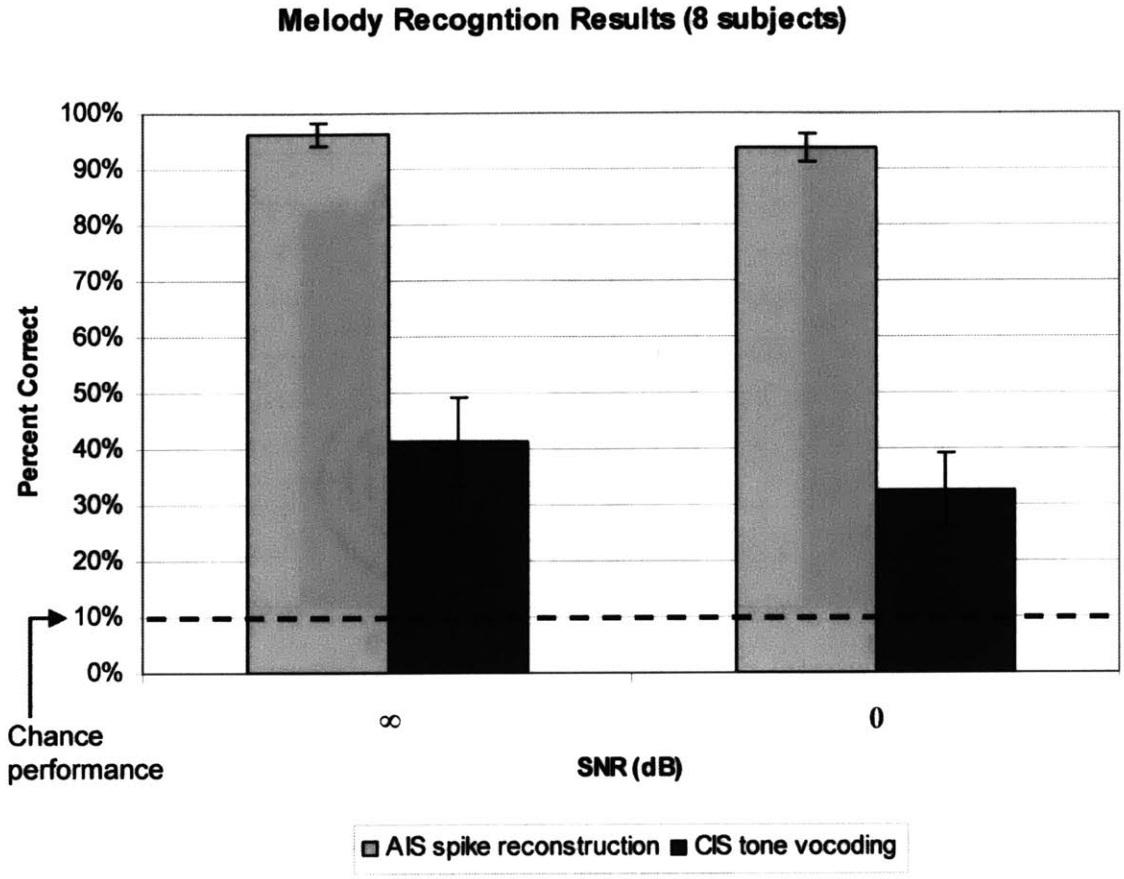


Figure 16. Melody recognition scores for AIS vs. CIS tone vocoding reconstructions in quiet and noise. Error bars show one standard error.

3.6. Discussion

Earlier stimulation strategies have recognized the importance of providing phase information. For example, the Peak Derived Timing (PDT) strategy presented in (van Hoesel and Tyler 2003; Vandali, Sucher et al. 2005) stimulates at times corresponding to positive peaks in the filter-band output, and the Spike-based Temporal Auditory Representation (STAR) strategy (Grayden, Burkitt et al. 2004) generates stimulation pulses at the zero-crossings. Both also encode phase information in the time of firing, but what truly matters is whether CI users are able to utilize the coded information. CI users in (Vandali, Sucher et al. 2005) were found to do no better at a pitch ranking task using PDT than other strategies that did not encode phase information. Many factors are likely to limit phase coding with CI users, such as perceptual dissonance (Loeb 2005) and widespread fiber synchrony which is endemic to electrical stimulation, and for these

reasons AIS may perform no differently from other strategies. However, in contrast with other strategies, AIS firing times are determined when E(t) in a channel is deemed in a uniquely bio-inspired way to be more pertinent than other channels, using a neuronal "integrate-to-fire" competition. Whether this or other details in AIS make any difference, however, remains to be proven in tests with real CI users.

3.7. Summary of contributions

We have demonstrated a simulation of an asynchronous interleaved sampling (AIS) strategy for neural stimulation in cochlear implants, that encodes both phase and envelope information known to be important in perceiving tonal languages and music, and for hearing in noise. Stimulus reconstructions with AIS using simple filtering-and-sum techniques show significantly higher correlation coefficients with the input for both speech and music than other stimulus reconstructions which use only envelope information. Perceptual tests in noise show that the improved correlation is reflected in normal-hearing listeners' ability to recognize both sentences and melodies more easily with AIS reconstructions than with more traditional envelope vocoding techniques. Our results confirm that phase information should make a greater difference for perceiving melodies than for speech in noise.

My particular contributions were in:

1. Creating the MATLAB simulation and investigating the parameters which mattered – for example, a low Vthresh of 35mV in simulation units.
2. Discovering how to do spike reconstructions, and moving from BPF to 1-stage LPF and finally to 2-stage LPF reconstructions to get the best performance.
3. Exploiting the Hilbert transform to separate envelope and phase correlations.
4. Quantifying the power savings from a reduction in AFR.
5. Qualifying the "naturalness" of firing by looking at ISIHs.
6. Designing the perceptual experiments on normal-hearing listeners.

3.7.1. Acknowledgements

The perceptual experiments conducted in section 3.5 were done with the help of Andrea Simonson and Andrew Oxenham from the Auditory Perception and Cognition Lab at MIT. The original idea for AIS was conceived by Mike Faltys at Advanced

Bionics and my advisor, Rahul Sarpeshkar.

4. Analog Circuit Implementation of AIS

4.1. Motivation

In order to deliver the temporal information in $\phi(t)$ within a pulsatile stimulation paradigm, the ability to stimulate at precise times is essential, and therefore high temporal resolution is required to take full advantage of asynchronous stimulation. If conventional discrete-time signal processing is used to generate asynchronous stimulation times, a high-rate sampling c·lock is required to achieve high temporal resolution, which can be costly in power. However, an event-based asynchronous digital or continuous-time analog signal processing scheme can provide high temporal resolution without the need for a fast sampling clock that is constantly running whether there are events or not.

Rather than waiting for custom asynchronous hardware to be developed to test the AIS strategy, we went ahead to develop our own in analog electronics. For the purposes of demonstrating the hardware feasibility and performance of the AIS algorithm, it made most practical sense for me to build upon the existing analog bionic ear processor (ABEP) described in section 3.1, which could readily provide continuous-time analog inputs in the form of half-wave-rectified (HWR) currents from the envelope detectors. The envelope detector circuit in the ABEP, more fully described in (Zhak, Baker et al. 2003), needed just a few modifications to output fast HWR currents and is shown in (Figure 17) below:

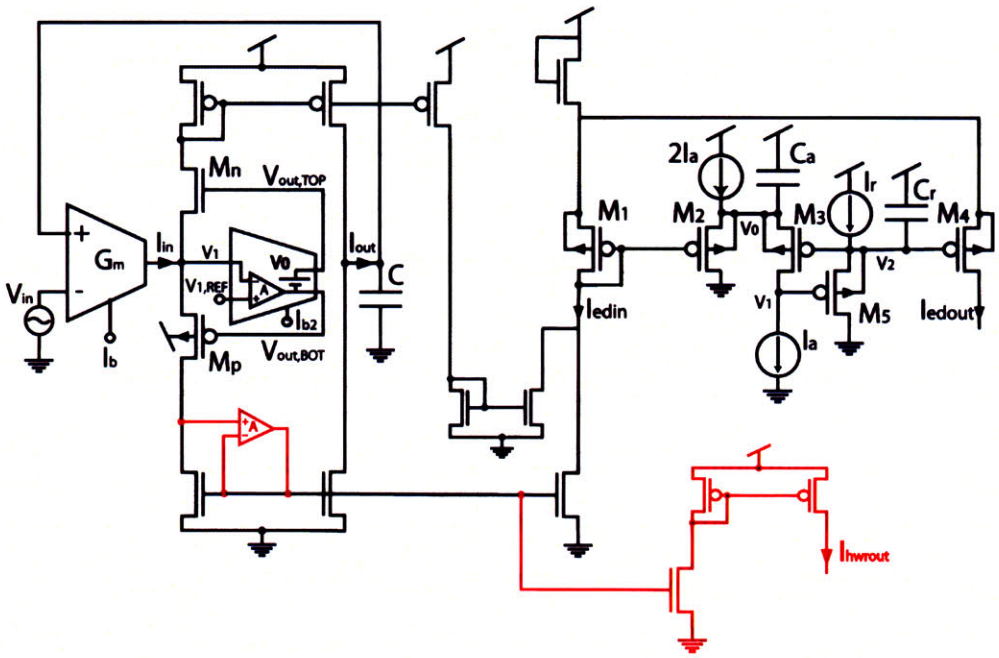


Figure 17. Envelope detector circuit with fast HWR output shown in red

A second compelling reason to use continuous-time analog signal processing is because it would be compatible with a very low-power implementation by using sub-threshold building blocks.

As a simple comparison with an asynchronous digital solution, listed below is the number of operations for each of the 16 channels that would be needed from the algorithm in section 3.1:

(1) 3 additions

(2) 2 multiplications (+1 more to vary channel thresholds independently)

(3) 17 comparisons

These operations need to be performed on state variables with at least 16 bits of precision to capture the full 60dB dynamic range in the current input before log compression. The maximum rate at which these operations need to be completed would be around the inverse of a single pulsewidth (~50-100μs), in the limiting case where a series of pulses might fire back to back. The bandwidth of these asynchronous blocks must therefore be greater than 10-20 kHz. The instantaneous power consumption might therefore be quite high, in the case where many bits needed to flip at this rate. Energy-wise, however, since these operations can be averaged out over time to the average firing rate of a few hundred Hz, even without using an aggressively scaled technology, the

average power consumption should be comparable to an analog processor.

A more unique advantage of using continuous-time analog processing (as opposed to asynchronous digital processing) is found in the low complexity of analog primitives that can be used to implement basic multiply-add and comparison-selection operations. These analog primitives will be described in the following sections.

4.2. Circuit Specifications

Before getting into a description of the circuit details, a summary of the specifications that are needed for both practicality and performance are listed below.

(1) **Individual tunability**: we should allow Vthresh to be set independently in each channel, to allow channel equalization, pre-emphasis, and also to accommodate variability between patients and neural populations within each patient. Provision to prefer a specific interpulse-interval (IPI) should also be made, to take potential advantage of the increased membrane noise during the relative refractory period. Finally the pulsewidth of each spike should be variable to allow for channels that need to maximize the amount of charge transfer by increasing time, rather than current, to stay within the voltage compliance of the supply rails.

(2) **Channel separation**: pulses in different channels should not overlap in time, to minimize channel interaction and current spreading effects. In other words, the interleaved stimulation constraint of having only 1 electrode active at any one time should still be enforced.

(3) **Temporal non-interference**: additional attention must be paid to the temporal implementation of the spike generating circuits. As suggested by the research described in Section 2.7, rate-discrimination limitations in cochlear implant users may be due (at least in part) to temporal processing issues like auditory filter ringing. If auditory filters can be modeled by constant-Q filters like what we implemented, the number of cycles spent ringing varies only with Q and thus remains roughly constant, so low frequency channels may ring for a longer time than high frequency channels. Thus it may be desirable to have constraints in the separation of stimulation pulses in time (for

example, to guarantee a minimum delay between stimulation of adjacent channels) in order to minimize temporal interference.

(4) **Low power**: it is crucial that the final system be a low power design, to be practical for implantation on a tight power budget.

A full 16-channel analog system was built to meet these specifications in an AMI 1.5 μm process, and will be called the AIS ABEP in the remainder of this thesis.

4.3. The "race-to-spike" circuit

Shown below is a schematic of a circuit which implements the race-to-spike algorithm, built from a voltage winner-take-all topology. Sixteen channels have been implemented, but only two channels are shown for illustrative purposes in (Figure 18).

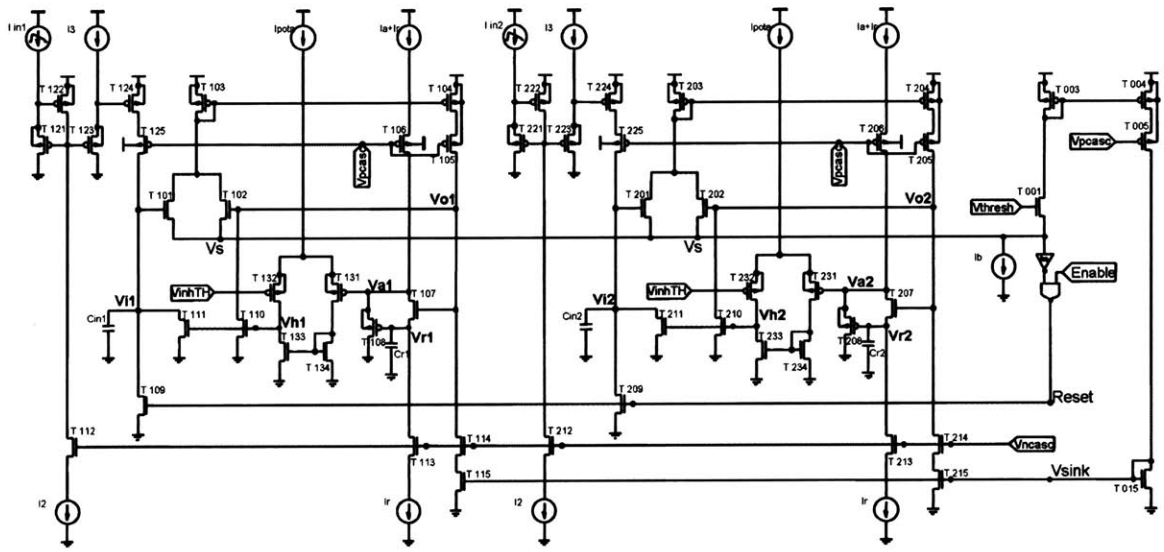


Figure 18. A two-channel race-to-spike circuit, built from a voltage winner-take-all topology.

Channel 1 transistors are labeled T1xx, channel 2 transistors are labeled T2xx and there is a "Channel 0" with transistors labeled T0xx which is required to set the nominal voltage threshold, Vthresh.

The circuit can be understood in 3 separate sub-sections:

(1) A translinear current gain input stage

(2) A voltage winner-take-all topology with positive feedback

(3) A super-buffer to set attack (Ta) and release (Tr) time constants

4.3.1. A translinear current gain input stage

The translinear current gain input stage is shown in (Figure 19) below.

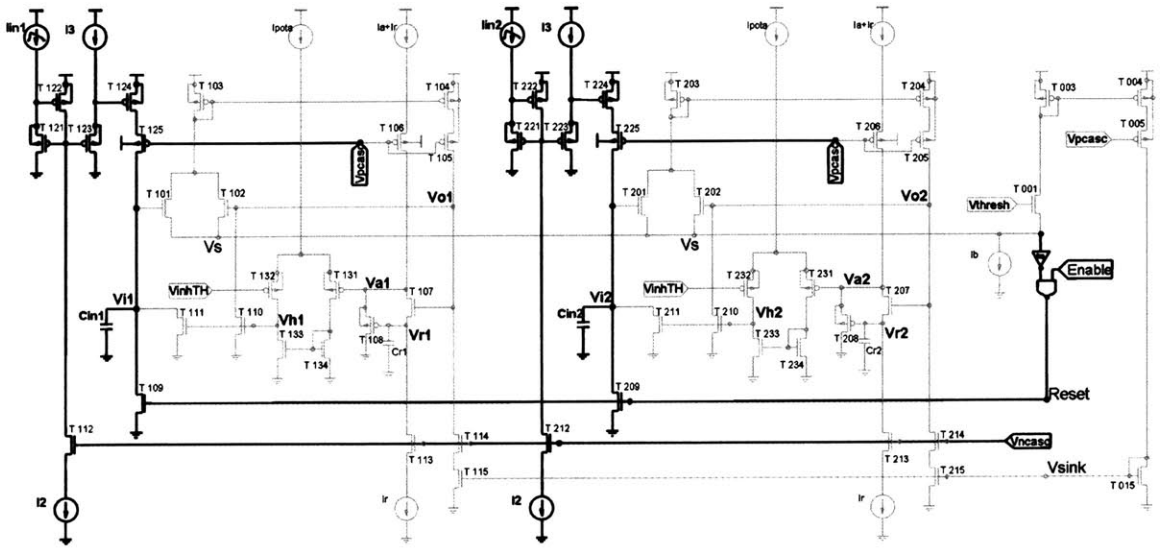


Figure 19. The 2-channel race-to-spike circuit with translinear current gain input stage in bold.

A programmable gain is applied to the input current of each channel by a translinear current mirror, shown in bold as Tn21, Tn22, Tn23, Tn24 and Tn25 (where n=channel number). Since the PMOS gate-to-source voltages are logarithmic functions of their input current when biased in the sub-threshold region, the output current is given by Iin·I2/I3.

The input current was chosen to be injected at Iin rather than I2 even though the DC transfer function is equivalent, because from an AC analysis, the input impedance seen at Iin is $1/g_m$ divided by a gain of $g_m r_o$ while the input impedance at I2 is just $1/g_m$. This is important for the input stage to have high input bandwidth, since there might be substantial input capacitance that would otherwise degrade the sharpness of the phase information in the HWR pulses. A formal analysis is proven in section 5.4.3.

The gained input current then charges the 'neuronal' input capacitance, Cin. Once any one of the channels 'wins' and causes Vs to go high, Reset transistors T109 and T209 then turn on and short all the input voltages (Vi) to ground.

The current gain is important for several reasons:

(1) It may be necessary to adjust the threshold of each individual channel, and weight some channels more strongly than others (pre-emphasis), or to equalize the relative strength/offsets of all channels. Since:

$$\int_0^{TTS(n)} A \cdot Iin(n)\, dt = Cin \cdot Vthresh(n) \quad (8)$$

where TTS(n) is the time-to-spike for channel n, and A is the translinear current gain I2/I3

Rather than fixing A and varying Vthresh(n), in this linear system it is functionally equivalent to fix Vthresh, and instead vary A(n) = I2/I3(n).

(2) Since the input capacitance of each channel Cin is fixed in fabrication and limited by the area available, it may also be necessary to scale all the currents up or down to vary the mean integration time TTS(n) between winning spikes, to match good simulation results at around 50-200μs.

In this design, Cin is fixed at 1pF, and Vthresh = VinhTH = Vncasc = 1.1V. To allow sufficient programming range and also set TTS(n) to be around the 50-200μs range, A(n) = I2/I3(n) was set using 3 digital bits which control I3 (Table 5):

Table 5. I3(n) and current gain A(n) programming table

Setting	[t2..0]	I3 (nA)	Current Gain A
0	000	5.714	7.9
1	001	5.000	9.0
2	010	4.286	10.5
3	011	3.571	12.6
4	100	2.857	15.8
5	101	2.143	21.0
6	110	1.429	31.5
7	111	0.714	63.0

The measured programmability of A(n) using I3 is shown in (Figure 20) below:

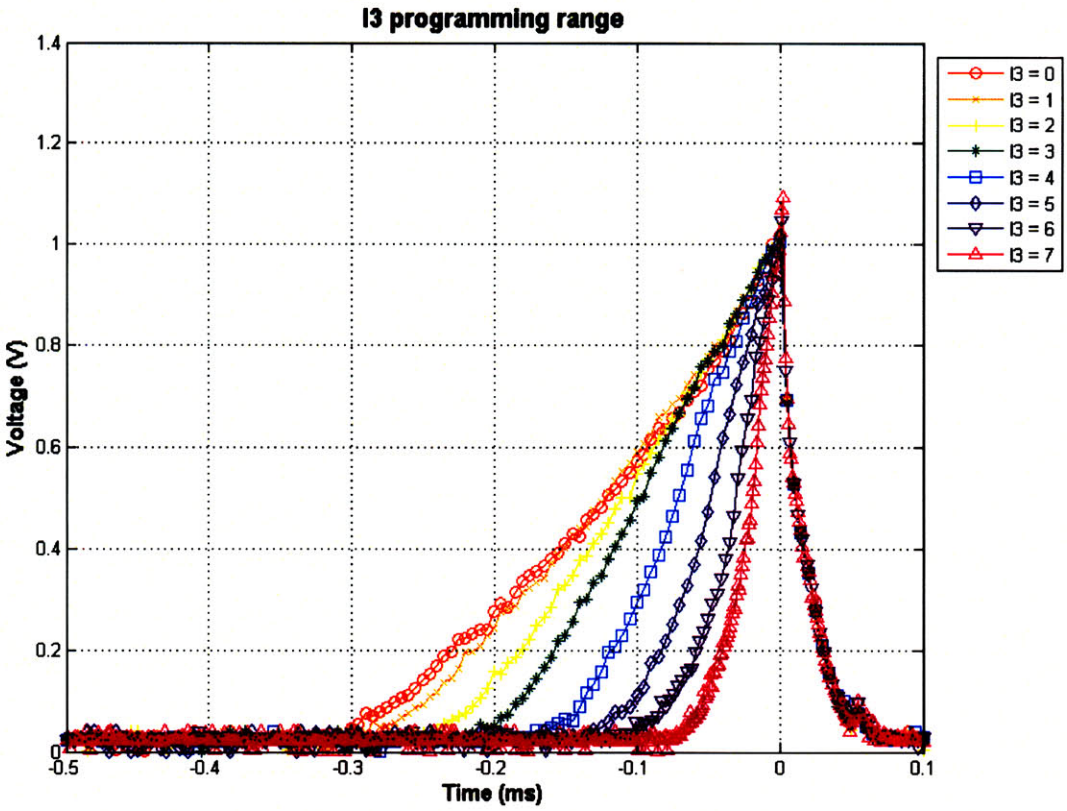


Figure 20. 3-bit programmability of A(n) using I3.

The figure was generated by applying a constant 50nA external DC input, which corresponds to a moderate sound level. The TTS thus covers the desired range.

A subtle point related to the current gain A(n) programmability is that it also determines a finite separation time between adjacent spikes. Since the Reset signal will force all channels to integrate up from ground, the minimum time-to-spike (TTS), which in turn is governed by A(n), sets the minimum separation time between spikes. This is important for 2 reasons:

(1) Since we want to report a single channel address (a 4-bit number from 0..15) and a single logA/D value (a 7-bit number) at every spiking event, we will use tristates to pass an 11-bit digital word onto a common bus. As with all pass-transistor logic, we need to guarantee that the tristates for two or more channels are never on at the same time, since they may "clash" and pass a lot of shoot-through current (a lesson learned from the original ABEP). Hence it is very important to guarantee that the spikes are distinctly non-overlapping.

(2) As mentioned in section 4.2(3), we may want to guarantee that the minimum separation time between spikes is long enough for auditory filter ringing to die out, especially on low frequency channels. We can do this by setting A(n) to a lower value, which lengthens the minimum TTS(n).

4.3.2. A voltage winner-take-all topology with positive feedback

The voltage winner-take-all (WTA) with positive feedback (PFB) is shown in (Figure 21) below.

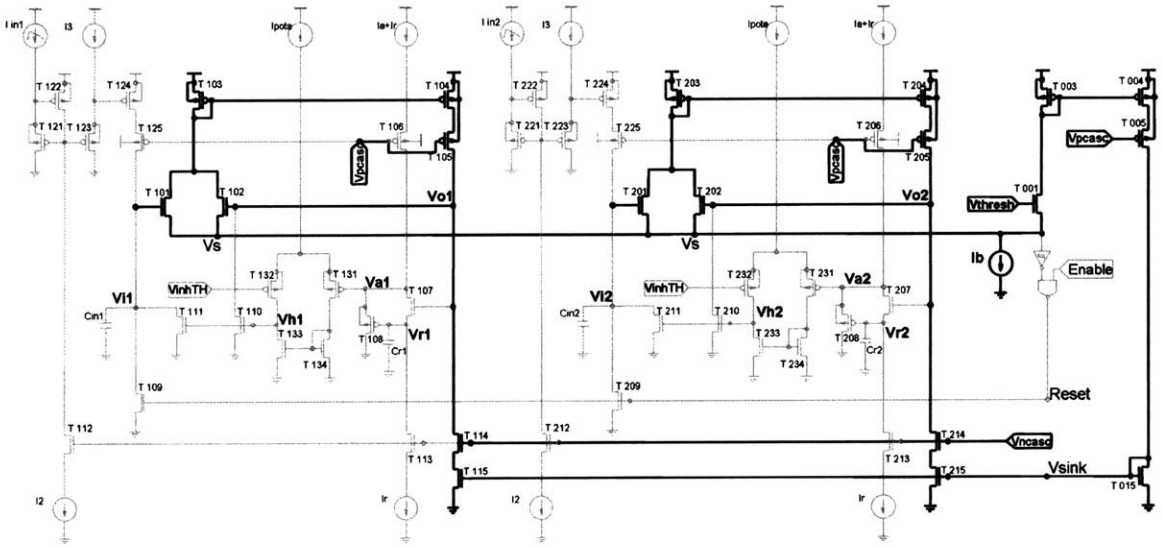


Figure 21. The 2-channel race-to-spike circuit with voltage winner-take-all (WTA) in bold.

This WTA topology with PFB was adapted from (Cauwenberghs and Pedroni 1994), with one improvement: by actively turning off the NMOS current sink transistors of Tn15 when one channel starts to win, the PFB gain is doubled. A very strong WTA action hence enforces the interleaved stimulation requirement in this circuit.

The circuit functions as follows: Vi1 and Vi2 are the voltage-mode inputs to the WTA circuit. Vthresh is the threshold voltage that marks the 'finish-line' in the race-to-spike algorithm. The first channel with an input voltage exceeding Vthresh will initiate a positive feedback loop around Tx03, Tx04, Tx05 and Tx02 (where x denotes the 'winning' channel number) that very quickly drives its output to the rail. Note that initially, Tx15 will be sinking Ib, but as Tx01 and Tx02 start to steal more current away

from T001 by pulling up on Vs, the current sunk by Tx15 will decrease and cause Von to pull up even faster. Hence the PFB gain is effectively doubled (in a small-signal sense). The first channel to pull up on Vs and be reinforced by PFB will thereby suppress all other channels from winning. The strong PFB minimizes the chance that more than one channel can simultaneously steal current from T001, and "win" the race at the same time.

4.3.3. A superbuffer to set attack and release time constants

The superbuffer together with a threshold comparator is shown below in (Figure 22).

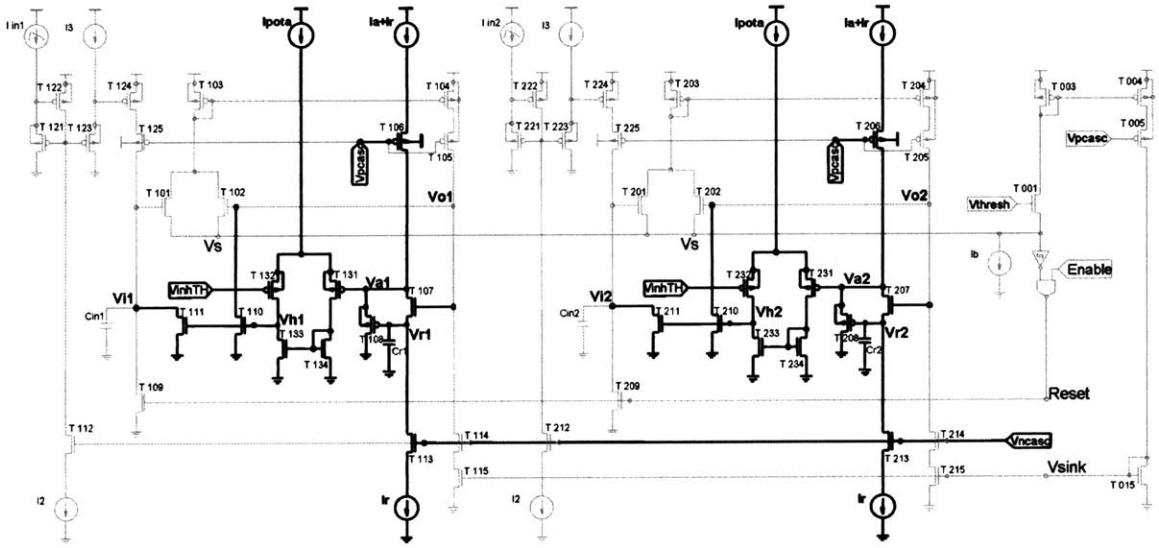


Figure 22. The 2-channel race-to-spike circuit with superbuffer (cascaded N and P source followers) in bold.

This circuit works as follows: once a channel has won (call it channel x), and Vox shoots quickly to the rail due to PFB regeneration, this rising edge is coupled into the superbuffer (cascaded N and P source followers) consisting of transistors Tx07 and Tx08. Due to the presence of a right-half-plane zero, the output from the superbuffer, Vax, actually shoots down (undershoots) before it starts to charge up at a rate of Ia/Cr (all Cr capacitors are set to be the same size of 1pF in this implementation). Intuitively, the right-half-plane zero comes from the negative gain from the gate to drain of Tx07 if the source does not move instantaneously due to Cr holding its value. From a large-signal perspective, once Vox shoots up, Tx07 quickly triodes and shorts Vax to Vrx. Since this causes the Vsg of Tx08 to collapse from ~0.9V to almost 0, Vax undershoots before it

actually starts to rise. As we shall see, this undershoot is essential to set up Ta.

Vax initially sits below VinhTH by design, and is compared against VinhTH by a simple OTA comparator (Tx31, Tx32, Tx33 and Tx34). Vax will charge up from the minima of the undershoot, called Vax(min), and when it crosses VinhTH, the comparator output Vhx goes high and starts the inhibition phase by turning on transistors Tx10 and Tx11. The time from the rising edge on Vox to the tripping of the comparator defines the attack time, $Ta \simeq \frac{Cr \cdot (VinhTH - Vax(\min))}{Ia}$, assuming negligible undershoot time. In this simplest implementation, Tx10 is just a switch that shorts input Vix to ground, much like reset transistor Tx09.

Tx11 is identical to Tx10 except it shorts the output Vox to ground. This is necessary because the PFB has now clamped Vox at Vdd, and without a reset phase on Vox, the WTA would be permanently locked up even after Vix is lowered.

By the same token, once Vox falls due to the action of Tx11, the falling edge is coupled into the superbuffer, which then undershoots (Vax shoots up) to a maxima of Vax(max) before it falls at a rate of Ir/Cr. From a large-signal perspective, the Vsg of Tx08 inflates again now that Tx07 shuts off, and steps up from ~0V back to ~0.9V. Vhx therefore remains high until Vax discharges back down below VinhTH. The time that Vhx remains high, and the inhibition remains on, is defined as the release time $Tr \simeq \frac{Cr \cdot (Vax(\max) - VinhTH)}{Ir}$, again assuming negligible undershoot time. Once Vhx falls, the inhibition on both Vix and Vox turn off rapidly, allowing the channel to receive inputs again.

Measured waveforms for Va are shown in (Figure 23) below:

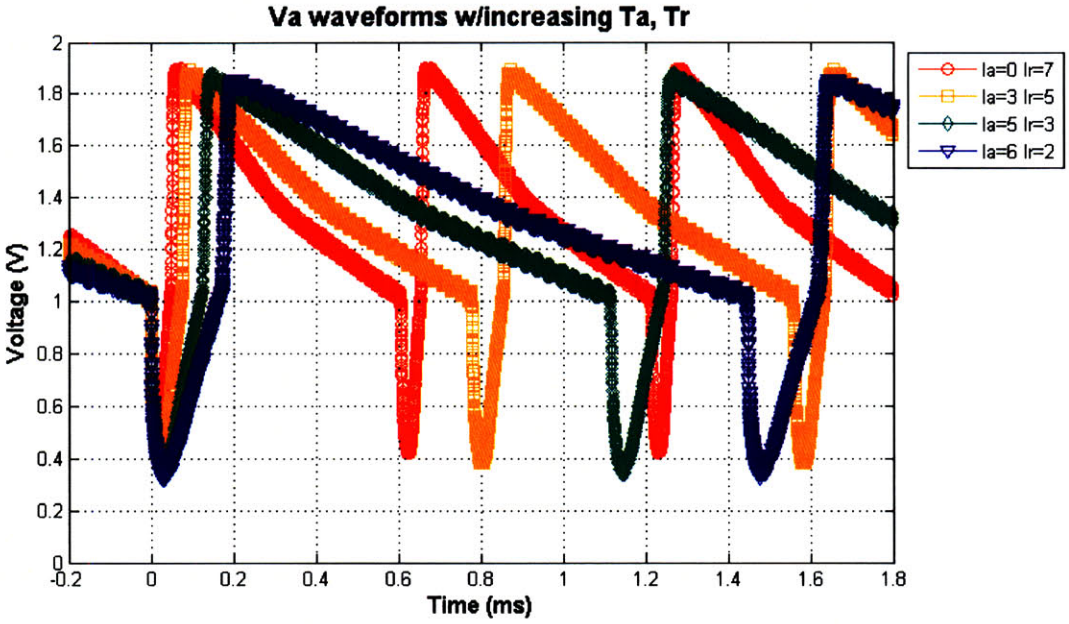


Figure 23. Superbuffer operation showing effect of Ta, Tr programmability on Va

Conceivably, rather than being a switch to ground, Tx10 could be a pass-gate to control a current source instead, which would then implement a soft current-limited inhibition (i.e. a relative refractory period, where a strong input can overcome the inhibition). This was not done simply due to lack of die area for the wiring of an additional current source, and sacrificed because MATLAB simulations predicted only a minor drop in performance without a relative refractory period (i.e. only an absolute refractory period). This implementation thus has a "brick-wall" inhibition-current roll-off, and to model this in MATLAB the roll-off slope k was set to an order of magnitude larger at 10^5, as shown below in (Figure 24).

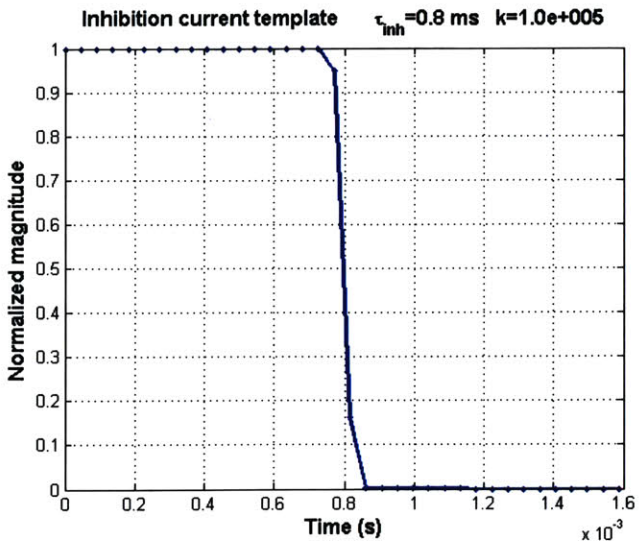


Figure 24. Time profile of inhibition current, with "brick-wall" exponential rolloff (k=1e5)

Similar to A(n), 3-bit programmability of Ta(n) and Tr(n) is given in (Table 6):

Table 6. Ia and Ir programming table

Setting	[bit2..0]	Ia+Ir (nA)	Ta (us)	Ir (nA)	Tr (us)
0	000	45.0	39	0.36	4284
1	001	39.4	45	0.71	2142
2	010	33.8	53	1.07	1428
3	011	28.1	63	1.43	1071
4	100	22.5	79	1.79	857
5	101	16.9	107	2.14	714
6	110	11.3	162	2.50	612
7	111	5.6	334	2.86	536

Ta in (Table 6) was computed at a minimum Ir (bit setting 0) so Ia would not be much different from Ia+Ir. A subtlety in getting this circuit to work robustly was to make sure that Ir was set on its own, since it is much smaller than Ia, and not as the difference between (Ia+Ir) and (Ia), which would be the case if the N source-follower and P source-follower were flipped.

The measured programmability of Ta is shown in (Figure 25) below:

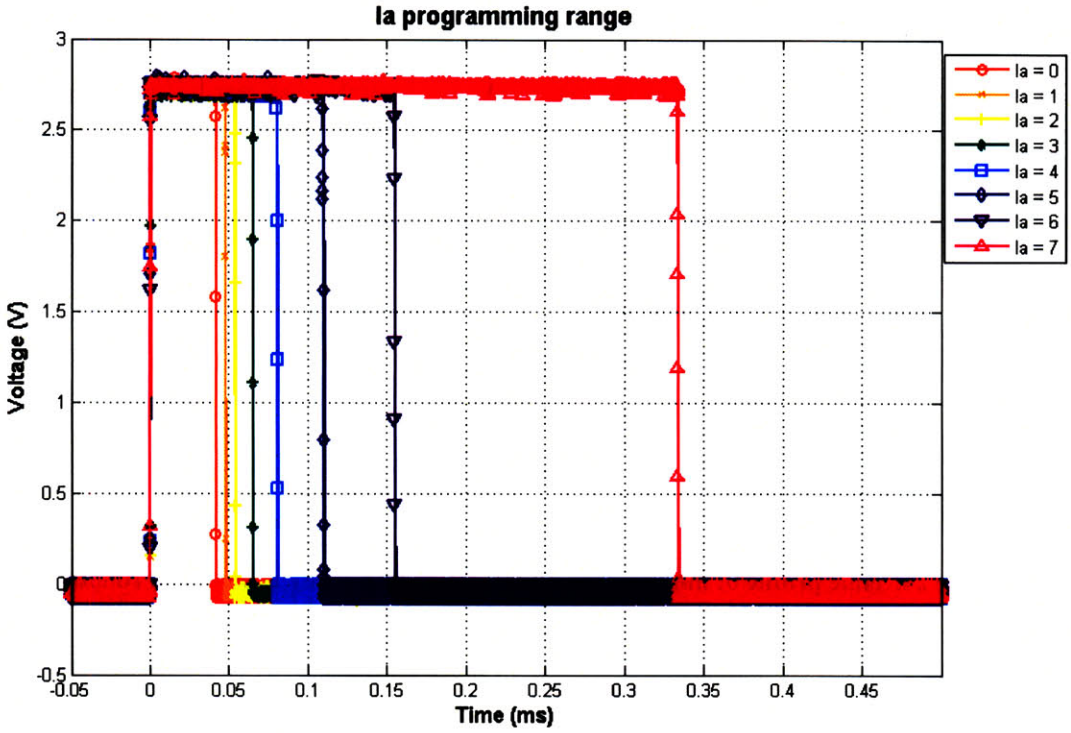


Figure 25. 3-bit programmability of Ta

And the measured programmability of Tr is shown in (Figure 26) below:

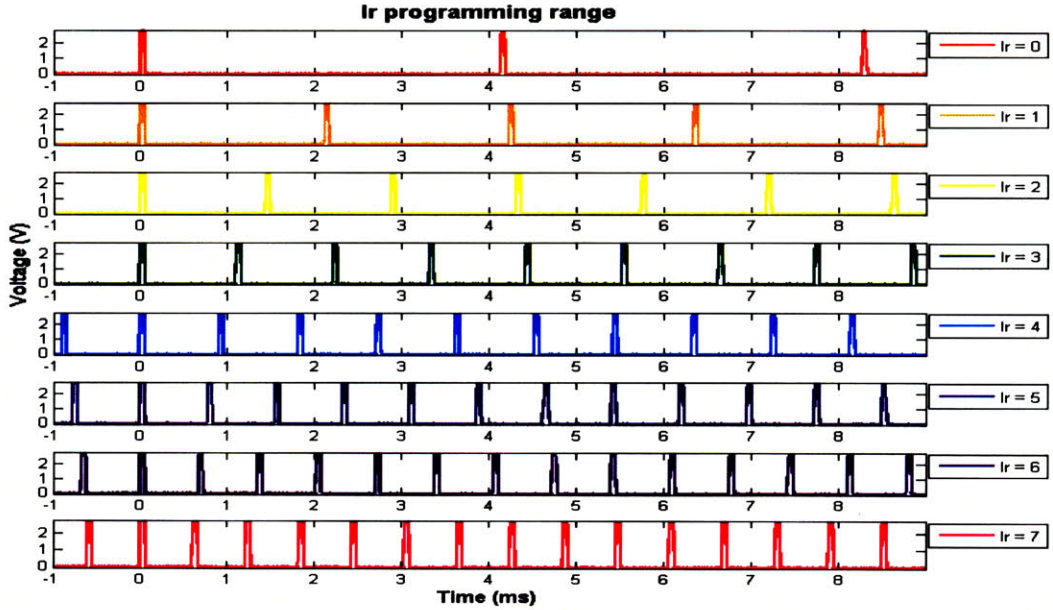


Figure 26. 3-bit programmability of Tr

4.3.4. Superbuffer small signal analysis

An important point was noted during the design and simulation of this circuit, that the parasitic capacitance on node Va (which we will call Ca) needs to be minimized for there to be an undershoot. A "back-of-the-envelope" justification for this can be seen in the following page of my notebook, shown in (Figure 27) below:

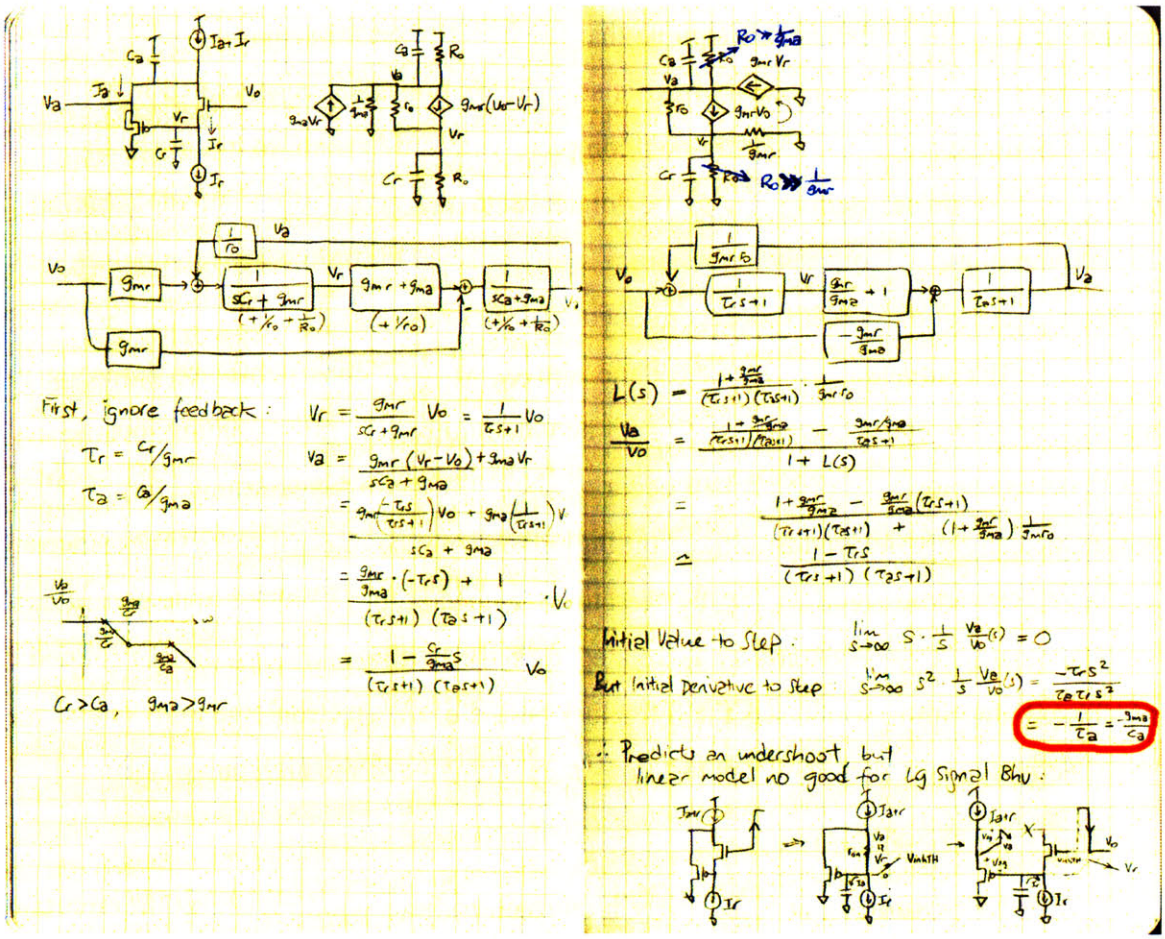


Figure 27. Superbuffer small-signal analysis, scanned from notebook

The right-half-plane zero is derived, and the important result (circled in red) is that the undershoot will have a size which is indicated by its initial slope, inversely proportional to Ca. If Ca is too large, the undershoot may not even appear, causing Ta and Tr to be ill-defined, since the undershoot sets up both Vax(min) and Vax(max). Hence I needed to be mindful in the layout to keep the parasitic Ca small.

It was also important to minimize the parasitic capacitance between Va and Vo to

avoid instability, especially when g_{ma} is small.

4.3.5. Interpretation of Ta and Tr, and "spikes" vs. "pulses"

The attack time constant Ta defines the pulsewidth of a spike. Although this is not defined by the AIS algorithm in MATLAB, a pulsewidth defaults to the sampling period (around 23μs for 44kHz sampling). In a truly asynchronous implementation, however, there is no sampling period and hence a finite pulsewidth must be defined by the hardware. It needs to be long enough so that the Reset transistors can fully discharge all input capacitors Cin(1..16), and a biphasic stimulation pulse that is generated in response to this spiking event must also fire and complete within this time, to avoid overlapping with the next pulse. Setting Ta to be longer than the actual duration of a biphasic stimulation pulse is another way to enforce a longer separation time between pulses such that the specification in 4.2(3) is accommodated.

To avoid confusion, as we have done above, we will consistently be referring to internal spiking events as "spikes" and actual stimulation pulses going out to the electrodes as "pulses," but both will occur at about the same asynchronous time (pulses will be just slightly delayed by a fixed timing offset from the internal spike times since an 11-bit channel address and data value need to be latched before a pulse can be generated).

The release time constant Tr is related directly to the inhibition time τ_{inh} in the MATLAB algorithm, which sets the refractory period of the electronic neuron. Tr and τ_{inh} are effectively equivalent in the circuit implementation due to the high value of $k = 10^5$.

4.3.6. Device Sizing

Device sizing turned out to be fairly important, to have good matching in important components, and also to minimize parasitics in the devices where speed was critical. A schematic with device sizes for channel 1 and channel 0 on the right is shown in (Figure 28) below:

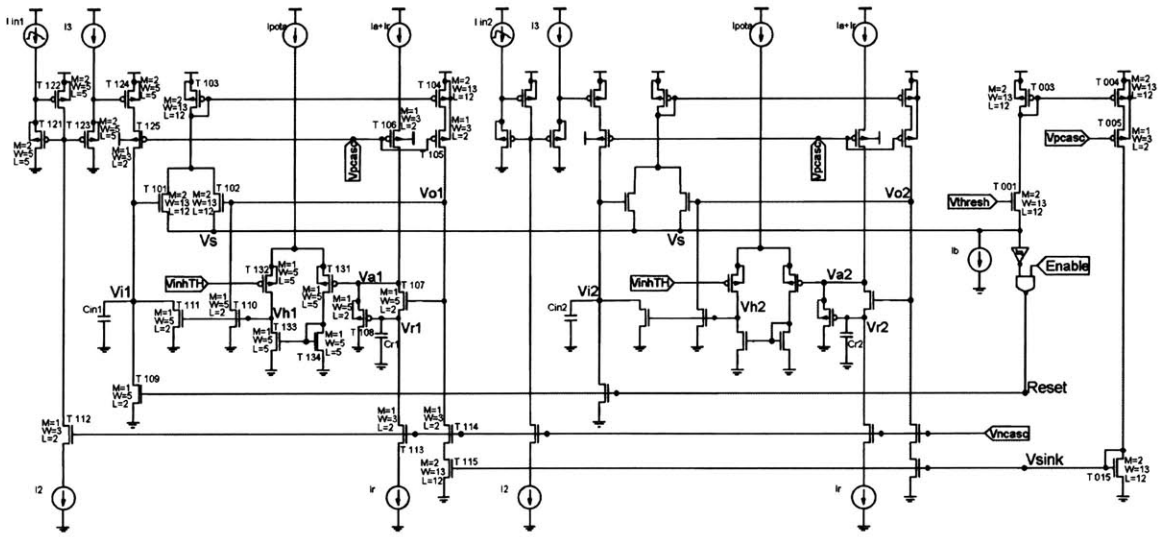


Figure 28. The 2-channel race-to-spike circuit with device sizes shown. Units of W and L are in lambda, with $\lambda = 0.8\mu m$.

Of note:

1. all cascode devices were minimum sized (3/2) for minimum drain capacitance
2. all devices in the threshold comparator are 5/5 for matching
3. superbuffer N and P devices are 5/2 to have well-controlled thresholds, but also small to minimize parasitic capacitance on node Va
4. input NMOS devices have the largest area at 13/12, M=2 for Vth matching across the chip
5. current gain devices are 5/5 for matching but small to minimize parasitic capacitance that would limit the bandwidth of the fast HWR input current

4.4. Sensitivity Analysis

It will be informative to evaluate the sensitivity of this race-to-spike circuit, to answer the following questions:

(1) at what level do closely matched inputs become indiscriminable due to noise and offset?

(2) what (if any) is the [absolute] minimum detectable signal of a single input?

To answer the first question, we must calculate the noise and offset contributions of the various parts. We shall start with the noise in the translinear current gain block.

4.4.1. Noise from input gain stage

The translinear current gain block is reproduced in (Figure 29) below, for ease of reference.

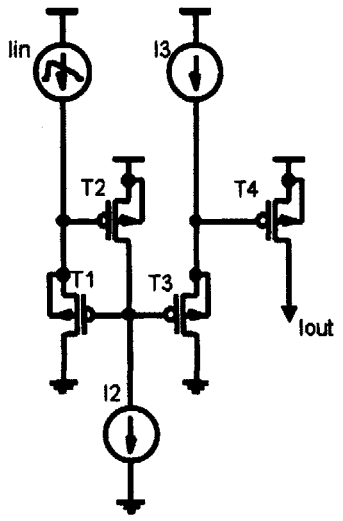


Figure 29. The translinear current gain circuit, in its simplest form.

We now compute N, the number of effective devices, that contribute to current noise at the output:

(1) **Iin**: there are actually 4 noise sources hidden in Iin, due to an N mirror and a P mirror coming from the envelope detector. We will ignore any noise sources from before these mirrors due to the effect of current noise rectification in the envelope detector. Adding T1 gives N1 = 5 noise sources from the input.

(2) **I2**: there are 13 noise sources hidden in I2, 5 from the current reference, and then plus 4 more mirrors with 2 devices each. Adding T2 gives N2 = 14.

(3) **I3**: depending on the DAC setting for I3, which we will call T (from 1..8), there are a total of (T+1) noise sources from the DAC plus 8 more from bias distribution mirrors. I am assuming that the current divider mirrors prior to that, with a factor of 1:7 and 1:18 in the current reference are negligible contributors because of heavy filtering due to low g_m and large gate capacitance in those mirrors. This gives N3 = T+9 = 17 in the worst case.

(4) **Iout**: a single device which carries (I2/I3)·Iin.

The sum of the current noise terms is as follows:

$$\left.\begin{aligned} PSD \text{ of } I_{out} &= \frac{I_2}{I_3} 2qI_{in}N_{in} + \frac{I_{in}}{I_3} 2qI_2N_2 + \left|\frac{\partial I_{out}}{\partial I_3}\right| 2qI_3N_3 + 2qI_{out} \\ I_{out,PSD} &= \frac{I_{in}I_2}{I_3} 2qN_{in} + \frac{I_{in}I_2}{I_3} 2qN_2 + \frac{I_{in}I_2}{I_3^2} 2qI_3N_3 + 2q\frac{I_{in}I_2}{I_3} \\ &= \frac{I_{in}I_2}{I_3} 2q\left(N_{in} + N_2 + N_3 + 1\right) \end{aligned}\right\} \text{Assuming all transistors are in subthreshold}$$

$$= 2q(5+14+17+1)I_{out}$$

$$= 2qNI_{out} \quad \text{where N=37}$$

Since I_{out} is integrated on capacitor C_{in} for time T, up to voltage V_{thresh},

$$Q(t) = \int_0^T I_{out}(t) \cdot dt$$

$$= I_{out}(t) * boxcar[0,T]$$

$$Q(f) = Iout(f) \times H(f) \qquad \text{where } H(f) = \frac{\sin \pi fT}{\pi f} e^{-j\omega\frac{T}{2}}$$

The variance of charge, σ_Q^2 that is integrated can then be calculated:

$$\sigma_Q^2 = \int_0^\infty I_{out,PSD} \times H^2(f) \cdot df$$

$$= I_{out,PSD} \times \frac{1}{2} \int_{-\infty}^{\infty} H^2(f) \cdot df \qquad \text{since } H(f) \text{ is even}$$

$$= I_{out,PSD} \times \frac{1}{2} \int_{-\infty}^{\infty} boxcar^2[0,T] \cdot dt \qquad \text{by Parseval's Theorem}$$

$$= I_{out,PSD} \times \frac{T}{2} = 2qNI_{out} \times \frac{T}{2} = qNI_{out}T = qNC_{in}V_{thresh}$$

$$Hence, \ \frac{\sigma_Q}{Q} = \frac{\sqrt{qNC_{in}V_{thresh}}}{C_{in}V_{thresh}} = \sqrt{\frac{qN}{C_{in}V_{thresh}}} = \sqrt{\frac{1.6e-19 \times 37}{1.1pC}} \Bigg\} for \ C_{in} = 1pF, \ V_{thresh} = 1.1V$$

$$= 0.23\%$$

Since the variance seen across 16 channels can be added:

$$\frac{\sigma_Q}{Q} \text{ (across 16 channels)} = \frac{\sqrt{16\sigma_Q^2}}{Q} = 4\frac{\sigma_Q}{Q} \text{ (across 1 channel)}$$

$$\simeq 0.92\%$$

4.4.2. Input-referred noise from the voltage WTA

The discriminability question askes, what is the input referred noise and offset appearing at each input Vin of each stage?

Let us assume that in the worst case, all the stages have Vin sitting right around Vthresh. We will neglect the noise from PFB transistors, which only serve to reinforce a winner once a winner is found. However, even they are included with Vout also near the balance point of Vthresh, the result is the same: there will be effectively 1 input device carrying Ib/17 in each stage.

The relevant devices are shown in (Figure 30) below:

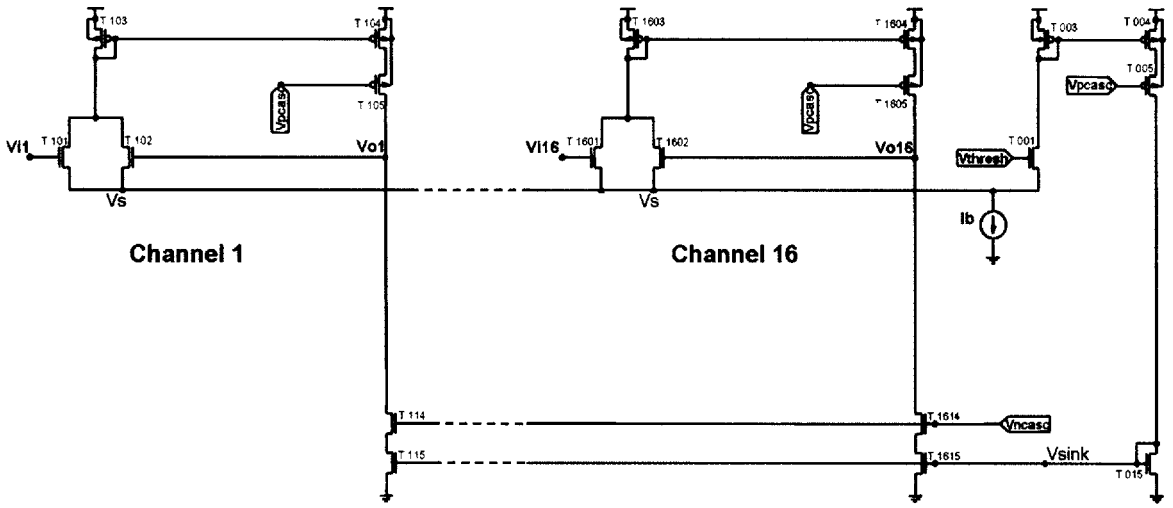


Figure 30. Schematic of WTA for noise analysis

In this worst possible scenario, there are 4 devices per channel (ignoring cascode devices), each carrying Ib/17, that contribute noise to its respective output node Vo1..16. At each output node Vo, we can assume there are only 8 devices contributing noise: 4 devices from the channel itself, plus 4 more from "channel 0."

The sum of the current noise terms is hence 8·2q(Ib/17), assuming all devices stay in weak inversion, and the input-referred noise variance can be calculated as follows:

$$V_{in,PSD} = \frac{8 \cdot 2q(I_b/17)}{g_m^{\,2}}$$

$$\sigma_{Vout}^{\,2} = \int_0^\infty V_{in,PSD} \times H^2(f) \cdot df$$

$$= V_{in,PSD} \times Gain^2 \times BW \times \frac{\pi}{2}$$

$$= \frac{8 \cdot 2q(I_b/17)}{g_m^{\,2}} \times g_m^{\,2}(g_m r_o)^2 r_o^{\,2} \times \frac{1}{2\pi(g_m r_o) r_o C} \times \frac{\pi}{2}$$

$$= \frac{8 \cdot 2q(I_b/17)(g_m r_o) r_o}{4C}$$

$$= \frac{4q(g_m r_o) V_A}{C} \qquad \text{where } V_A = (I_b/17) r_o \simeq 50V \text{ and } g_m r_o \simeq 200 \text{ for our process}$$

$$\sigma_{Vout} \simeq \sqrt{\frac{4 \times 1.6e-19 \times 200 \times 50V}{200fF}} \qquad \text{where } C \simeq 200fF$$

$$= 180\ mVrms$$

Input-referring σ_{Vout} back to the input by the gain $(g_m r_o)^2$

$$\sigma_{Vin} = \frac{\sigma_{Vout}}{(g_m r_o)^2} \simeq 4.5\ \mu Vrms$$

Since the variance seen across 16 channels can be added:

$$\sigma_{Vin} \text{ (across 16 channels)} \quad = \frac{\sqrt{16\sigma_{Vout}^{\ 2}}}{g_m r_o} = 4\sigma_{Vin} \text{ (across 1 channel)}$$
$$\simeq 18\ \mu Vrms$$

And thus $\frac{\sigma_{Vin}}{V_{thresh}} \simeq \frac{0.018mV}{1100mV} \simeq 0.002\%$

4.4.3. Input-referred offset from the voltage WTA

Assuming drain current mismatch errors on the order of $\sigma_{Id} / Id \approx 2\%$:

$$\sigma_{Vofs}^{\ 2} = 8 \text{ devices} \times \left(\frac{\sigma_{I_d}}{I_d} \text{ of } I_d\right)^2 \times \frac{1}{g_m^{\ 2}}$$
$$= 8 \text{ devices} \times (2\%)^2 \times \left(\frac{\phi_T}{\kappa}\right)^2$$
$$\sigma_{Vofs} = \sqrt{8} \times 0.02 \times \frac{\phi_T}{\kappa}$$
$$\simeq 2.1\ mVrms$$

Since the variance seen across 16 channels can be added:

$$\sigma_{Vofs} \text{ (across 16 channels)} \quad = \sqrt{16\sigma_{Vofs}^{\ 2}} = 4\sigma_{Vofs} \text{ (across 1 channel)}$$
$$\simeq 8.4 \text{ m}Vrms$$

And thus $\frac{\sigma_{Vofs}}{V_{thresh}} \simeq \frac{8.4mV}{1100mV} \simeq 0.76\%$

4.4.4. Vth variation due to doping

Due to layout constraints, each channel's input pair could not be laid out in common centroid or even in the same region of the die. This is because the output stage in the ABEP was previously entirely digital, and the only space available for the AIS circuitry was in the place of the old output stage.

Thus an extreme Vth variation could be as bad as ~100mV in the worst case, say between opposite ends of the chip. With Vthresh = 1.1V, the percentage error mismatch would then be ~0.1/1.1 ≈ 9.1%. This error will clearly dominate the ability of the WTA to discriminate between closely matched inputs, but is non-fundamental – a layout that dedicates an area for the entire WTA structure to be well matched would avoid this.

4.4.5. Summary of noise and offset analysis

A summary of the noise and offset contributions that limit the 16-channel discriminability of the circuit are shown in (Table 7) below:

Table 7. Summary of noise and offset errors in race-to-spike circuit

Error Source	Noise (random offset)	Mismatch (systematic offset)
Input gain stage	0.92%	
Input-referred WTA noise	0.002%	
2% current matching		0.76%
100mV Vth variation		9.10%
Total:	0.92%	9.86%

In conclusion, the noise is dominated by the input gain stage, but mismatch plays a much larger role than noise, and we can expect that the circuit may make an error when it has to decide between inputs which are within ~10% of each other.

4.4.6. The minimum detectable signal

A very interesting feature of the "race-to-spike" is that it has no minimum detectable signal. This may or may not be obvious, but a side-by-side comparison between an amplifier and an integrating neuron should make it very clear:

Consider an amplifier, with input u, gain of A and output voltage V:

$Au = V$

Assuming there is some fixed noise variance σ_V at the output, what is the input-referred σ_u?

$du = \frac{\partial u}{\partial V} dV$

$\sigma_u = \frac{u}{V}\sigma_V$ (variances treated as differentials)

$= \frac{u}{Au}\sigma_V = \frac{\sigma_V}{A}$

Note that as u decreases, the output V goes down with it as $V = Au$. Thus σ_u remains constant, fixed at $\frac{\sigma_V}{A}$. σ_u is the minimum detectable input below which output V is lost in the noise σ_V.

Consider an integrating neuron, with input i, current gain A, capacitance C, time-to-spike T and capacitor voltage V:

$Ai = \frac{CV}{T}$

Assuming there is some fixed noise variance σ_V on the capacitor voltage, what is the input referred σ_i?

$di = \frac{\partial i}{\partial V} dV$

$\sigma_i = \frac{i}{V}\sigma_V$ (variances treated as differentials about a mean)

$= \frac{C}{AT}\sigma_V$

Note that as i decreases, $V = Vthresh$ remains constant. Alternatively, as i decreases, T increases. Thus σ_i decreases together with the signal i. σ_i stays at a fixed fraction of the input ($\frac{\sigma_V}{V}$), so there is NO MDS.

The significance of this is that even very faint signals can be detected by an integrating neuron, given enough time to spike (T). So noise limits an integrating neuron not in signal detection, but only in discrimination between closely matched signals.

4.5. *Digital output protocol*

A custom digital protocol was designed to send the asynchronous timing events off-chip. Within each channel, the analog rising and falling edges on Vo were converted to digital edges using the following circuit in (Figure 31):

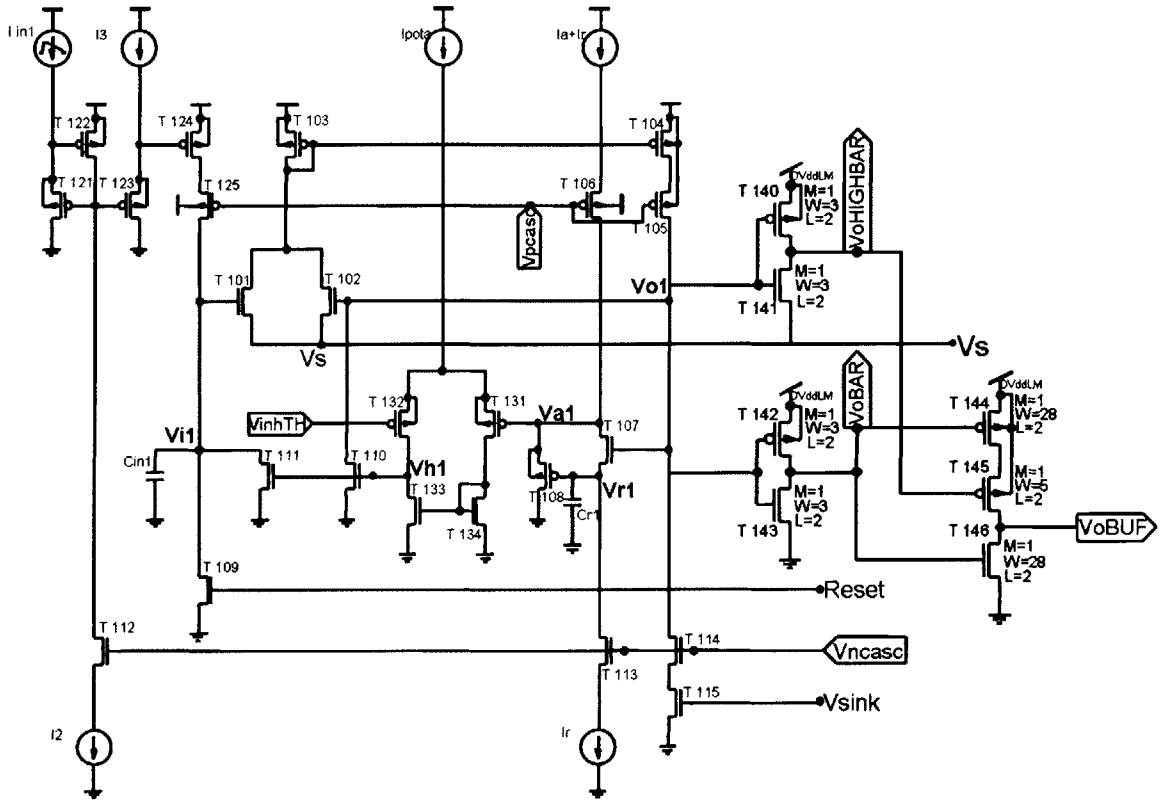


Figure 31. Creation of digital output signal, VoBUF

The addition of signal VoHIGHBAR was added as an extra safety measure to prevent two channels from ever spiking at the same time. It was noted that if two inputs, say ViA and ViB crossed Vthresh very close to each other, both their outputs VoA and VoB would rise. Thus VoA and VoB may both cross the inverter threshold before the PFB would have time to kick in and force the slightly lower one back down. To prevent a spurious digital event in this case, VoHIGHBAR only falls low enough to enable VoBUF if Vo rises sufficiently above Vs. It is highly unlikely for Vo to rise above Vs unless it is truly the winner. In simulation, this eliminated all the cases with multiple spikes, and appears to be borne out in measurement as well.

4.5.1. Multiplexing of channel data using a bank of tristates

In (Figure 32) below, we show how channel data (E6..0 and Ch3..0) is muxed onto a common bus (F6..0 and G3..0).

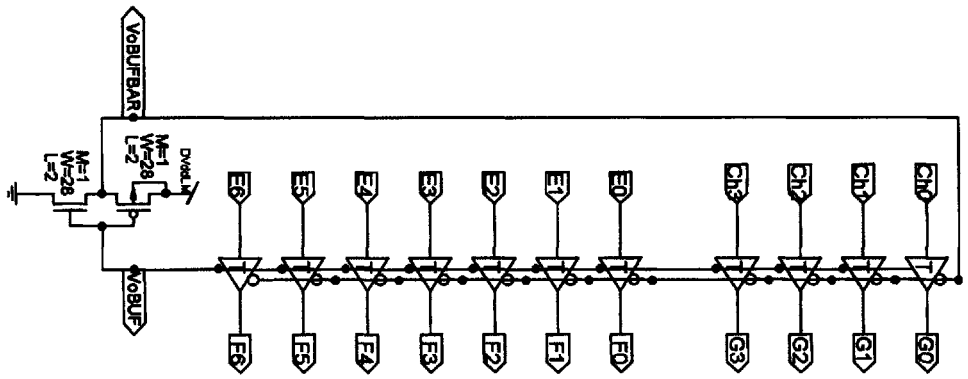


Figure 32. Tristates passing channel data onto a common bus

Ch3..0 are hardcoded to identify each channel with a channel address from 0..15, and so implements a 16-to-4 bit decoder without having to look at all 16 VoBUF signals. This then allows us to use a single pin (i.e. the Reset signal) to report the presence of an asynchronous event, and then have G3..0 identify the channel that spiking event was in. The inverse of Reset is brought off-chip as a signal called VresetBAR.

The combination of VresetBAR and the common bus (F6..0 and G3..0) provides an efficient way to report the asynchronous events together with channel data and address, for latching by an external microcontroller or instrument.

VoBUF1..16 are also reported off chip, for ease of interfacing with 16 independent biphasic pulse generators. Each pulse generator would then listen only to its own channel, and whenever a spike is received, latch F6..0 on the rising edge of VresetBAR and asynchronously generate a stimulation pulse of that magnitude.

4.5.2. Modification of LogADC output latches

Since the logADC in the ABEP is running on a clock that is independent of the AIS output stage (Sarpeshkar, Salthouse et al. 2005), the AIS tristates must be allowed to grab the most recent data bits (E6..0) out of the logADC at any time. To be compatible with this, the logADC needed to be modified slightly so that the CLEAR signal for the data

latches which is exerted during the Integration phase would only zero the input, but not clear the output. This modification is shown in (Figure 33) below:

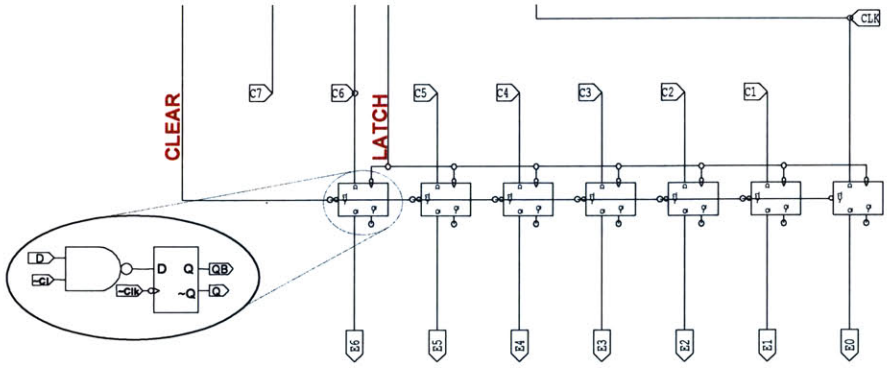


Figure 33. Modified LogADC latches to be compatible with AIS.

In this way, the old output is preserved until a new LATCH signal is generated to refresh the latches with new data. The AIS tristates are then guaranteed to receive valid data which is at most one cycle old. Since the LATCH signal occurs simultaneously on all 7 bits and is delayed past the clock edge, the bits are either all old or all new without any possibility of ripple error, and there is only a tiny probability that F6..0 will pass metastable data to be latched (i.e. if LATCH occurs just when VoBUF shuts off).

4.6. Testing strategy and results

A picture of the test setup is shown in (Figure 34) below.

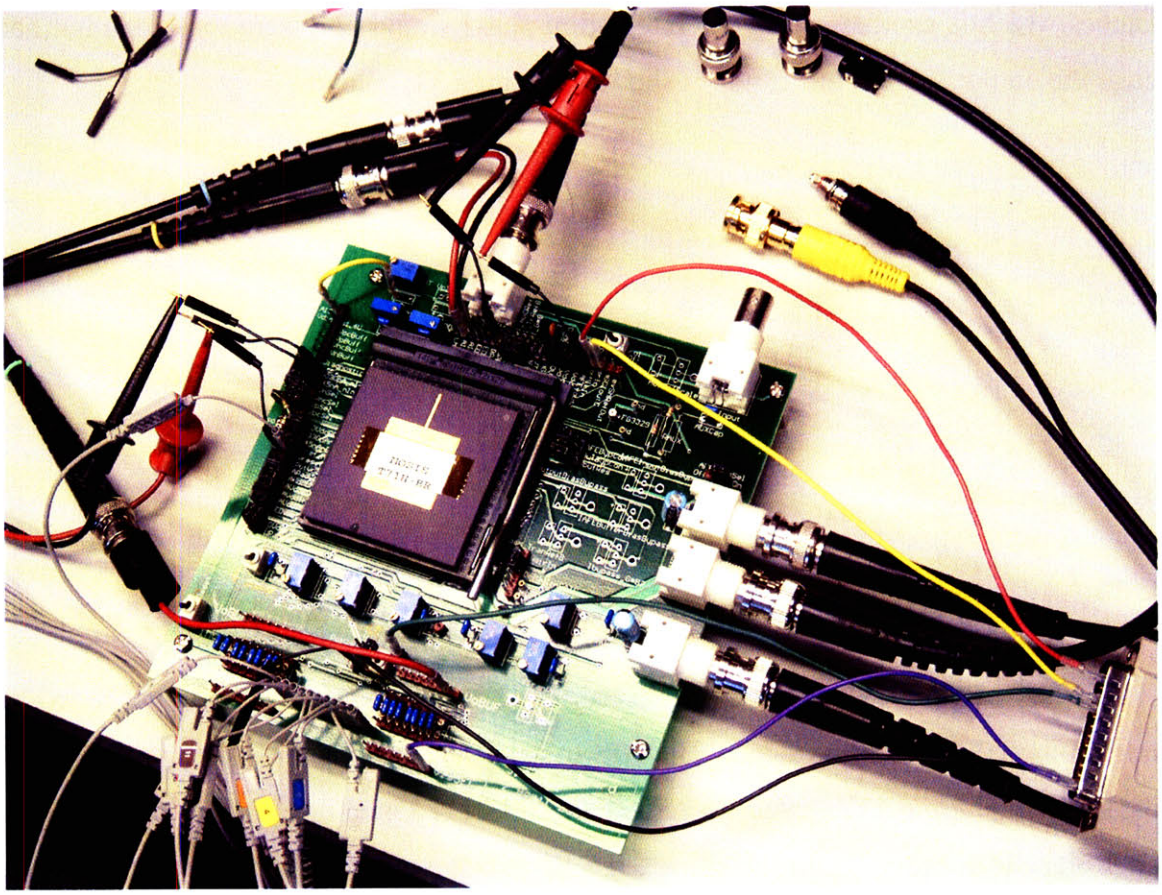


Figure 34. AIS ABEP test setup.

4.6.1. 3-wire Serial Peripheral Interface (SPI) programming

The chip was programmed using the DAQ (data acquisition) toolbox in MATLAB, which has native support for controlling all 17 lines in the parallel port as independent digital I/O lines. This is more than sufficient, since the AIS ABEP was designed to only utilize 3 wires for digital programming: srClkEnable, srClk and srIn (where sr stands for shift-register).

- srClkEnable is an active-low pin which gates srClk off when it is raised high (simply implemented with a NOR gate).
- srClk causes a new bit of data from srIn to be shifted into a 546-bit long shift register, on its rising edge.

The 546 bits which program the chip (18 global + 33×16 channels) are described in (Table 8) below:

Table 8. SPI bit order for AIS ABEP programming

Circuit parameter	# Bits	Bit order	Polarity
AGC	**16 global**		
lattack	4	0..3	n
lrelease	4	0..3	n
G1	4	0..3	n
Imax	4	0..3	n
BPF	**14/channel**		
taus	7	0..6	n
qs	7	0..6	n
ED	**7/channel**		
rel	3	0..2	p
atk	2	0..1	n
xp5	1	0	n
x2	1	0	n
LogADC	**3/channel**		
Iref	3	0..2	p
AIS	**9/channel**		
Ia	3	2..0	p
I3	3	2..0	p
Ir	3	2..0	n
LogADC stimrate	**2 global**		
SR	2	0..1	p

The MATLAB code used to prepare and shift in these bits is attached in an Appendix.

4.6.2. Using the Logic Analyzer

Using the Agilent logic analyzer 1672G in the "State" capture mode, a word of channel address and data can be captured on each rising edge of the VresetBAR signal, as described earlier. The logic analyzer is able to capture from 4096 to 65536 events in its memory, and I found that 4096 events is enough to encode between 0.5 – 1.5s worth of signal, depending on the signal complexity (more for music, less for speech). At 4096 events every second, this is about 11bits data + 32bits timing = 43 bits/event × 4096 events / 1s = 176 kbps bit rate.

However, this assumes that the timing data needs to be transmitted as bits (like the output from the logic analyzer). If the timing information is extracted from the signal itself as time between asynchronous events (like the input to the logic analyzer) then the

timing information is provided by the receiver's internal clock. In this case, the bit rate goes down to 11 bits/event × 4096 events / 1s = 45 kbps. This is favorably compared to conventional fixed-rate encoding (e.g. PCM), which at ~minimum rate and precision of 8 kS/s × 8 bits/sample, will take 64kbps.

An idea similar to this has been exploited in vocal-tract vocoding in cellphones, which can reduce the bit-rate to ~3kbps for efficient transmission by projection onto a dictionary of kernel filters with coefficients adaptively defined by Linear Predictive Coding (LPC). So even though we don't do as well as LPC in cellphones, as a benchmark we are still doing better than conventional PCM.

4.6.3. Filter calibration curves

By playing a pure tone into the AIS ABEP and recording 4096 events in the logic analyzer over a range of frequencies from 100 to 10kHz, the frequency response curves for each channel filter were captured. At each frequency and for each channel, the mean of the samples collected was taken, and in cases where a channel did not put out a spike within the 4096 events, a point would not be plot but just connected with the next available data-point.

After some manual adjustment of the tau and iref bits, a plot of roughly equally-spaced and equal-height filter channels was generated, shown in (Figure 35) below.

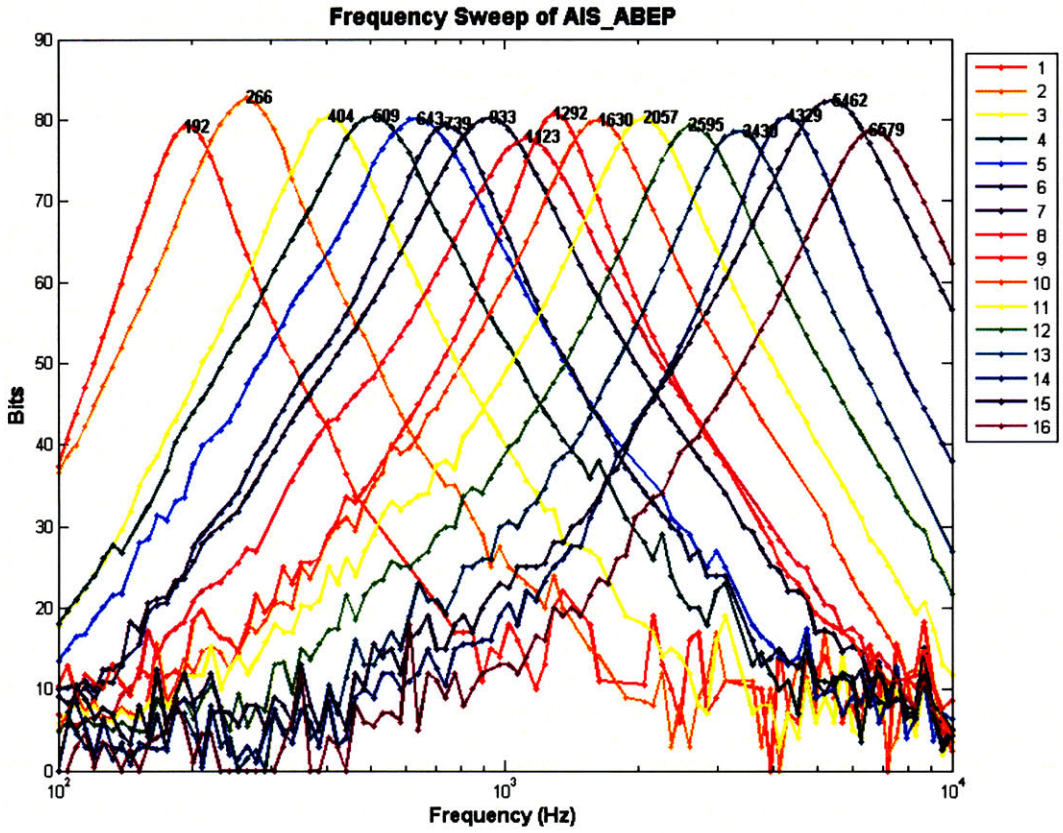


Figure 35. Plot of calibrated filters, taken using 200mVpp tones input directly to the channels.

4.6.4. Sound reconstruction results

The Agilent 1672G has a mixed acquisition mode called "State+Timing" capture, which was used to obtain asynchronous events timestamped by the logic analyzer's 100MHz internal clock. It thus amply meets the timing resolution required (~50μs or faster).

Playing a variety of sound clips into the chip, I performed a side-by-side comparison with the AIS algorithm in MATLAB. The same sound clip 'die' as in section 2 was used. The MATLAB simulation results are slightly different from before because of the 'brick wall' inhibition doing away with the relative refractory period.

In (Figure 36) below, individual channel reconstructions are shown, with the original spikes overlaid in gray:

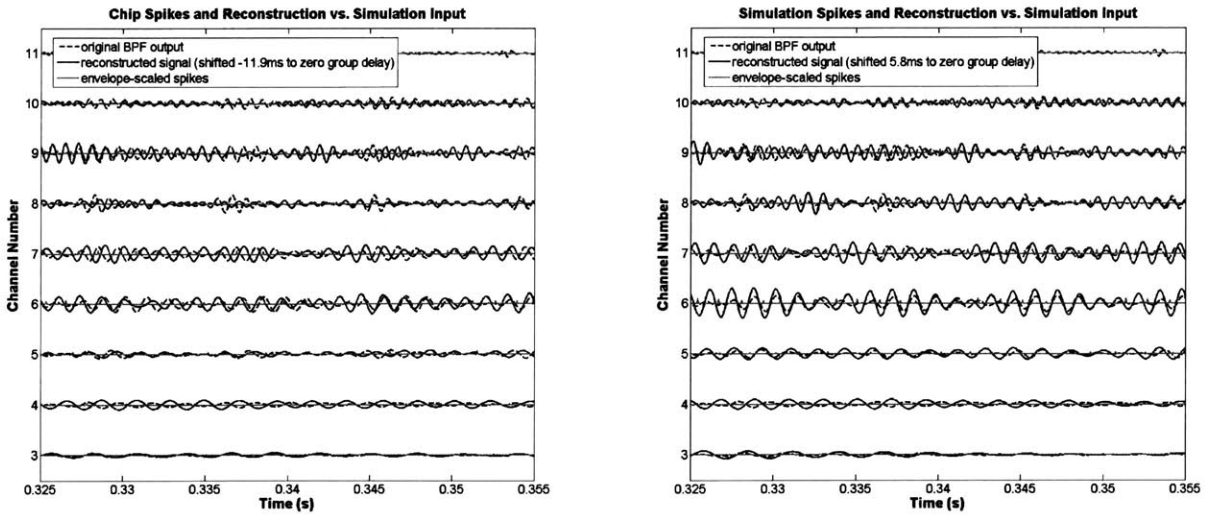


Figure 36. Channel by channel reconstructions, comparing chip data (on left) with simulation data (on right).

In (Figure 37) below, we zoom out to see the spectral features at full-scale, and we see that they are also quite similar.

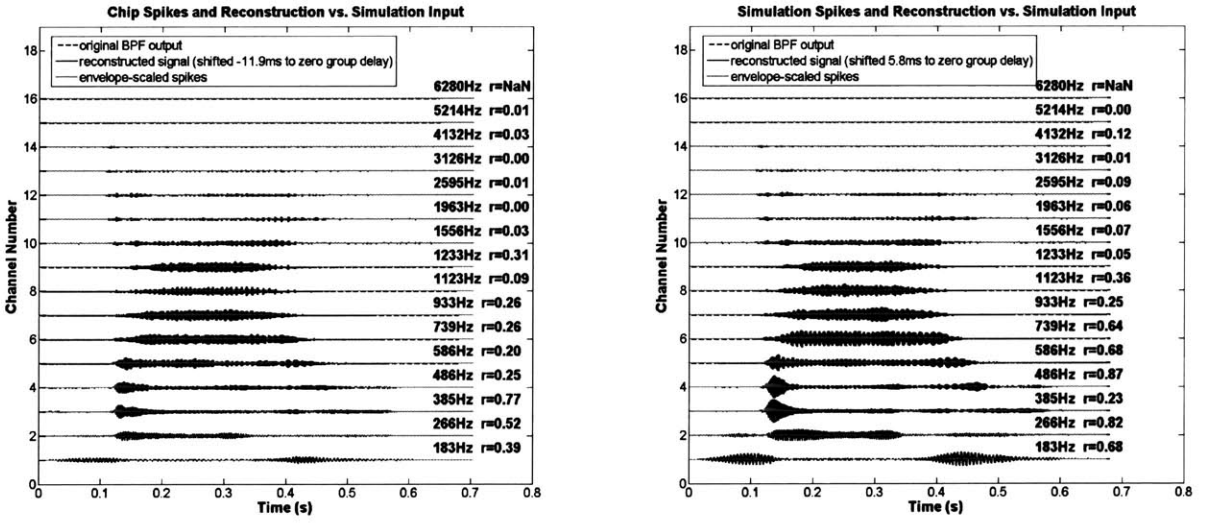


Figure 37. Channel by channel reconstructions, zoomed to fullscale and showing correlation coefficients for each frequency.

Finally, we sum the channels together to look at the composite reconstruction and its correlation coefficient. The two pictures are very comparable, shown in (Figure 38) below:

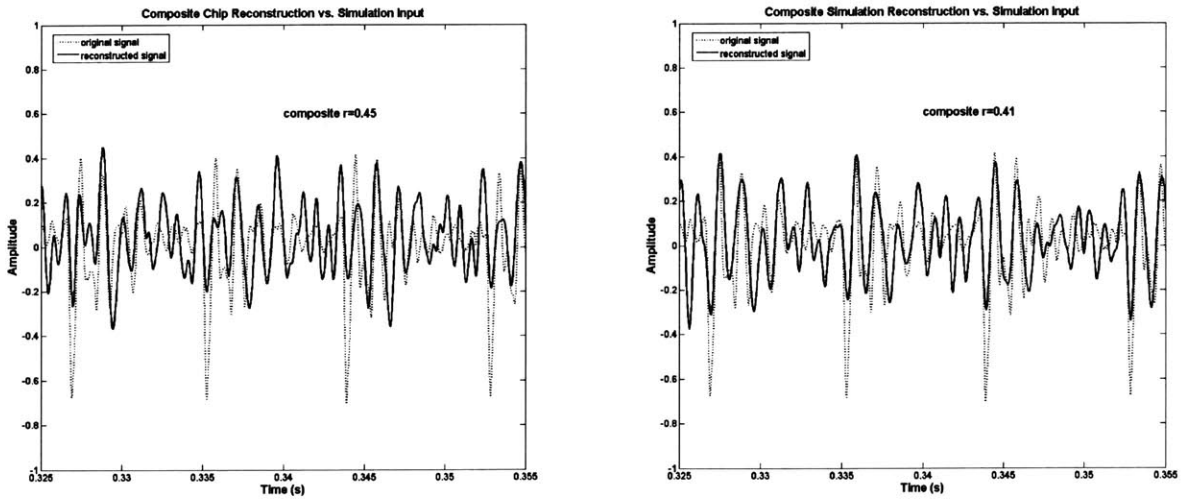


Figure 38. Comparison of composite reconstruction (all channels summed together).

The composite correlation coefficient is actually a little higher for the chip reconstruction, but should only indicate that the two are of comparable quality.

(Table 9) below shows a summary of the results from 11 sound clips processed by the chip and by MATLAB simulation:

Table 9. Summary of results comparing chip performance against MATLAB simulation

Sound clip	Chip		Sim	
"Die"	Env: 0.59	AFR: 231 Hz	Env: 0.56	AFR: 252 Hz
	Phase: 0.28	kbps: 40.7	Phase: 0.17	kbps: 44.3
	Total: 0.45	Acq: 0.705s	Total: 0.41	Len: 0.680s
"Buy"	Env: 0.53	AFR: 221 Hz	Env: 0.59	AFR: 256 Hz
	Phase: 0.22	kbps: 38.8	Phase: 0.11	kbps: 45.1
	Total: 0.42	Acq: 0.621s	Total: 0.41	Len: 0.547s
Jazz "Highway Blues"	Env: 0.29	AFR: 161 Hz	Env: 0.22	AFR: 228 Hz
	Phase: 0.30	kbps: 28.3	Phase: 0.27	kbps: 40.2
	Total: 0.41	Acq: 3.53s	Total: 0.39	Len: 3.50s
Beethoven instrumental (Symphony #9)	Env: 0.37	AFR: 386 Hz	Env: 0.41	AFR: 446 Hz
	Phase: 0.26	kbps: 67.9	Phase: 0.27	kbps: 78.5
	Total: 0.46	Acq: 3.01s	Total: 0.40	Len: 3.00s
Handel chorus "Hallelujah"	Env: 0.28	AFR: 439 Hz	Env: 0.43	AFR: 511 Hz
	Phase: 0.27	kbps: 77.3	Phase: 0.31	kbps: 90.0
	Total: 0.46	Acq: 1.50s	Total: 0.54	Len: 1.47s
"The boy did a handstand"	Env: 0.68	AFR: 212 Hz	Env: 0.67	AFR: 269 Hz
	Phase: 0.14	kbps: 37.3	Phase: 0.21	kbps: 47.3
	Total: 0.47	Acq: 1.82s	Total: 0.53	Len: 1.79s
"The kitchen clock was wrong"	Env: 0.67	AFR: 175 Hz	Env: 0.65	AFR: 226 Hz
	Phase: 0.13	kbps: 30.8	Phase: 0.13	kbps: 39.8
	Total: 0.43	Acq: 1.72s	Total: 0.56	Len: 1.68s
"The nervous driver got lost"	Env: 0.65	AFR: 169 Hz	Env: 0.59	AFR: 231 Hz
	Phase: 0.07	kbps: 29.8	Phase: 0.12	kbps: 40.7
	Total: 0.44	Acq: 2.07s	Total: 0.48	Len: 1.98s
"The young people are dancing"	Env: 0.66	AFR: 150 Hz	Env: 0.55	AFR: 201 Hz
	Phase: 0.15	kbps: 26.3	Phase: 0.16	kbps: 35.4
	Total: 0.45	Acq: 1.63s	Total: 0.46	Len: 1.59s
"They took some food outside"	Env: 0.61	AFR: 187 Hz	Env: 0.63	AFR: 242 Hz
	Phase: 0.11	kbps: 32.9	Phase: 0.13	kbps: 42.6
	Total: 0.44	Acq: 1.66s	Total: 0.41	Len: 1.64s
"They waited for an hour"	Env: 0.68	AFR: 198 Hz	Env: 0.56	AFR: 262 Hz
	Phase: 0.15	kbps: 34.9	Phase: 0.12	kbps: 46.2
	Total: 0.46	Acq: 1.52s	Total: 0.39	Len: 1.47s

Env, Phase and Total refer to correlation coefficients. kbps is the bit rate, Acq is the acquisition time, and Len is the total duration of the sound clip.

These clips can also be listened to via a link from http://avnsl.mit.edu/AIS, and are of demonstrably high quality.

4.6.5. Interspike interval histograms

One final measure of performance that needs to be considered is the interspike interval histogram, to show differences between the distribution of spikes among the channels. The ISI histograms for 'die' are shown in (Figure 39) below:

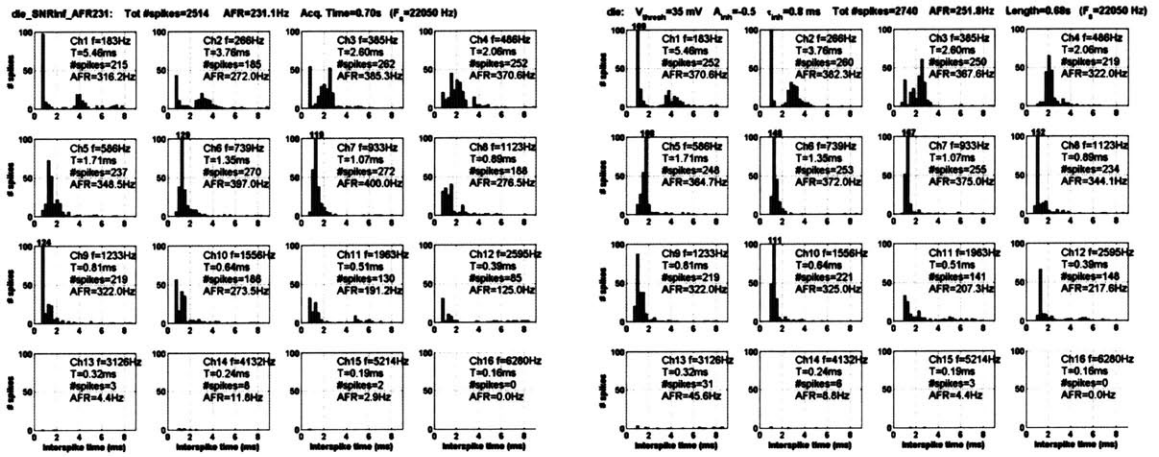


Figure 39. ISIH comparison for the sound clip 'die'

The distributions look fairly similar, with only minor variations. Comparing the ISIH for a piece of music, we find in (Figure 40) that the histograms are much denser and an exponential roll-off is much clearer, but chip and simulation are also similar:

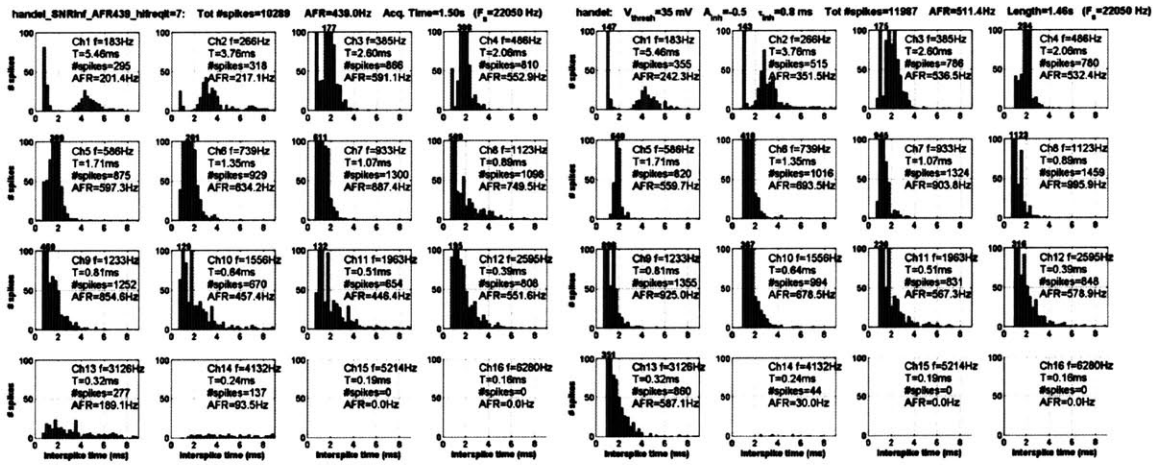


Figure 40. ISIH comparison for the sound clip of Handel's Messiah

Since the original sound clip of Handel's Messiah was sampled at 8 kHz, it seems reasonable to expect no frequency content higher than 4 kHz. However, the chip's 3 kHz filter (Ch13) seems to fare worse than simulation (which has a strong peak and smooth exponential drop), which might account for the slightly poorer performance for music. This might be due to a bandwidth limitation in the half-wave-rectified input, Ihwrout, but does not impair performance by too much.

4.6.6. Power Consumption

The power consumption of the AIS ABEP is shown in (Table 10):

Table 10. Power consumption in various subsystems of the AIS ABEP.

Subsystem	Current (uA)	Power (uW)	Increase over ABEP (uW)
Microphone Front End	38	106.4	0
Automatic Gain Control	10	28.0	0
Bandpass Filters	13.8	38.6	1
Envelope Detectors	12.6	35.3	17
Off-chip input buffers	2.4	6.7	7
Current Biasing	16.5	46.2	42
Log ADC + AIS circuits	15.6	43.7	40
Digital circuits	18.6	52.1	
Total	127.5	357.0	107

The increase in 107 μW over the ABEP is partly due to lack of optimization in the bias distribution – no attempt to conserve power was made when distributing current biasing to the new AIS circuits. Hence there were several more layers of current mirrors added in each channel, which quickly multiplied up to a large increase in biasing power.

The WTA in the AIS circuit consumes 2 μA, or 5.6 μW, with 1 μA supplied externally from a pot (the internal 1 μA bias was insufficient).

The digital power is quoted at a clock frequency of 128 kHz for 1 kHz sampling, and would be doubled for 2 kHz sampling, and halved for 0.5 kHz sampling.

Finally, as in the ABEP, an additional 2 μA needed to be added to the voltage buffers to lower the noise coupled onto the internal voltage references. The 2 NMOS references, Vn and Vnc in particular cause circuit performance to be affected by the clock unless the buffers are given more power.

4.7. Summary of contributions

We have demonstrated that the AIS architecture is both feasible and practical to be implemented in custom analog electronics. The design is sufficiently low-noise and is sufficiently programmable to achieve performance which is comparable to simulation. With a few refinements, it should have the potential to be used in patient stimulation.

My particular contributions were in:

1. Designing the WTA circuit with enhanced PFB (from Vsink) to perform channel selection.
2. Designing the super-buffer timing circuit to set spike pulsewidth (Ta) and absolute refractory period (Tr).
3. Designing the translinear gain input stage for equalization / pre-emphasis of channels.
4. Solving all the system integration, programming and output interface issues.

5. Charge Balanced Stimulation

5.1. Background on electrodes

As mentioned in the introduction, the second part of my thesis deals with providing charge-balanced stimulation at physiologically safe levels, without using a DC blocking capacitor but instead performing precise balancing of current pulses in electronics.

When thinking about the problem of delivering balanced charge in a biphasic current pulse, it is helpful to have in mind the electrical model of an electrode, shown in (Figure 41):

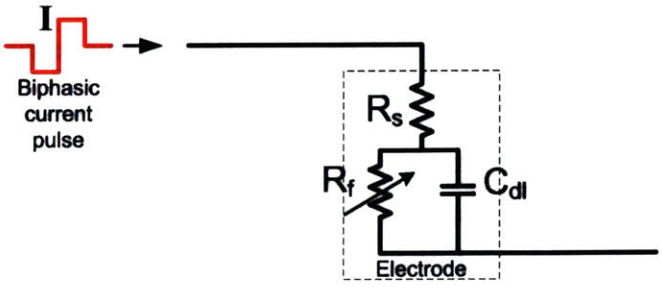


Figure 41. Simple circuit model for a typical electrode.

In this simple model, Rs is the solution spreading resistance, well determined by the

resistivity of the fluid (set by ionic species in solution). Cdl is the double-layer capacitance, created by the accumulation of tightly adsorbed ions near the electrode surface and more loosely attracted ions in a "diffuse layer" behind it. Finally Rf is the Faradaic resistance, which is governed by diffusion of reactive species to the electrode for charge-transfer reactions. It seems that Rf is not well modeled at present, being non-linear and time-varying over different time scales depending on the time constants of the reactions taking place. Hence it is indicated as a variable resistor in (Figure 41).

In the literature, to take reaction time constants into account, the Faradaic resistance Rf is also often modeled as a complex impedance called the Warburg impedance, which has a frequency dependence of $\omega^{-0.5}$. The physical origin of the capacitive behavior in the Warburg impedance can be derived from the Nernst diffusion layer thickness $\delta = (\pi Dt)^{0.5}$, and its dependence on the square root of time (Taylor and Gileadi 1995). As we shall see later, this time-varying behavior should be an important consideration in charge-balanced stimulator designs that employ the strategy of shorting out charge in the electrode.

The values of Rs and Cdl are more predictable, and in the devices we are designing for, should be between 5-20 kΩ and 5-15 nF respectively.

We should also make a distinction between Faradaic current flowing through Rf, and capacitive current flowing onto Cdl. Faradaic current is DC current flowing through the electrolyte, and signifies an electrolysis reaction taking place (which may be irreversible, like electroplating of tissue, and should be avoided). However, AC current which only charges and discharges Cdl would characterize non-Faradaic processes like ion adsorption, which are often quickly reversible and biologically less harmful.

5.2. Specifications

To assess the design, it is important to know the leakage current levels which are considered harmful to the body. Even a high quality blocking capacitor will have leakage, and may exhibit ~10GΩ of shunt resistance across its terminals. Thus with an average of 10V across the device, such a blocking capacitor would contribute ~1nA of leakage current.

Some cochlear implant companies have reported and tested patients with up to 25nA of leakage current in their device, without perceptible ill effects. However, above an

average DC level of 100nA, physiological damage is projected to occur (Xu, Shepherd et al. 1997). Our goal is therefore to get the average DC leakage current below 25nA.

The second important specification is the output voltage compliance. Due to high electrode impedances, clinical applications actually require between 9-20V of compliance (Bhatti and Wise 2006). For example, if we need to drive 1mA into an electrode with only 7 kΩ solution-spreading resistance, that alone requires at least ±7V rails and 14V of output compliance. The current levels we are required to operate over vary from 1μA to 1mA, depending on many user factors like electrode placement and etiology of deafness. In practice, the voltage compliance limits the maximum current amplitude we can drive. Since the percept evoked by stimulation varies with charge and not with current per se, we have to get around large electrode impedances by increasing the pulsewidth in order to deliver the same charge at lower current, albeit at the cost of timing resolution. Higher voltage compliance therefore enables higher timing resolution. To this end, we chose to design in an AMI 0.7μm, 30V process to achieve ~20V of output compliance.

5.3. Dynamic Current Balancing

The strategy that we have employed is a simple one, similar to a dynamic current mirror: by storing the level of current put out by an NMOS current sink on a PMOS current source, the two currents should be matched to within the accuracy of our storage cell. This eliminates the order of 1-2% matching error because the same PMOS device which stores the NMOS current is later used to output the P current pulse.

The hold time on the analog storage does not need to be very long since a single pulse of electrical stimulation in cochlear implants needs to happen on the order of ~100μs or less, so the precision of an analog storage element can be quite high over that time scale.

5.3.1. Overview of strategy to achieve charge balance

The strategy we will employ to achieve charge balance is two-fold: first, a dynamic current-balancing phase is performed to match a PMOS current source with an NMOS current sink. Second, after the biphasic pulse is generated, any residual charge error left on the electrode is then short to ground, with the details of shorting as an aid to charge balance covered in section 5.5. This order of operations is shown in (Figure 42) below:

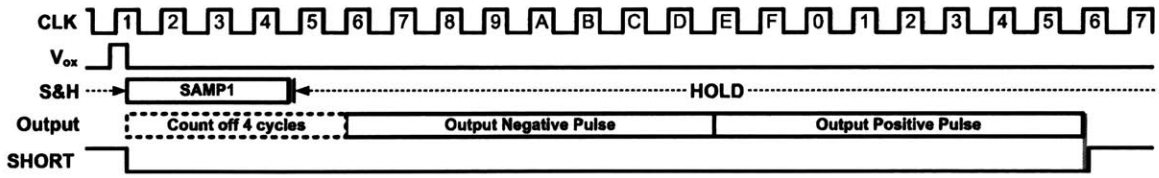


Figure 42. Timing diagram of dynamic current balancing, followed by biphasic pulse generation and shorting.

Once the asynchronous input pulse Vox arrives, the sample phase of a sample and hold (S&H) starts, which is shown as SAMP1 in (Figure 42). In this phase, the PMOS gate voltage required to support an NMOS current Iin is sampled on a hold capacitor C, as shown in (Figure 43) below:

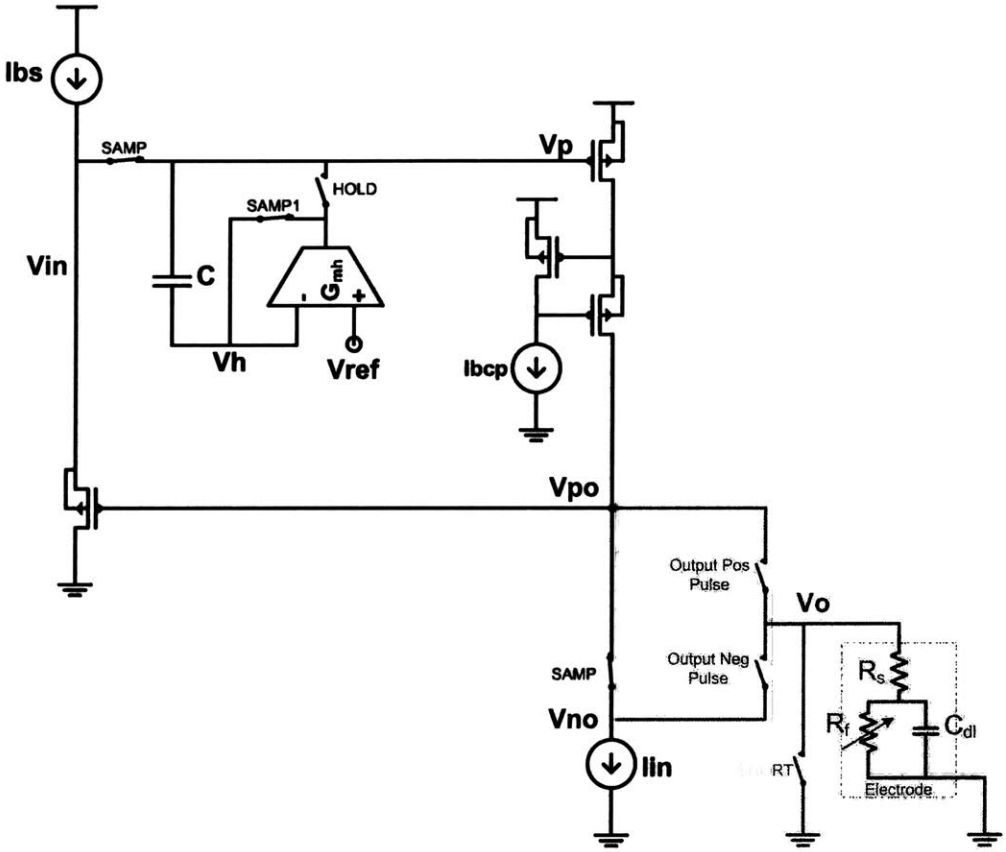


Figure 43. Sample phase in dynamic current balancing.

After the circuit has settled, the SAMP switches are opened, and the HOLD switch is closed for the remainder of the time, as shown in (Figure 44):

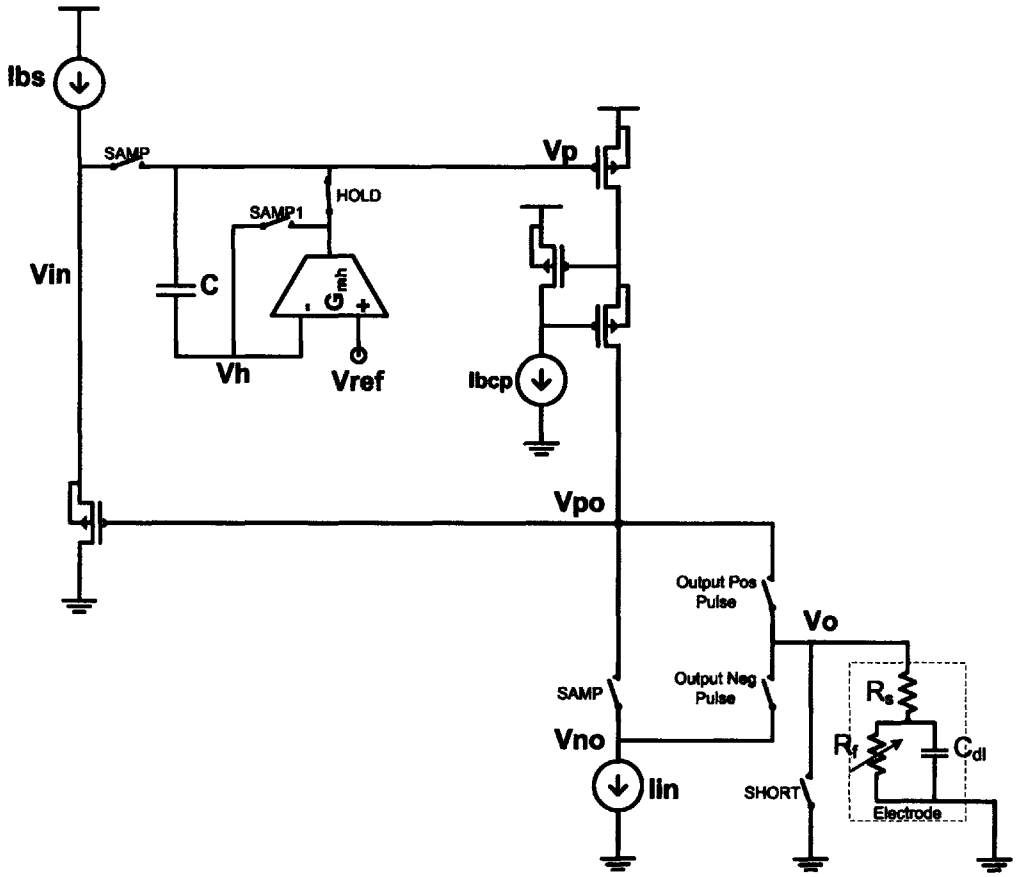


Figure 44. Hold phase in dynamic current balancing.

In this way, the gate voltage Vp is held through the duration of the positive pulse, and therefore the PMOS source puts out the same current which was sensed during the sample phase. Soon after the hold phase starts, the Output Neg Pulse, Output Pos Pulse and SHORT switches are each closed and opened in turn to complete the remaining operations. Details of the circuit are now described in the following sections.

5.3.2. Closed-loop sample and hold with low-leakage storage

Taking advantage of work that has already been done to achieve high-precision analog storage, the final circuit which we implemented is simply the integration of a low-leakage analog storage cell (O'Halloran and Sarpeshkar 2004) inside a dynamic current mirror feedback loop. The analog storage cell is the same one which achieves 12-bit precision on 10nW of power, by using low-leakage analog switches in a differential closed-loop sample and hold. To understand its operation, a conventional closed-loop sample and hold circuit is shown in (Figure 45) below:

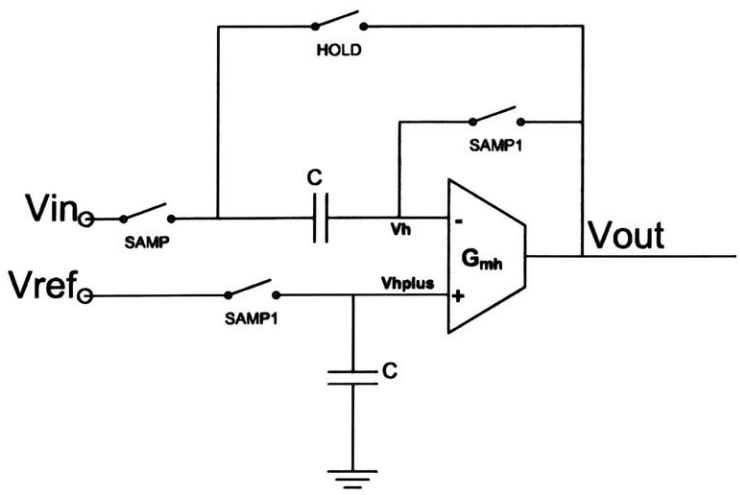


Figure 45. A conventional closed-loop sample and hold circuit.

On the sample phase, SAMP1 and SAMP are closed, and HOLD is open. Vin - Vref is sampled across the capacitor connected to Vh, while Vref is sampled on Vhplus. On the hold phase, SAMP1 opens first, to implement bottom-plate sampling (at a fixed common mode of Vref) before SAMP opens and HOLD closes. Vout then holds Vin in closed loop feedback, which attenuates noise and other disturbances to Vout by the gain of the loop. Gmh is a transconductance amplifier with cascoded outputs for high gain.

The improvement to this topology using a low-leakage switch is shown in (Figure 46) below.

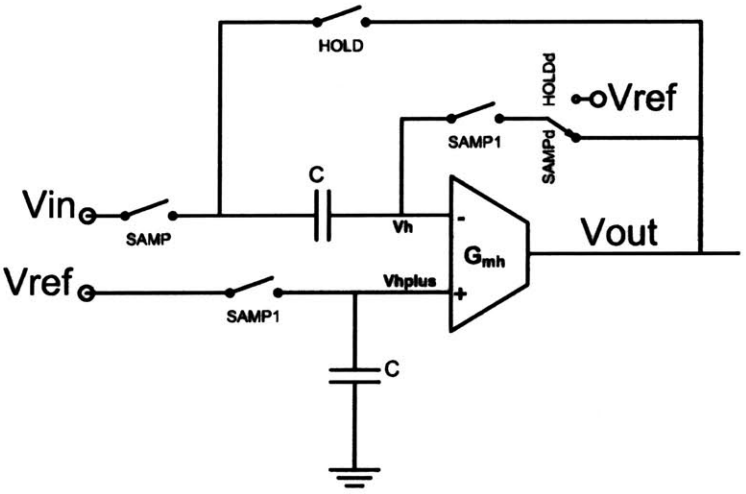


Figure 46. The closed-loop sample and hold, including a low leakage switch.

It was noted in (O'Halloran and Sarpeshkar 2004) that minimizing the switch leakage

on node Vh is crucial to extending the hold time of the closed loop sample and hold. The key to minimizing leakage across the switch is to match both source and drain voltages so even sub-threshold diffusion currents would go down to aA current levels. While this is done automatically for the Vhplus switch, an additional switch, called the "low-leakage switch" is added to drive the opposite side of the Vh switch to Vref during the HOLD phase. This low-leakage switch is controlled by the switching signals SAMPd and HOLDd.

One change I made to the circuit operation of (O'Halloran and Sarpeshkar 2004) is to delay the switching of SAMPd and HOLDd at the ***start*** of the SAMP phase, as indicated by the "d" suffix, in order to improve the settling time during the SAMP phase. This will be elaborated on later. This delay is not to be confused with the delay between SAMP1 and SAMP, where SAMP (and also SAMPd) are delayed from switching at the ***end*** of the SAMP phase, to minimize charge injection errors on the sampled voltage.

5.3.3. Analog circuit implementation

The complete circuit which implements the sample and hold within a dynamic mirror is shown in (Figure 47) below:

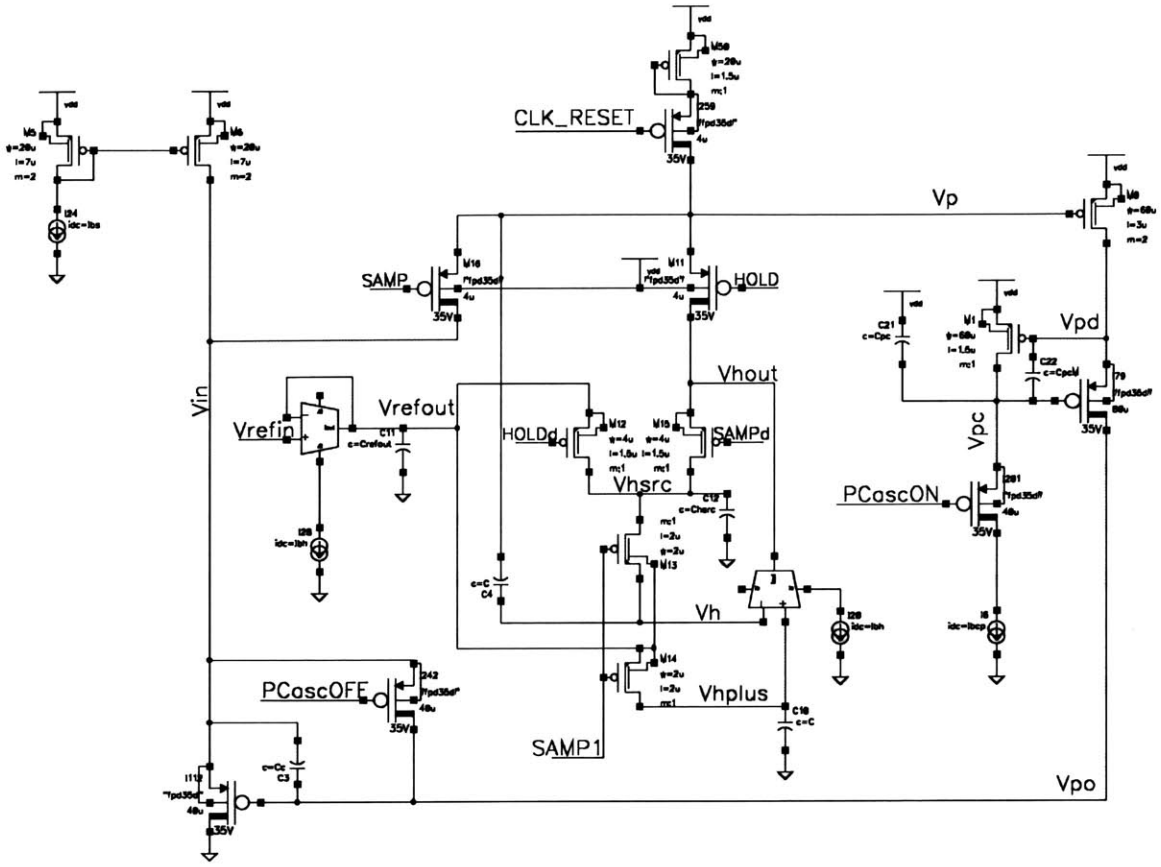


Figure 47. Dynamic mirror including an analog storage cell

The circuit works as follows: during the SAMPLE phase, current is pulled out of node Vpo when PCascON is low and PCascOFF is high. Vin is thus the output of a source-follower that tracks Vpo, Vp is connected to Vin via the switch SAMP, and Vp completes a feedback loop back to Vpo by servoing the current source transistor to match the current pulled out of Vpo. The transistors M1 and I79 form an active cascode, turned on and off by PCascON. The gate voltage Vp is stored across capacitor C4 in this phase.

The analog storage cell within this loop should be recognizable, as Vrefout provides a buffered version of Vrefin for noise rejection, Vhout is the output of the sample and hold, and Vh and Vhplus are the storage nodes that need to be quiet and symmetric in layout for differential cancellation. During the HOLD phase, the storage cell will hold the gate voltage Vp and therefore put out a current which matches the input current during the SAMPLE phase.

When the circuit is inactive and waiting for a signal to put out a stimulation pulse, the transistor gated by CLK_RESET is on and pulls Vp up to a diode drop below the rail.

This is to assist the settling time during the SAMPLE phase, by biasing Vp to a minimum-current level, since it is faster for the source follower to pull down on Vp than to pull up. During this inactive phase, PCascON is off (saving power by shutting off the cascode) while PCascOFF is on. This shorts out the source-follower, and pulls Vpo up to Vdd. This is also to minimize the settling time by starting at the same initial condition on Vpo every time, and once more it is faster for Vpo to pull down than to pull up.

Another minor addition to the implementation in (O'Halloran and Sarpeshkar 2004) was the addition of a 1pF capacitor, Chsrc, at the node Vhsrc. This was to dampen a voltage spike on this node when switching from SAMP to HOLD, which might inject an error charge into Vh.

The P transistors shown with solid 35V drain symbols are DMOS (diffusion-MOS) transistors with lightly-diffused drains, to handle large breakdown voltages. This circuit was designed to operate on rail voltages as high as ±10V, to achieve 20V of output compliance.

The transconductor Gmh which is used in the analog storage cell is shown in (Figure 48) below. It is used both for the closed-loop feedback amplifier and also as the buffer for Vref, and biased with the same bias current, since that matches the impedance seen at Vref with the impedance seen at Vh when the switch SAMP1 is turning off. The matching of impedances should help the differential charge injection onto the capacitors to also be well matched, to minimize the hold error.

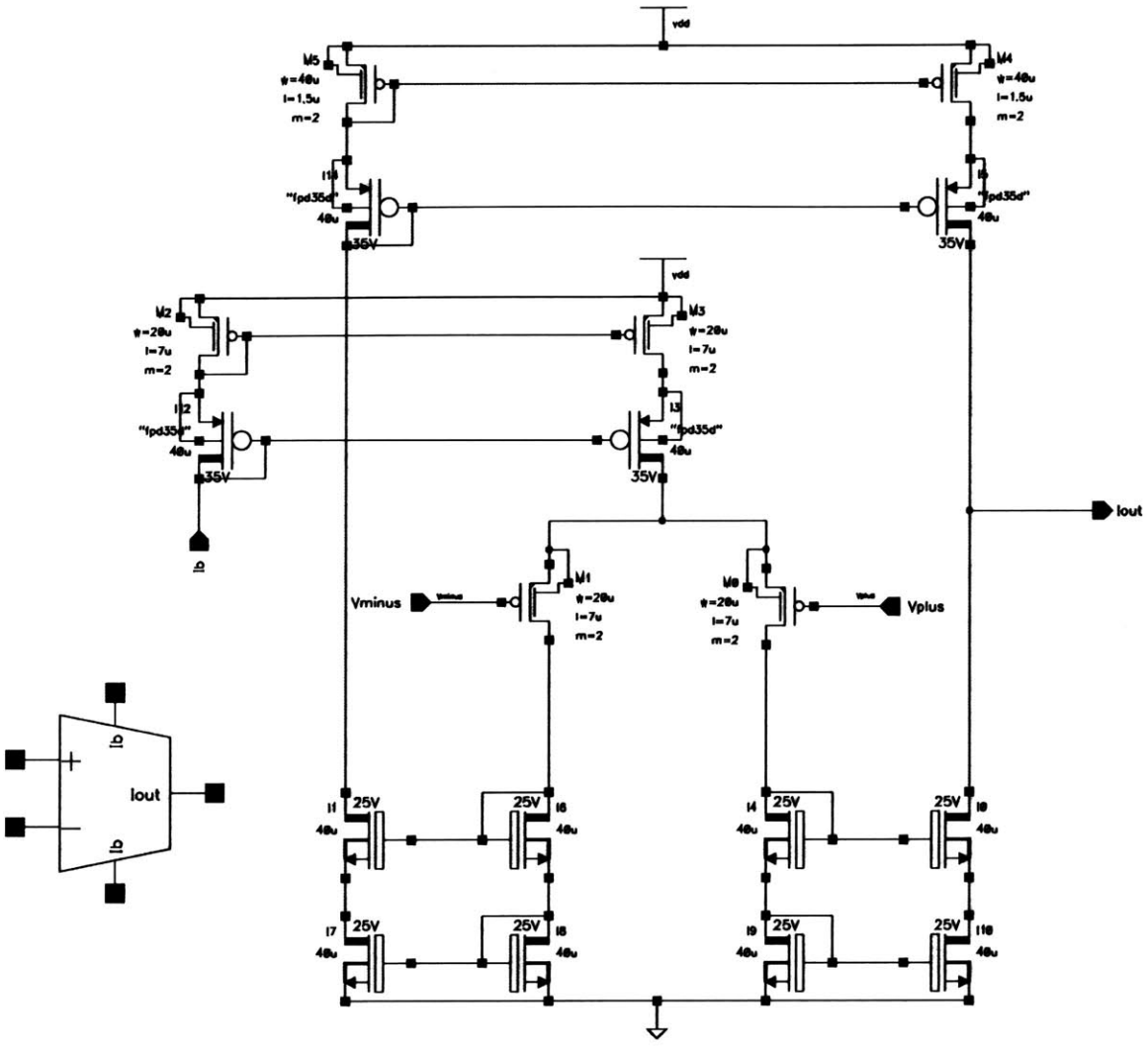


Figure 48. Topology of high-output impedance transconductor Gmh, used in the analog storage cell. The output is cascoded and protected with DMOS output devices to allow a wide voltage swing from Vref to Vin without breaking down.

The full analog section which implements the biphasic pulse generation was fabricated in a 0.7μm, high-voltage AMI process and is shown in (Figure 49) below:

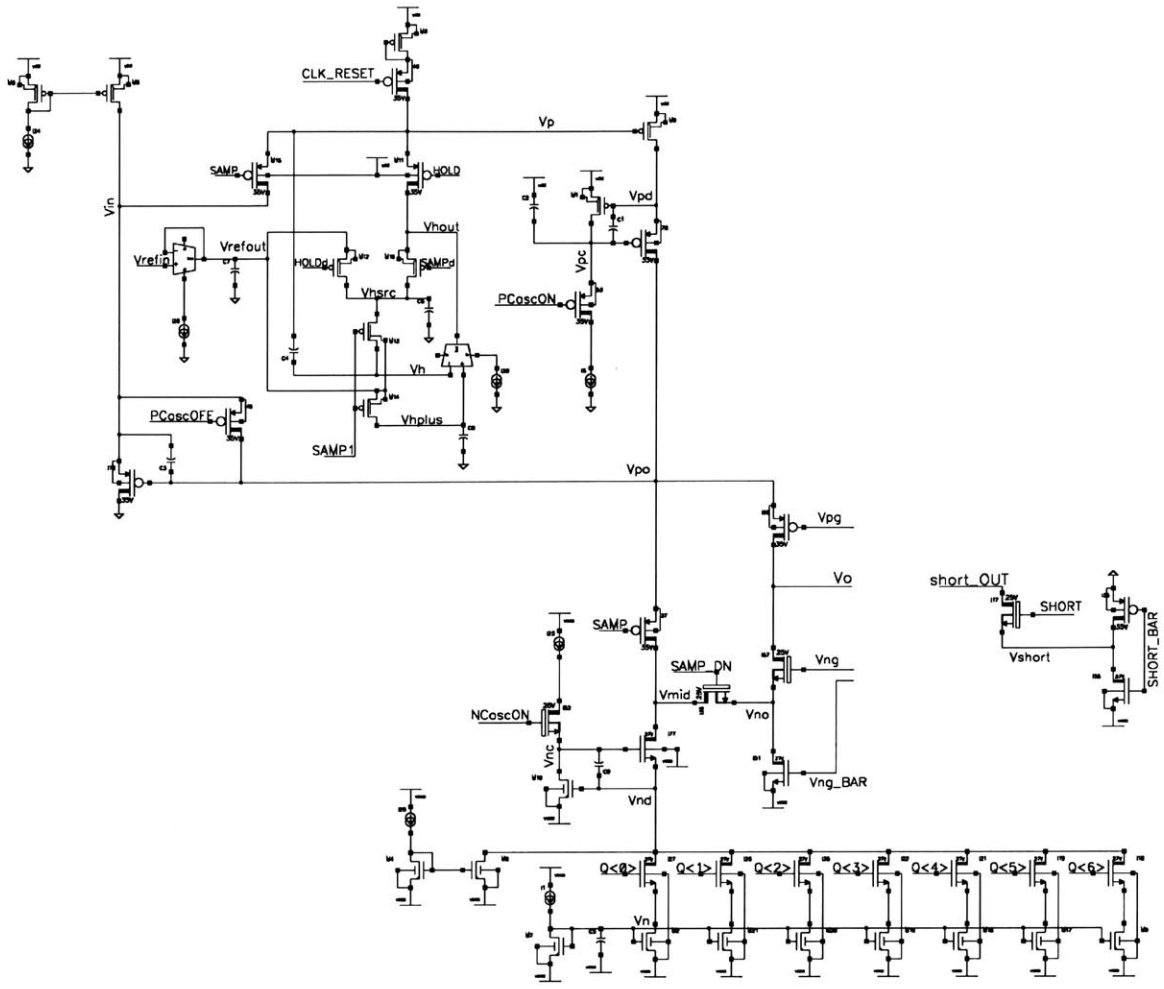


Figure 49. Analog section of biphasic pulse generator. The bottom-right half of the circuit is a 7-bit NMOS current DAC, with switches to convey current either to the storage cell or to the output Vo. The top-left shows the PMOS dynamic mirror and storage cell loop. Both NMOS and PMOS current sources utilize active cascodes to achieve high output impedance.

5.3.4. NMOS current DAC and control switches

A close-up of the bottom half of the circuit is shown in (Figure 50) below:

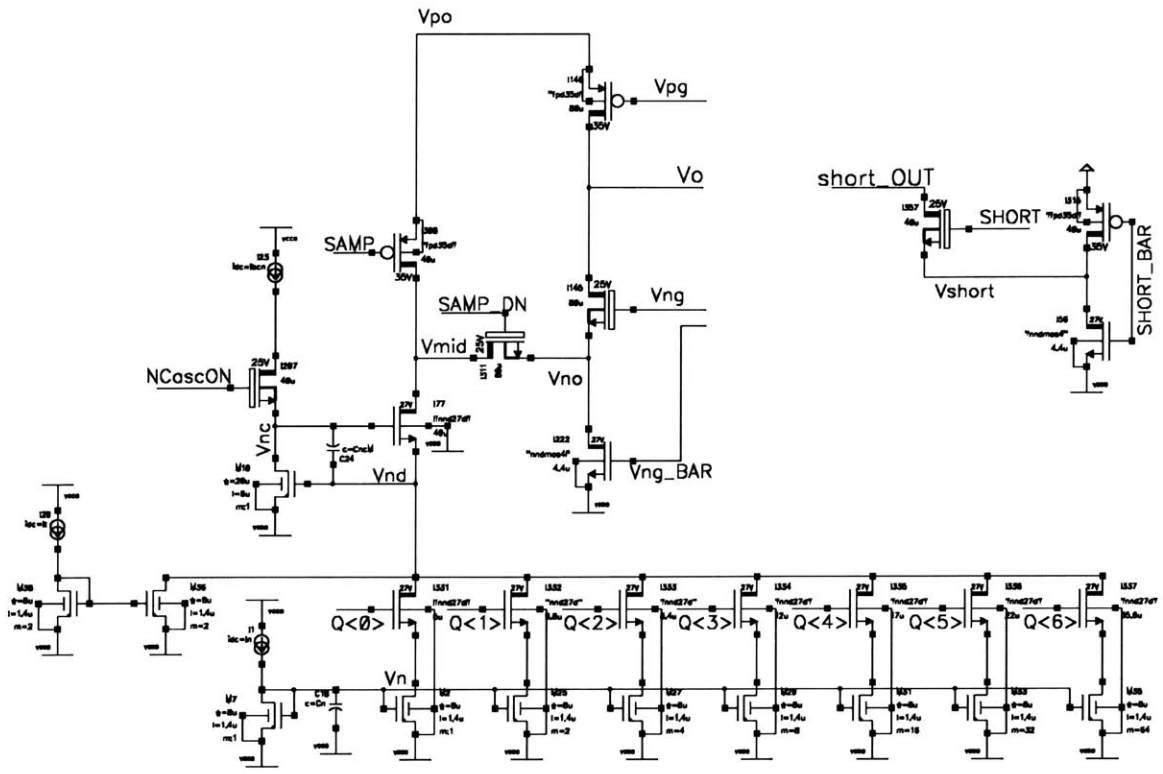


Figure 50. 7-bit current DAC and switches to steer current up in the SAMPLE phase or off chip during stimulation. A shorting circuit short the output to ground is also shown on the far right.

This portion of the circuit functions as follows: The current In which sets up gate voltage Vn goes to a common-centroid 7-bit current DAC, switched by control signals Q<6..0>. There is an additional current It which is added to the DAC output to set the "T-level" or minimum detectable threshold current in that electrode channel. NCascON controls an active cascode, and can be switched off to conserve power when not in use.

During the SAMPLE phase, SAMP is on, while Vpg, Vng and SAMP_DN are off to disconnect the output. Current is therefore pulled out of Vpo from the dynamic mirror up above. During this time, Vng_BAR is on to hold Vno at the negative rail Vssa. Note that Vssa might be -10V, while Vmid could rise to be near Vdd, which may be almost a 20V drop. Without the SAMP_DN transistor, reverse current may flow out of the Vng gating transistor, and so Vno must be held at Vssa. The SAMP_DN transistor therefore protects the Vng gating transistor from the voltage surge on Vmid. The NDMOS transistors with lightly-doped drains to handle high voltages are labeled with solid black drain symbols and come in 2 varieties: 25V (floating P-body NDMOS) or 27V (non-floating P-substrate NDMOS).

During stimulation, which starts with a negative current pulse, SAMP and Vng_BAR are turned off, while Vng and SAMP_DN are turned on (Vpg stays off during this time). The DAC current is then pulled out of Vo through Vng and SAMP_DN for a fixed time which is set by a high-precision clock. After the negative pulse, Vng and SAMP_DN turn off, while Vng_BAR and Vpg turn on. Concurrently, in the top half of the circuit PCascON also turns on, and the value of current which was stored on Vp during the SAMPLE phase gets sourced through the Vpg transistor for another fixed time period equal to the negative pulse time.

Finally, after the positive pulse has ended, SHORT goes high while SHORT_BAR goes low, which shorts its output, short_OUT to ground. This stays on until the next stimulation pulse arrives. Shorting is a technique used to discharge any residual charge error that is left on the electrode capacitance after the stimulation pulse. By providing a path for residual charge to flow out of the electrode capacitance, and not through the Faradaic resistance into the electrolyte, this method corrects for some of the original imbalance in the charge pulses. Its effectiveness will be analyzed in a later section. In practice, short_OUT is connected to the electrode output Vo, but was separated in this circuit to investigate the difference with and without shorting.

A timing diagram which summarizes in sequence all the above timing relationships is shown in (Figure 51) below:

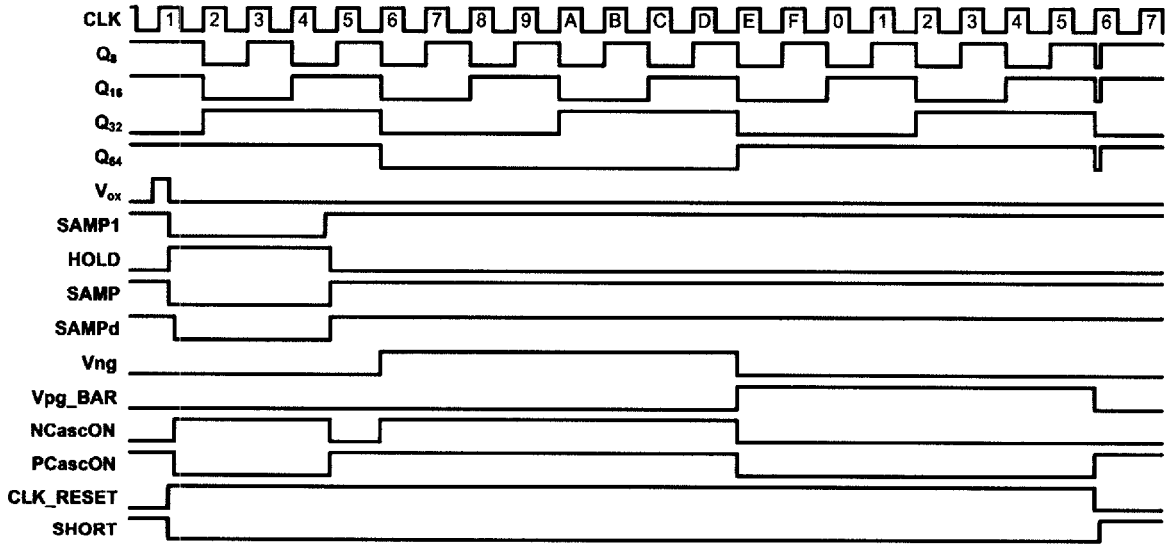


Figure 51. Timing diagram for operation of the dynamic current balancer

5.3.5. Logic Circuits

This dynamic current balancer was designed to be fully compatible with receiving the AIS ABEP asynchronous spiking outputs, and thus the initiation of a pulse occurs on the falling edge of an input spike (Vox in the timing diagram above). This latches a 7-bit DAC value into registers, and the 7-bits are then level-shifted from (0,3V) logic levels down to (Vssa,0V) as shown in (Figure 52) below:

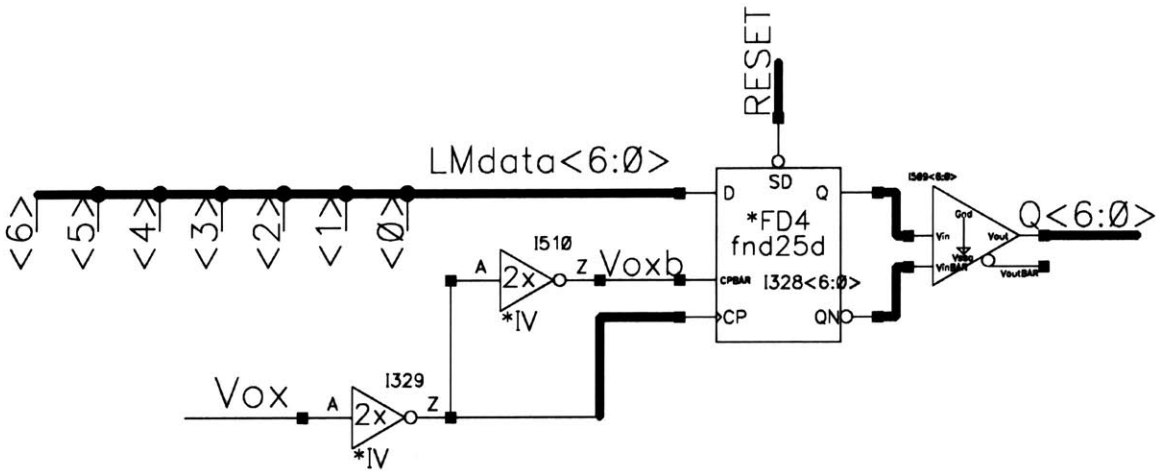


Figure 52. Input latches triggered by asynchronous spike on Vox.

The necessary level shifters to go from (0,3V) logic levels up to (0,Vdd) and down to (Vssa,0) for controlling analog switches are shown in (Figure 53):

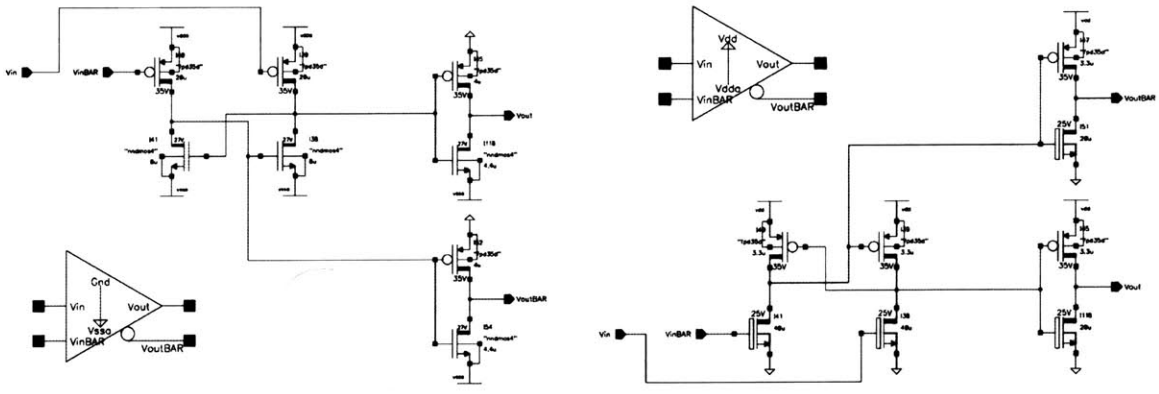


Figure 53. Level shifters to create correct voltage levels to turn analog switches on and off.

Vox also raises the signal CLK_RESET (active low), which allows a cascade of 4 toggle flip-flops to count off 4 cycles for sampling (clock cycles 2 to 5 in the timing

diagram), followed by 8 samples for the negative pulse (clock cycles 6 to D) and another 8 samples for the positive pulse (clock cycles E to 5). Nominally, the CLK frequency was designed to be 256 kHz, to allow 16μs for sampling and 32μs for each pulse. This allows the CLK to be shared with the AIS ABEP, for its dual-slope converters to sample 7 bits at 2 kHz. At the end of stimulation, CLK_RESET is lowered to deactivate the clock dividers and save power. This clock divider is shown in (Figure 54) below:

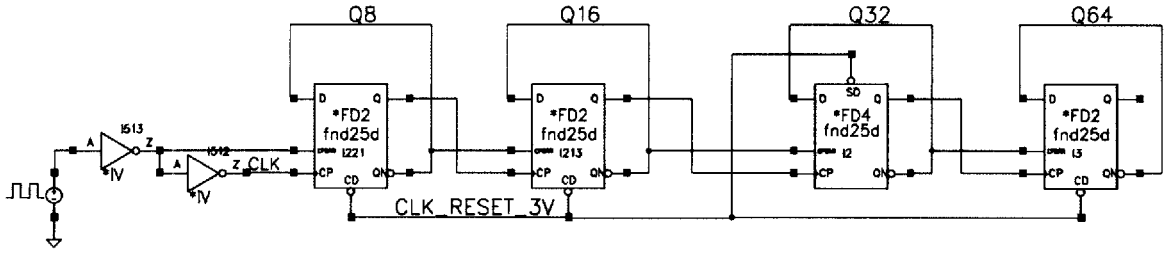


Figure 54. Clock divider, activated when CLK_RESET goes high.

When Vox raises CLK_RESET, it also lowers SAMP1, and SAMP1 in turn starts a single-slope analog timer to count off a time for the SAMPLE phase, shown in (Figure 55) below:

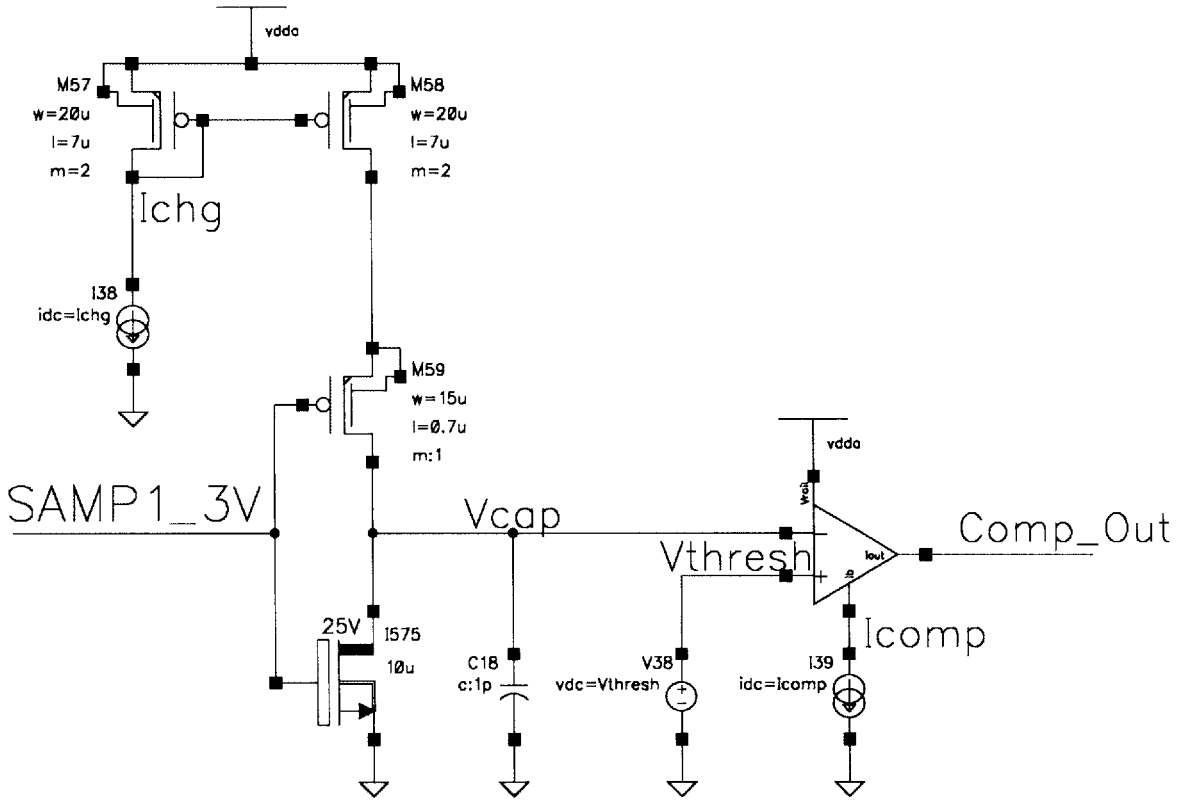


Figure 55. Single-slope analog timer (current starved inverter charging a capacitor to a voltage threshold).

This analog timer does not need to be very precise, but it must signal the SAMPLE phase to end some time before the start of the negative pulse. This is because the N current sink must have some time to settle into an "off" state after the connection to the dynamic mirror is broken. Both N and P current sources must be configured to turn on from the same state of being "off" initially, and then coming "on" into saturation when the gating transistors are opened, in order to achieve good charge balance. So the charge balance will be adversely affected if the N current sink has not settled sufficiently into an "off" state that matches the P current source.

By using this asynchronous analog timer, we are able to set a sufficient amount of time for feedback loops to settle in the SAMPLE phase, but also to end the SAMPLE phase sufficiently before the start of stimulation (in <4 clock cycles or ~16μs). Note in the timing diagram that the SAMPLE phase ends even before clock cycle 5, because the Vox pulse is shown to arrive just after a rising edge on the clock, which gives almost one extra clock cycle before the start of stimulation. We have to design for the worst case, however, where Vox may arrive just before a rising edge on the clock, and hence the

SAMPLE phase must still end within less than 4 clock cycles from the arrival of Vox.

5.3.6. Delay Circuits

Finally, I will describe the delay circuits and their importance in this section. There are 3 short analog delays in this circuit, which should be noticeable in the timing diagram:

1. SAMPd and HOLDd are delayed from the start of SAMP1. In addition, because NCascON = Vng OR HOLDd, and PCascON = Vpg_BAR NOR HOLDd, both NCascON and PCascON are also delayed from the start of SAMP1. It was found that these delays improved the settling time of the SAMP phase by many μs, which is significant, due to the following reason: When the two halves of the circuit are first connected, Vmid surges up to be short with Vpo (near Vdd) and couples a large voltage spike onto Vn. There are 4 feedback loops that will all need to reject this disturbance: the 2 active cascodes, the loop within the analog memory, and the dynamic mirror major loop itself. Since they are all interconnected, they can be thrown into a long period of ringing. The delay makes sure that the 4 feedback loops remain broken during the voltage spike, and turn on only after the DAC current has settled. Adding a bypass capacitor of Cn = 5pF to Vn also helped dampen the voltage spike and improved the initial transient behavior.
2. SAMP / HOLD and SAMPd / HOLDd are delayed from the end of SAMP1 so that charge injection onto the hold capacitor voltages is minimized.
3. SHORT rising high is delayed from the end of CLK_RESET falling low. This is because we want to turn the SHORT phase on only after the positive pulse has fully ended. If the SHORT phase and overlaps with the positive pulse, the charge balance will be adversely affected.

These three delays are implemented using analog delay blocks as in (Figure 56) below:

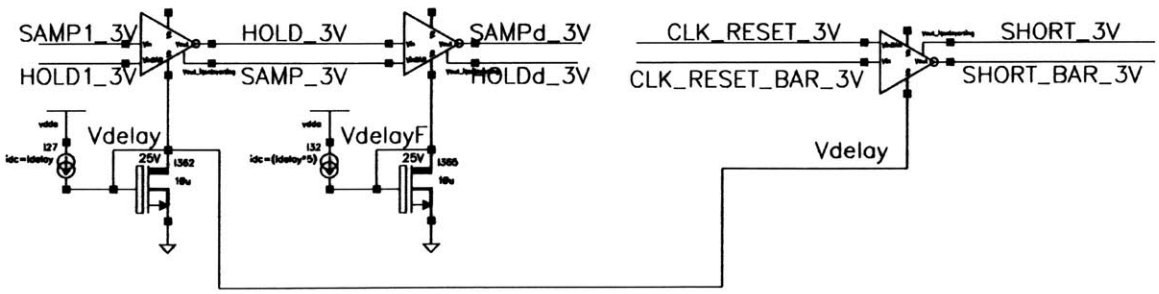


Figure 56. Implementation of three short delays using an analog delay circuit. The delay SAMPd / HOLDd can be much shorter than the other two, and so it has a separate bias.

The analog delay circuit is the same one which was implemented in (O'Halloran and Sarpeshkar 2004), and shown in (Figure 57) below:

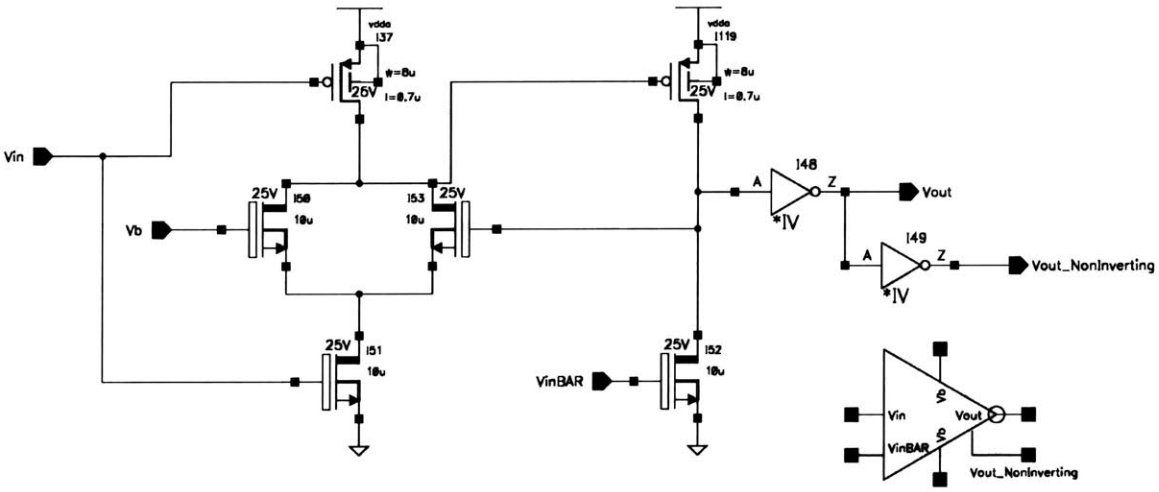


Figure 57. Analog delay of a rising edge on Vin. Falling edges on Vin are not delayed.

The circuit delays only a rising edge on Vin, by having to wait for the bias current from Vb to discharge the gate capacitance of I119. There is a positive feedback loop which quickly regenerates a sharp rising edge on the drain once the gate of I119 nears its threshold. By choosing the right polarity of Vin and VinBAR, we can choose the analog delay to be on any edge that we want.

5.4. Feedback and Noise analysis

It might already be evident from the circuit operation that there is a tight constraint on meeting the settling time during the sampling phase, because 3 circuit techniques had to be employed to minimize aberrant behavior in the initial transient response that would

otherwise hurt the settling time. Those techniques are summarized again below:

1. SAMPd, HOLDd, PCascON and NCascON are delayed at the start of SAMP1
2. PCascOFF switch initializes Vpo to Vdd before SAMP begins
3. CLK_RESET switch initializes Vp to a diode drop below Vdd before SAMP begins

I chose to allow the sampling phase to take up to a quarter of the stimulation time, i.e. 16μs prior to the 64μs stimulation time, trading off a lower power consumption (not settling very fast) at the expense of some latency.

Even with well-behaved large-signal transient behavior, the small-signal settling time for the feedback loop must be considered, to make sure that the circuit can settle over many time constants to achieve good precision.

In addition, we will analyze the noise performance of this circuit to see how much the sampled thermal noise affects the charge balance.

5.4.1. Closed loop stability and bandwidth

A simplified schematic of the dynamic mirror in closed-loop sampling is shown in (Figure 58) below:

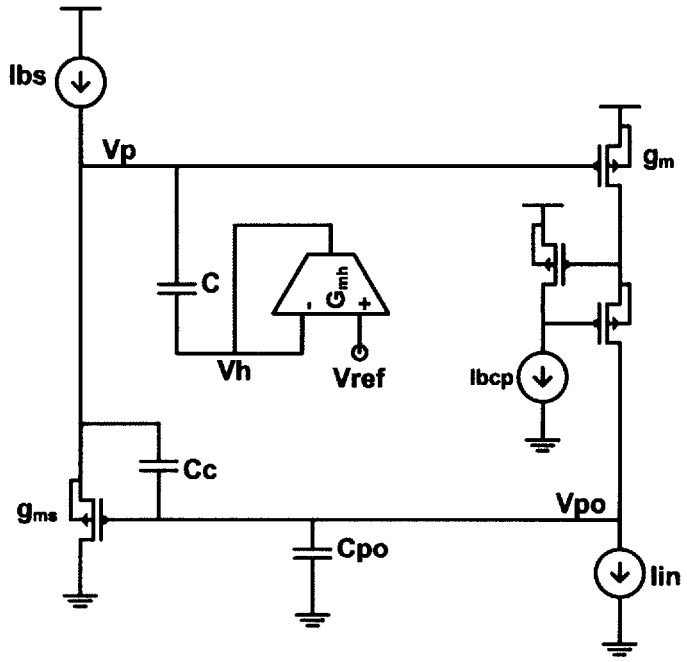


Figure 58. Simplified schematic of dynamic mirror in the SAMPLE phase.

Cpo is not shown in earlier schematics because it is a parasitic capacitance, but is important for this analysis. The block diagram for this loop can be written out as in (Figure 59) below:

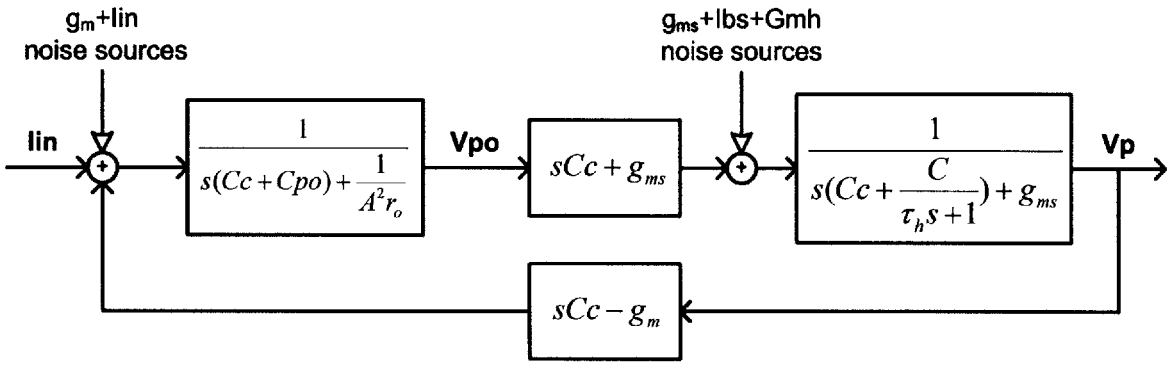


Figure 59. Block diagram of sampling loop.

A simplifying assumption is made here, that the active cascode is fast and does not contribute to the loop dynamics, and only increases the output resistance by a factor of A^2 to A^2r_o. τ_h is defined as C/Gmh, and all other terms should be readily identifiable.

To find the loop stability and closed loop response, we can simplify the block diagram to find Vp/Iin, as shown in (Figure 60) below:

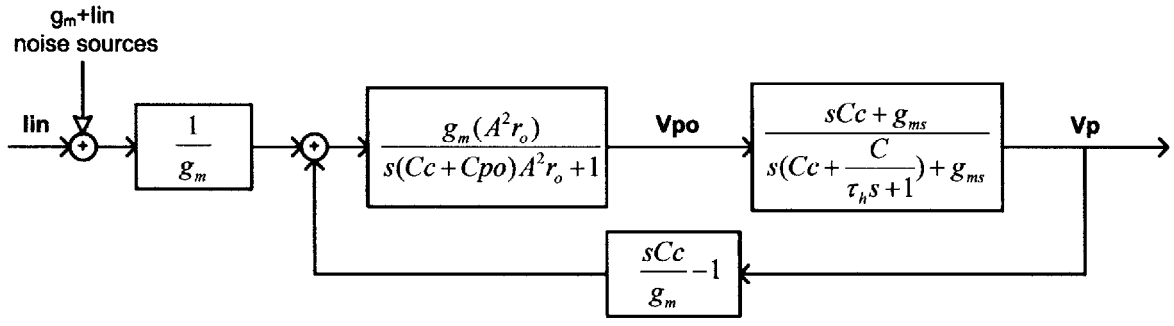


Figure 60. Simplified block diagram, to read off L(s) and loop bandwidth.

Observing that the transfer function Vp/Vpo can be treated as ~unity (which will be verified more formally later), it becomes easy to plot the forward transmission G, the inverse of the feedback transmission 1/H, and also L(s). Since L(s) = G·H, it is indicated by the continuation of G with the dotted line in (Figure 61) below:

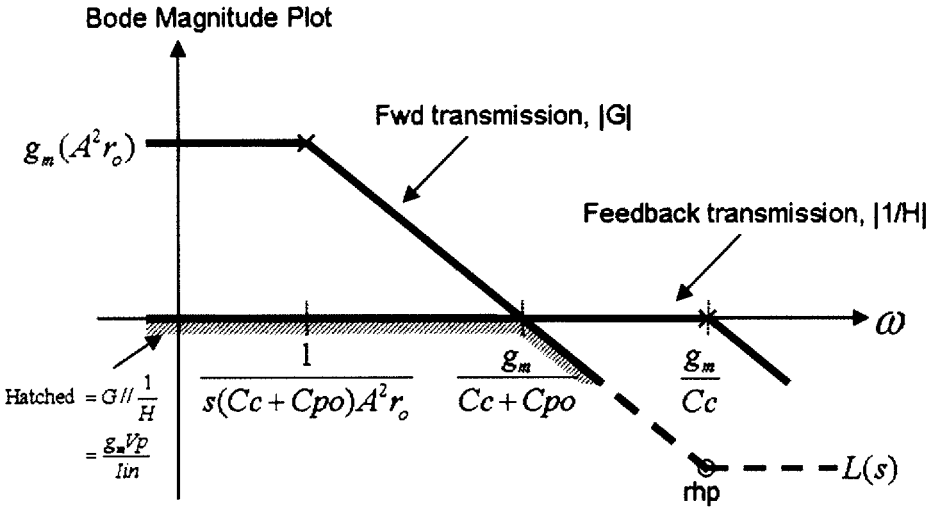


Figure 61. Plot of closed loop transfer function G//(1/H) and loop transmission L(s).

Thus we can see that the loop is stable, with a unity gain crossover at $g_m/(Cc+Cpo)$, and a destabilizing right-half-plane (rhp) zero at g_m/Cc. However, since the rhp zero comes after crossover by a factor of Cpo/Cc, it should not adversely affect the loop stability. However, this is one reason to not make Cc too large, and was set to be 0.5pF.

We can identify a few key insights from this loop: Firstly, without compensation, the non-dominant pole would come primarily from the source-follower pole at g_{ms}/C in Vp/Vpo, which can be quite slow. The compensation capacitor helps a lot by adding a zero at g_{ms}/Cc to cancel out the non-dominant pole. In addition, the large sampling capacitor C, which is sized at 2pF, would be a bigger problem for stability if not for the zero introduced by Gmh in series with it: at high frequencies, rather than seeing the capacitive rolloff continue forever, the impedance flattens out at 1/Gmh and hence inserts a zero at Gmh/C. The combination of both these zeroes makes the loop stable over a large range of input currents, down to the minimum Iin of 1μA.

We can also see that the closed-loop bandwidth is the same as the unity-gain bandwidth of $g_m/(Cc+Cpo)$, and since the minimum input current Iin is set by our application at 1μA, we can expect one time constant to be at most ~1pF · 200mV / 1μA = 200ns. There may be some degradation from this value due to a lag effect (drop in gain) from Vp/Vpo, but the loop should still be fast enough to settle to within 10 bits (7 time constants) even with an order of magnitude reduction in speed (to 2 μs).

5.4.2. Effect of non-dominant poles from source follower

To formally assess the effect of the non-dominant poles coming from Vp/Vpo, we can write out the transfer function as follows:

$$\frac{Vp}{Vpo} = \frac{sCc + g_{ms}}{sCc + \frac{sC}{\tau_h s + 1} + g_{ms}}$$

$$= \frac{(sCc + g_{ms})(\tau_h s + 1)}{(sCc + g_{ms})(\tau_h s + 1) + sC}$$

$$= \frac{(\tau_s s + 1)(\tau_h s + 1)}{(\tau_s s + 1)(\tau_h s + 1) + \frac{sC}{g_{ms}}}$$

The first thing to note is that at high frequencies, Vp/Vpo recovers to unity gain. However, this may not be entirely true because we neglected parasitic capacitance on Vp, which we can call Cp. Thus at high frequencies there may be a loss of gain (a lag effect) which reduces the bandwidth by a factor of Cc/(Cc+Cp).

Secondly, the two zero locations are well defined at Gmh/C and g_{ms}/Cc as mentioned in the previous section. The poles are split to go higher and lower than the zero frequencies by the term sC/g_{ms} which increases the effective "damping" of the second order system in the denominator (the poles are real, so they stay on the real axis and just and split apart). The bode plot for this is shown in (Figure 62) below:

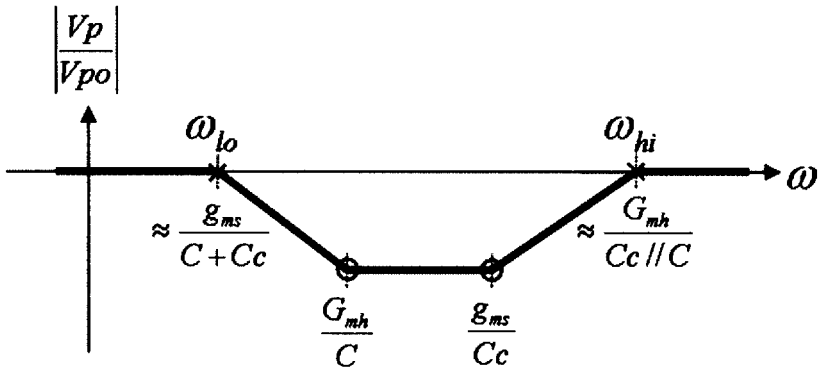


Figure 62. Bode plot of Vp/Vpo, showing two stabilizing zeroes and a transfer function of unity above the high frequency pole.

The approximate locations of the split poles, ω_{lo} and ω_{hi} are shown. Since in practice we will bias g_{ms} and G_{mh} with the same bias current (~1.2μA), this sets $g_{ms} \approx 2G_{mh}$, and the ordering of poles and zeroes will be as shown, dominated mainly by the capacitors.

Since g_m will be set by a minimum bias current of 1μA, we can expect that Vp/Vpo will return to ~unity gain before the crossover of $g_m/(Cc+Cpo)$, in all but the lowest current settings. Even in the lowest current settings, however, there should be some loss of bandwidth but not loss of stability.

5.4.3. Noise estimate

Because the voltage Vp which is stored needs to be converted back to a current through the transconductance $Ip = g_m Vp$, we can expect that the error will get worse with higher g_m. Fortunately, g_m should eventually become constant at high currents when the device enters velocity saturation and the current increases only linearly with gate voltage. For the size of our current source transistor, we can expect to enter velocity saturation above 500μA.

From the block diagram in (Figure 60), we can see that the noise from the g_m and I_{in} current sources will see the full bandwidth of the loop. To estimate the contribution of the other devices, we can re-draw the block diagram of (Figure 59) in (Figure 63) below:

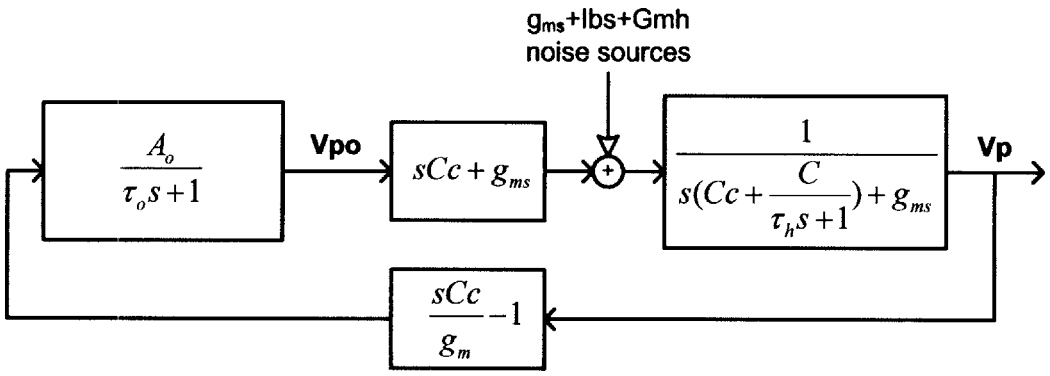


Figure 63. Block diagram showing noise contribution from Ibs and similar current sources.

A_0 is used to represent the loop gain of $g_m(A^2r_o)$ and τ_o is the corresponding dominant pole at $(Cc+Cpo)A^2r_o$. Since the noise sources shown above see the loop gain in the feedback path, we can expect their contribution to the output to be greatly attenuated. To show this formally, we can rewrite the block diagram as in (Figure 64) below:

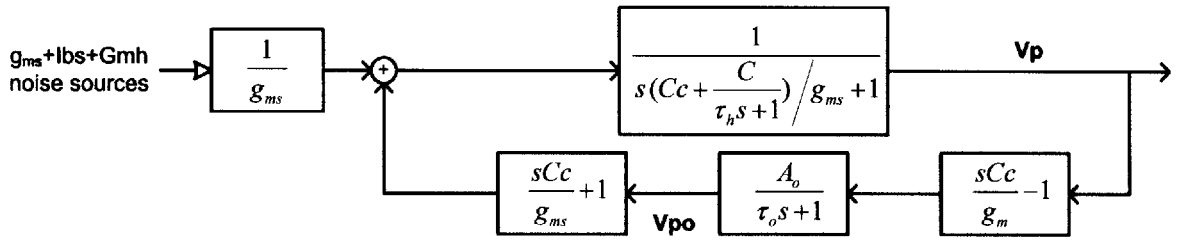


Figure 64. Block diagram showing attenuation of Ibs and similar noise sources.

Finally, we can plot the transfer function from Ibs and similar noise sources to Vp, shown in (Figure 65) below:

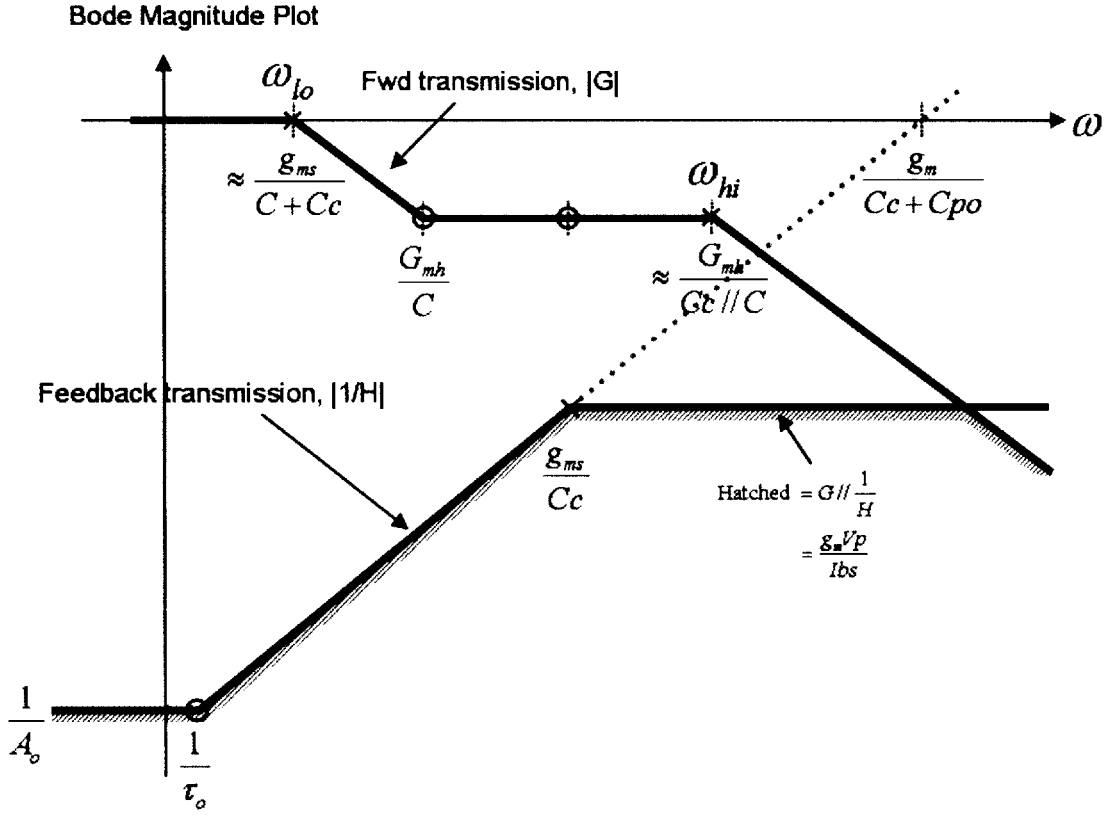


Figure 65. Bode plot of transfer function from Ibs and similar noise sources to Vp.

G can be observed to be Vp/Vpo without the zero at g_{ms}/Cc, and 1/H shows the strong attenuation of 1/Ao at low frequencies. At higher frequencies, it will intersect with the rolloff from G, and hence the noise injected at Vp should be attenuated by the loop over all frequencies. Although the current noise from Gmh does not inject directly into Vp, it is capacitively coupled through C, and thus should have a high-pass nature which further

attenuates low-frequency components. Hence we can expect the noise contribution from these sources to be negligible compared to g_{ms} and Iin injected at Vpo.

Finally, we can note that the low-frequency impedance looking into Vp is $1/g_m$ divided by the loop gain A_o. This proves the result that was mentioned in section 4.3.1.

Now we have all we need to estimate the noise:

$$
\begin{aligned}
Vp_{,PSD} &= \frac{N \cdot 4kT\left(\frac{2}{3}g_m\right)}{g_m^{\,2}} \\
\sigma_{Vp}^{\,2} &= \int_0^{\infty} Vp_{,PSD} \times H^2(f) \cdot df \\
&= Vp_{PSD} \times BW \times \frac{\pi}{2} \\
&= \frac{N \cdot 4kT\left(\frac{2}{3}g_m\right)}{g_m^{\,2}} \times \frac{g_m}{2\pi(Cc+Cpo)} \times \frac{\pi}{2} \\
&= \frac{\frac{2N}{3}kT}{Cc+Cpo} \qquad \text{where } N \simeq 2 \text{ and } Cc+Cpo \simeq 1pF \\
\sigma_{Vp} &\simeq \sqrt{\frac{4/3 \times 1.6e-19 \times 26mV}{1pF}} \\
&= 75\ \mu Vrms
\end{aligned}
$$

So even with a spread of 6σ, and using a maximum g_m of ~1mA/V, this works out to a total spread in the current noise of ~ 450μV × 1mA/V ≈ 450nA, which is better than 11 bits of precision on a full scale signal of 1mA. In terms of charge, using a pulsewidth of 32μs, the standard deviation is ~ 75μV × 1mA/V × 32μs ≈ 2.4pC. This is therefore a very low noise circuit, and is borne out in measurement as well: the noise observed (shown in section 5.6.5) is only a small modulation on top of the final residual charge error.

5.4.4. Final error estimate

Assuming that the final error would just be dominated by errors in the stored analog voltage, the maximum current error we would be able to achieve is then the voltage error multiplied by the maximum g_m.

At the highest current level, g_m is estimated to be around 1mA/V. The full scale variation in gate voltage Vp is around 1.4V. If we achieve a maximum error of ~1.4mV or ~10 bits of precision on the analog storage, we would then achieve ~1.4 μA or just over 9 bits in current terms. If the maximum g_m was higher, we would have higher current error and even lower precision. Hence we can see that g_m reaching a maximum

due to velocity saturation actually helps us by limiting the current error.

Simulations predicted that the circuit would be able to achieve a charge balance of ~0.15% error, which is 1.5μA out of 1mA. As will be shown in the data of section 5.6.3, we achieve in practice a 0.4% matching error, due to other error sources besides the analog storage alone.

5.5. RC analysis on shorting (when is shorting beneficial?)

It is widely believed that shorting an electrode to ground between stimulation pulses is sufficient to ensure safe (<100 nA) long-term electrical stimulation (Bhatti and Wise 2006).

In this section we will analyze the usefulness of shorting and its limitations.

5.5.1. Effectiveness of shorting

The basic operation that occurs during shorting is shown in (Figure 66) below:

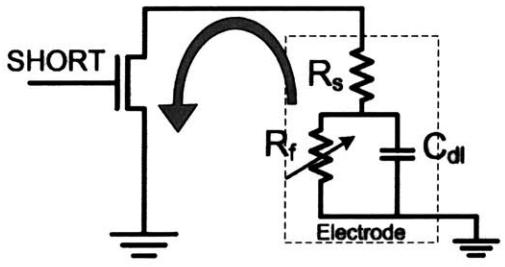


Figure 66. The basic operation in shorting: discharge of Cdl.

When SHORT is raised high, Cdl is allowed to discharge through the shorting transistor, which is usually sized to have low enough resistance that the time constant of discharge is dominated by Rs and Cdl.

The effectiveness of discharging Cdl to achieve charge balance during the shorting phase is thus determined primarily by two factors:

(1) How much time relative to the discharge time constant (nominally, Rs·Cdl) is given for shorting

(2) The impedance of Rf during this time, relative to Rs.

Designing for the worst case in (1) where Rs·Cdl is 20 kΩ · 15 nF = 300 μs, if we enforce a minimum interpulse interval of ~1ms (as is done in the AIS ABEP), then the

electrode only discharges for ~3 time constants. This reduces any charge error down to a factor of e^{-3}, which is roughly 0.05.

Hence our strategy to get less than 25nA of DC current error was to follow a moderately precise (~8-9 bit) dynamic charge balancing scheme with a shorting phase, and reduce an error on the order of 0.4% by a factor of 20 down to ~0.02%.

At this point readers may note that 0.02% of 1mA is 200nA, which is not less than 25nA. However, the current error of 0.02% is not made for a continuous DC current but effectively only during the pulsewidth of a single phase (which is nominally 32 μs). The total biphasic pulse time of 64 μs is actually chosen so that within a 1ms refractory period, all 16 channels will have an opportunity to fire. Thus there is an effective duty-cycling of the error by a factor of about 2×16 = 32. Dividing the 0.02% error, or 200nA out of 1mA by 32× thus gives an average error of only ~6nA. In terms of charge, we arrive at the same result by dividing the 120pC error by a factor of 20 due to shorting, giving 6pC out of every 1ms, which is a DC average of ~6nA.

Note that without shorting, our DC average error with 0.4% error in the stimulation pulses would be a factor of 20 higher, or ~120nA, which is unacceptable. However, rather than getting more bits of precision in the current balance, the ability to use shorting to reduce the error is the simplest solution that allows us to meet the specification.

As for (2), up to now it was assumed that the value of Faradaic leakage due to Rf is not a concern. This assumption is also pervasive in the literature. Perhaps the typical time scale over which stimulation occurs (~100μs) and over which shorting occurs (~1ms) is still considered short compared to the time for electrochemical reactions, and thus no charge is assumed to leak out through Rf.

However, a formal analysis will be done in the following section to derive the limits over which this assumption remains valid for a biphasic pulse. It turns out that if Rf does not remain high enough during the duration of the pulse, then shorting will hurt, rather than help the final charge balance. Ultimately, electrochemical experiments to verify reaction time constants *in vivo* and the corresponding change in electrode impedance must be measured before we can safely rely on shorting to help, rather than hurt, the charge balance.

5.5.2. Limitation of shorting

Consider the simple case of a biphasic current pulse that is injected into a capacitor C with a leak resistor R in parallel, as shown below in (Figure 67):

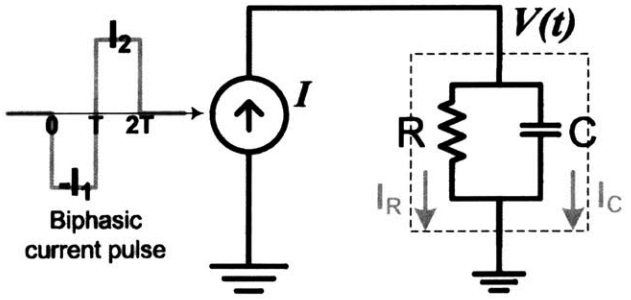


Figure 67. Simple model of biphasic current pulse injected into a parallel RC circuit.

This is the case for the electrode, which has Rf in parallel with Cdl, and just assumes that Rf is a constant R over the duration of the pulse.

Since shorting can only remove the residual charge which is left on C after the pulse, we can start by finding the final voltage V(T) after the negative pulse and then V(2T) after both pulses:

$$
\begin{aligned}
I &= I_R + I_C \\
&= \frac{V(t)}{R} + C\frac{dV(t)}{dt} \\
-RI_1 - V(t) &= RC\frac{dV(t)}{dt} \\
\int_0^{V(T)} \frac{1}{RI_1 + V(t)} dV(t) &= \int_0^T \frac{-1}{RC} dt \\
\ln\left[RI_1 + V(t)\right]_0^{V(T)} &= \frac{-T}{RC} \\
\frac{RI_1 + V(T)}{RI_1} &= e^{-\frac{T}{RC}} \\
V(T) &= -RI_1(1 - e^{-\frac{T}{RC}})
\end{aligned}
$$

$$
\begin{aligned}
I &= I_R + I_C \\
&= \frac{V(t)}{R} + C\frac{dV(t)}{dt} \\
RI_2 - V(t) &= RC\frac{dV(t)}{dt} \\
\int_{V(T)}^{V(2T)} \frac{1}{-RI_2 + V(t)} dV(t) &= \int_0^T \frac{-1}{RC} dt \\
\ln\left[-RI_2 + V(t)\right]_{V(T)}^{V(2T)} &= \frac{-T}{RC} \\
\frac{-RI_2 + V(2T)}{-RI_2 + V(T)} &= e^{-\frac{T}{RC}} \\
V(2T) &= RI_2(1 - e^{-\frac{T}{RC}}) + V(T)e^{-\frac{T}{RC}} \\
&= RI_2(1 - e^{-\frac{T}{RC}}) - RI_1(1 - e^{-\frac{T}{RC}})e^{-\frac{T}{RC}}
\end{aligned}
$$

At this point we should note that even if there is perfect charge balance, i.e. if $I_1 = I_2 = I_0$, V(2T) will still be non-zero:

$$
\begin{aligned}
V(2T) &= RI_2(1-e^{-\frac{T}{RC}}) - RI_1(1-e^{-\frac{T}{RC}})e^{-\frac{T}{RC}} \\
&= RI_0(1-e^{-\frac{T}{RC}}) - RI_0(1-e^{-\frac{T}{RC}})e^{-\frac{T}{RC}} \\
&= RI_0(1-e^{-\frac{T}{RC}})(1-e^{-\frac{T}{RC}}) \\
&= RI_0(1-e^{-\frac{T}{RC}})^2
\end{aligned}
$$

Where does this net positive charge come from? It will become clear if we analyze the three relevant charge quantities:

(1) Error Charge, Qe (from input I):

$$Qe \triangleq (I_2 - I_1)T$$

(2) Residual charge, Qr (left on C):

$$
\begin{aligned}
Qr &\triangleq CV(2T) \\
&= RC(1-e^{-\frac{T}{RC}})(I_2 - I_1 e^{-\frac{T}{RC}})
\end{aligned}
$$

(3) Faradaic charge, Qf (lost through R):

$$
\begin{aligned}
Qf &\triangleq \int_0^{2T} I_R(t)dt \\
&= \int_0^{2T} \frac{V(t)}{R}dt
\end{aligned}
$$

The net positive charge delivered into the RC "black box" is Qe. Thus by conservation of charge, Qe = Qr + Qf. If $I_2 = I_1 = I_0$, meaning that Qe=0, then:

$$
\begin{aligned}
Qr &= RCI_0(1-e^{-\frac{T}{RC}})^2 \\
&= -Qf
\end{aligned}
$$

tells us that there is a net non-zero Faradaic charge pulled through R, even though we have perfect charge balance at the input.

So we can ask: with perfect $I_2 = I_1$ charge balance, why is the negative pulse not perfectly cancelled by the positive pulse, causing some net Qf lost through R, and exactly –Qf left behind on C? If we evaluate Qf as the integral defined above, we can see where the net Qf comes from:

$$
\begin{aligned}
Qf &= \int_0^T -I_1(1-e^{-\frac{t}{RC}})dt && \text{from } \int\frac{V(t)}{R} \text{ over } [0,T] \\
&+ \int_0^T I_2(1-e^{-\frac{t}{RC}}) + \frac{V(T)}{R}e^{-\frac{t}{RC}}dt && \text{from } \int\frac{V(t)}{R} \text{ over } [T,2T]
\end{aligned}
$$

Now it is clear that if $I_2=I_1$, the first integral is cancelled only by the first term in the second integral. It is the second term, i.e. $\frac{V(T)}{R}e^{-\frac{t}{RC}} = -I_1(1-e^{-\frac{T}{RC}})e^{-\frac{t}{RC}}$ that fails to match

even in perfect charge balance, and thereby causes a Faradaic leakage which cannot be neglected.

To interpret this term intuitively, we can separate the effects of the first and second pulses by superposition as shown in (Figure 68) below:

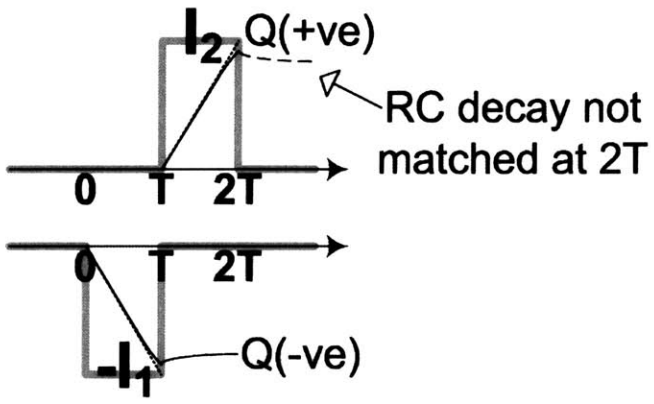


Figure 68. Mismatch of RC leakage causing net Faradaic charge error at time 2T.

Thus, even with perfect charge balance, we should note the following:

(1) charge-buildup will be matched even in RC losses during the accumulation phase of both pulses (the ramp curvatures will be the same).

(2) However, at time 2T, the RC leakage of charge from the first pulse that occurred during the time of the second pulse will not have yet been matched.

To perform a sanity check, we can evaluate the integral of Qf above and show that it is indeed equal to Qe – Qr:

$$
\begin{aligned}
Qf &= \int_0^T -I_1(1-e^{-\frac{t}{RC}})dt && \text{from } \int\frac{V(t)}{R} \text{ over } [0,T] \\
&+ \int_0^T I_2(1-e^{-\frac{t}{RC}}) + \frac{V(T)}{R}e^{-\frac{t}{RC}}dt && \text{from } \int\frac{V(t)}{R} \text{ over } [T,2T] \\
&= \int_0^T (I_2-I_1)(1-e^{-\frac{t}{RC}}) - I_1(1-e^{-\frac{T}{RC}})e^{-\frac{t}{RC}}dt && (\text{let } \tau = RC) \\
&= (I_2-I_1)\left[t+\tau e^{-\frac{t}{\tau}}\right]_0^T + I_1(1-e^{-\frac{T}{\tau}})\left[\tau e^{-\frac{t}{\tau}}\right]_0^T \\
&= (I_2-I_1)T + (I_2-I_1)\tau[e^{-\frac{T}{\tau}}-1] + I_1(1-e^{-\frac{T}{\tau}})\tau[e^{-\frac{T}{\tau}}-1] \\
&= (I_2-I_1)T - \tau[1-e^{-\frac{T}{\tau}}](I_2-I_1e^{-\frac{T}{\tau}}) \\
&= Qe \qquad - Qr
\end{aligned}
$$

The above expression for Qf is useful for another reason, namely to compare the relative contribution between Faradaic leakage due to:

(1) an error in the input (i.e. if $I_2 \neq I_1$), defined as Qfe, and

(2) the systemic (unavoidable) error due to I_1 arriving first and leaking over [T,2T], defined as Qfs.

To do this, we rewrite the expression for Qf as:

$$\begin{aligned} Qf &= (I_2 - I_1)\left[t + \tau e^{-\frac{t}{\tau}}\right]_0^T + I_1(1 - e^{-\frac{T}{\tau}})\left[\tau e^{-\frac{t}{\tau}}\right]_0^T \\ &= (I_2 - I_1)\left(T - \tau[1 - e^{-\frac{T}{\tau}}]\right) - I_1\tau(1 - e^{-\frac{T}{\tau}})^2 \\ &= Qfe \qquad\qquad + Qfs \end{aligned}$$

Since T is nominally 32μs and should be relatively short compared to τ=RC (e.g. τ only needs to be greater than ~20kΩ · 15nF = 300μs for this to be well valid), we can approximate the exponentials in Qfe and Qfs as follows:

$$\begin{aligned} Qfe &= (I_2 - I_1)\left(T - \tau[1 - e^{-\frac{T}{\tau}}]\right) \\ &= (I_2 - I_1)T\left(1 - \frac{\tau}{T}[1 - e^{-\frac{T}{\tau}}]\right) \qquad \text{let } x = \frac{T}{\tau} \\ &= (I_2 - I_1)T\left(1 - \frac{1}{x}[1 - e^{-x}]\right) \\ &\simeq (I_2 - I_1)T\left(1 - \frac{1}{x}[1 - (1 - x + \frac{x^2}{2})]\right) \quad \text{since } x \text{ is small} \\ &= (I_2 - I_1)T\left(1 - \frac{1}{x}[x - \frac{x^2}{2}]\right) \\ &= (I_2 - I_1)T\left(\frac{x}{2}\right) \\ &= (I_2 - I_1)T\left(\frac{T}{2RC}\right) \end{aligned}$$

$$\begin{aligned} Qfs &= -I_1\tau\left(1 - e^{-\frac{T}{\tau}}\right)^2 \\ &\simeq -I_1\tau\left(1 - (1 - \frac{T}{\tau})\right)^2 \quad \text{since } \frac{T}{\tau} \text{ is small} \\ &= -I_1\tau\left(\frac{T}{\tau}\right)^2 \\ &= -I_1 T\left(\frac{T}{RC}\right) \end{aligned}$$

This allows us to interpret Qfe and Qfs as a full scale charge times a fractional error term in parentheses. To compare Qe, which is an error we make in the electronic current balance, with Qfs, we can rewrite $Qe = (I_2 - I_1)T$ as $d \cdot I_0 T$, where I_0 is the mean of I2 and I1 and d is the differential error above and below the mean:

$$\begin{aligned} Qe &= d \cdot I_0 T \qquad \text{where } d \cdot I_0 = I_2 - I_1 \\ &\qquad\quad \textit{i.e.} \qquad I_2 = I_0 + \frac{d}{2} \\ &\qquad\quad \text{and} \qquad I_1 = I_0 - \frac{d}{2} \end{aligned}$$

Now rewriting Qf in terms of I_0 and Qe:

$$
\begin{aligned}
Qf &= Qfe + Qfs \\
&\simeq (I_2 - I_1)T\left(\frac{T}{2RC}\right) - I_1 T\left(\frac{T}{RC}\right) \\
&= d \cdot I_0 T\left(\frac{T}{2RC}\right) - (I_0 - \frac{d}{2})T\left(\frac{T}{RC}\right) \\
&= d \cdot I_0 T\left(\frac{T}{RC}\right) - I_0 T\left(\frac{T}{RC}\right) \\
&= Qe\left(\frac{T}{RC}\right) - Qtot\left(\frac{T}{RC}\right) \qquad \text{Defining } Qtot \triangleq I_0 T, \text{ (the mean total charge)} \\
&= d \cdot Qtot\left(\frac{T}{RC}\right) - Qtot\left(\frac{T}{RC}\right)
\end{aligned}
$$

Conclusion #1: Since both terms are multiplied by the same factor (T/RC), and since Qe is only a small percentage of Qtot, any error (d) that we make in failing to balance I_2=I_1 precisely will ALWAYS contribute a much smaller Faradaic leakage than I_1 arriving first, and leaking over [T,2T].

There is nothing that shorting can do about this Faradaic leakage Qf, since shorting can only reduce the residual charge Qr (any Qf at the point of shorting is irretrievably lost to the electrolyte). To see the effectiveness of shorting Qr, we can express Qr in terms of Qe (the error made by electronics) and Qtot (unavoidable systemic error from I_1 arriving first):

$$
\begin{aligned}
Qr &= Qe - Qf \\
&\simeq Qe\left(1 - \frac{T}{RC}\right) + Qtot\left(\frac{T}{RC}\right) \\
&= d \cdot Qtot\left(1 - \frac{T}{RC}\right) + Qtot\left(\frac{T}{RC}\right)
\end{aligned}
$$

The first term is the fraction of Qr which is due to electronic error (d). If the first term is smaller than the second term, then the dominant charge left on Qr is not due to electronic error but more an "inverse image" of the extra charge which was leaked over [T,2T], namely Qtot(T/RC). We can see this easily if we set d=0 for both Qf and Qr. In this case, we would do better to correct for the large value of Qf by simply letting Qr discharge through R on its own.

However, consider the other extreme where R is enormous (say 1 GΩ). In this case, the fraction T/RC will be on the order of 32μs/10s, or 3.2ppm and completely insignificant compared to the error term d which is more like 0.3% or 3000ppm. It is then much better to short Qr right away, since the dominant charge in Qr will be due to

electronic error Qe.

We can therefore find the Faradaic resistance R above which shorting is beneficial: it is when the first term in Qr is larger than the second:

$$d \cdot Qtot\left(1-\frac{T}{RC}\right) > Qtot\left(\frac{T}{RC}\right)$$

$$d > \frac{\frac{T}{RC}}{1-\frac{T}{RC}}$$

In our case where $d \approx 0.004$, given T and C we can find R:

Let $x = \frac{T}{RC}$, $x(1+d) < d$

Hence $$\frac{T}{RC} < \frac{d}{1+d}$$

$$R > \frac{1+d}{d} \cdot \frac{T}{C} \approx \frac{32\mu s}{0.004(10nF)} = 0.8M\Omega$$

Conclusion #2: The effectiveness of shorting relies significantly on the Faradaic resistance Rf being high, above ~1MΩ over the duration of the pulse.

5.6. Test Results

The dynamic current balancer tested has the following parameters shown in (Table 11):

Table 11. Parameter values for dynamic current balancer

Integrated Capacitors		**Off-chip biases**			
CncM	2pF	Ichg	130nA	In	7.7uA
Cpc	1pF	Icomp	640nA	It	1uA
CpcM	1pF	Vthresh	1.67V		
Cn	5pF			Ibcp	640nA
				Ibcn	640nA
Cc	0.5pF	Ibref	1.2uA		
C	2pF	Ibs	1.2uA	Idelay	150nA
Chsrc	1pF	Ibh	1.2uA	IdelayF	150nA

The test setup is shown in (Figure 69) below:

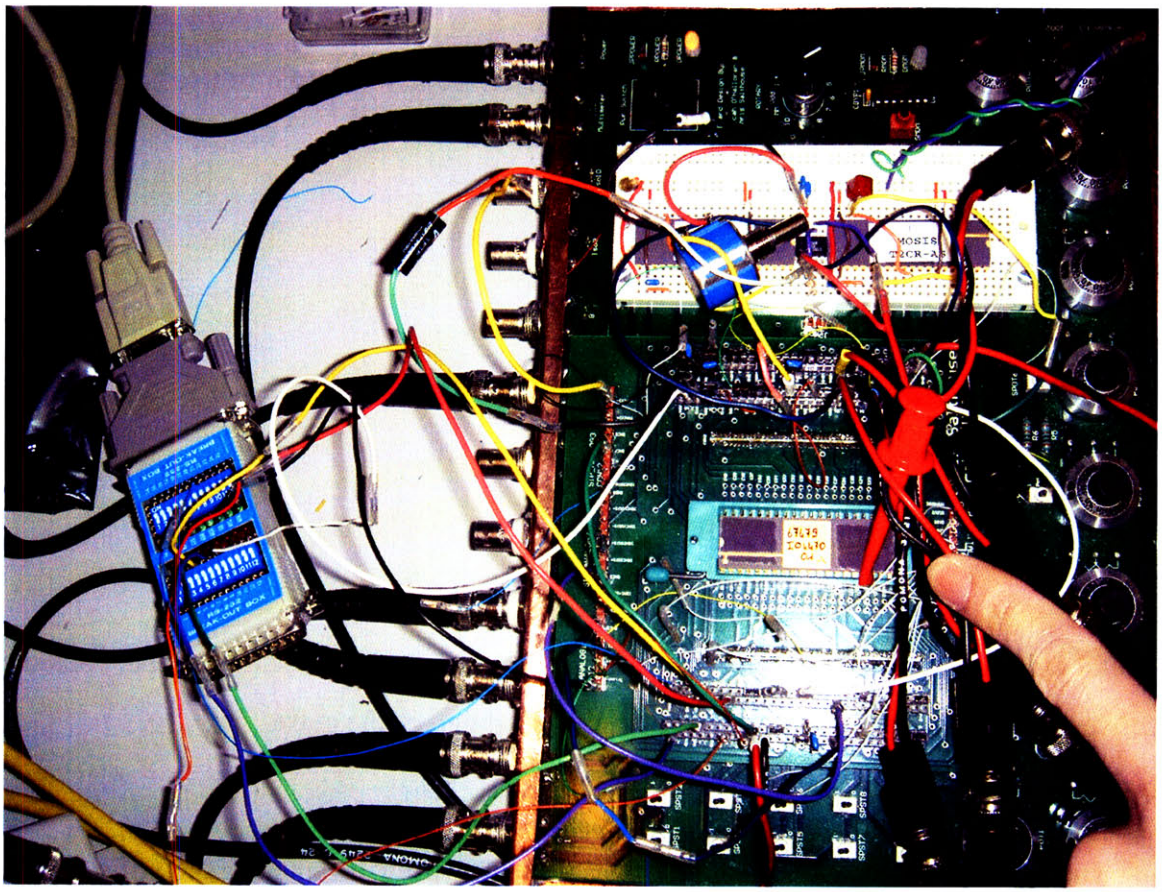


Figure 69. Dynamic current balancer test setup.

Features of note:

- 7 bits from the parallel port were used to automate the setting of Din (the 7-bit current level), brought out from the DB-25 break-out box shown on the left.
- The 4μs clock is brought directly to the pin to minimize clock injection (using a red grabber, shown next to my finger).
- An AD820 buffer on the protoboard is used to buffer the output voltage to the scope (to avoid loading the capacitor with the scope's 1MΩ load).
- A 10nF Teflon capacitor (shown with green and red leads) is used to integrate the biphasic pulse and hence measure the residual error. It has >50 MΩ of shunt resistance when measured at 1kΩ on the LCR meter.
- Pot box design courtesy of Micah O'Halloran and Chris Salthouse.

5.6.1. Biphasic pulse on a resistor and capacitor load

A typical electrode looks much like a resistor and capacitor in series. According to

the design values of Rs and Cdl provided, I used a 10nF capacitor to simulate Cdl and a resistor in the kΩ range to simulate Rs. Plot below in (Figure 70) are the output waveforms when the biphasic pulse at maximum current (Din=127) is delivered to the RC load:

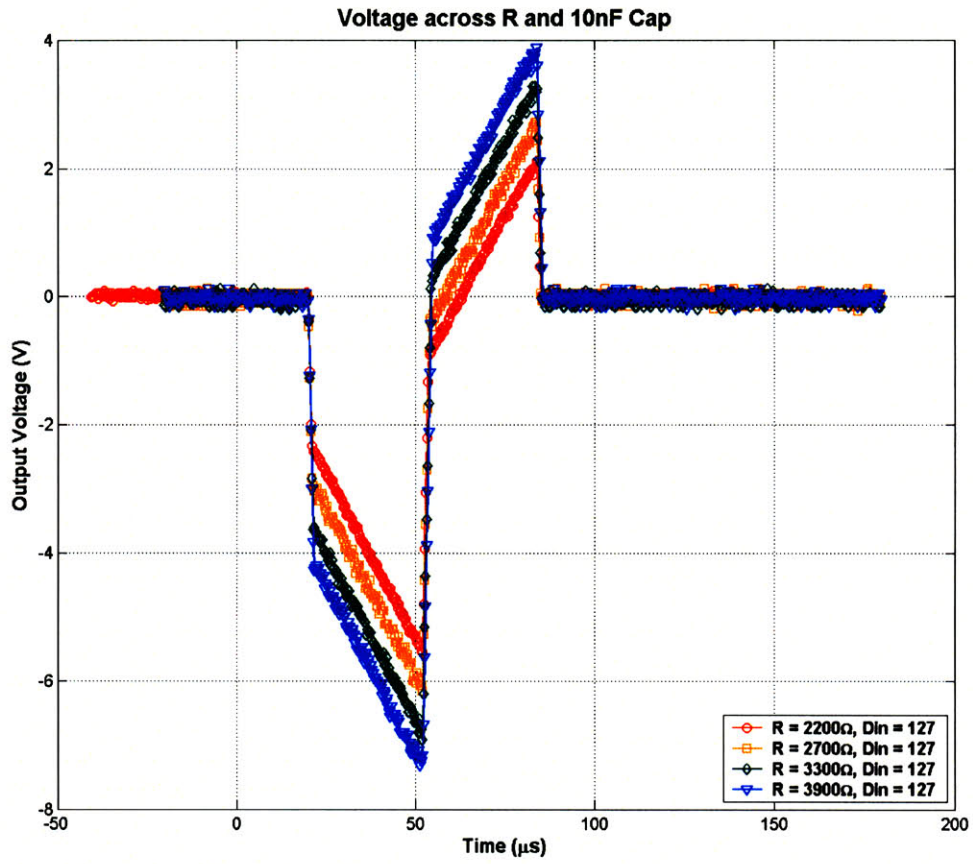


Figure 70. Output voltage Vo measured with various resistive loads.

A few notes on taking this data:

(1) An input pulse on Vox was delivered every 1ms, causing a biphasic pulse to be generated at that rate.

(2) The short_OUT pin had to be connected to Vo, otherwise the capacitor would slowly integrate the net charge error (which was positive) and shift the waveforms up to be centered at ~1V where the reduction in positive current due to early voltage variation balanced out the charge error. So

under these conditions, shorting works very well to reduce the charge error.

(3) Vdd rails of ±7V were used initially for R=2.2kΩ, but to accommodate higher resistance, rails of +6V and -9V were used. Note that the negative rail needs more compliance due to the resistive drop and the capacitive ramp acting in the same direction. On the positive phase, the accumulated negative charge reduces the maximum voltage attained.

Unfortunately, I was not able to get higher than 15V of compliance with this chip. Above a 15V drop from positive to negative supply, the voltage waveforms for internal nodes started to oscillate and look increasingly unstable. I suspect this could be due to an increase in substrate current due to breakdown of the ESD protection diodes: although the drain-source junctions are rated to have over 25V breakdown, I was cautious to include high-voltage protection diodes by tying DMOS gates to sources, but unfortunately the gate-drain junctions are not rated over 5.5V.

A process like the AMI 0.7μm I2T100E with a thick gate oxide layer may allow true 20V compliance to be achieved, but >15V may not be necessary anyway because the power consumption on 20V rails becomes quite prohibitive for bionic implants.

5.6.2. Capacitor voltage waveforms

To measure the DAC linearity, I took a sweep of capacitor voltage waveforms over all 128 levels for Din, to extract the output current level. Sample waveforms for the midpoint (Din=64) and end point (Din=127) are shown in (Figure 71) below:

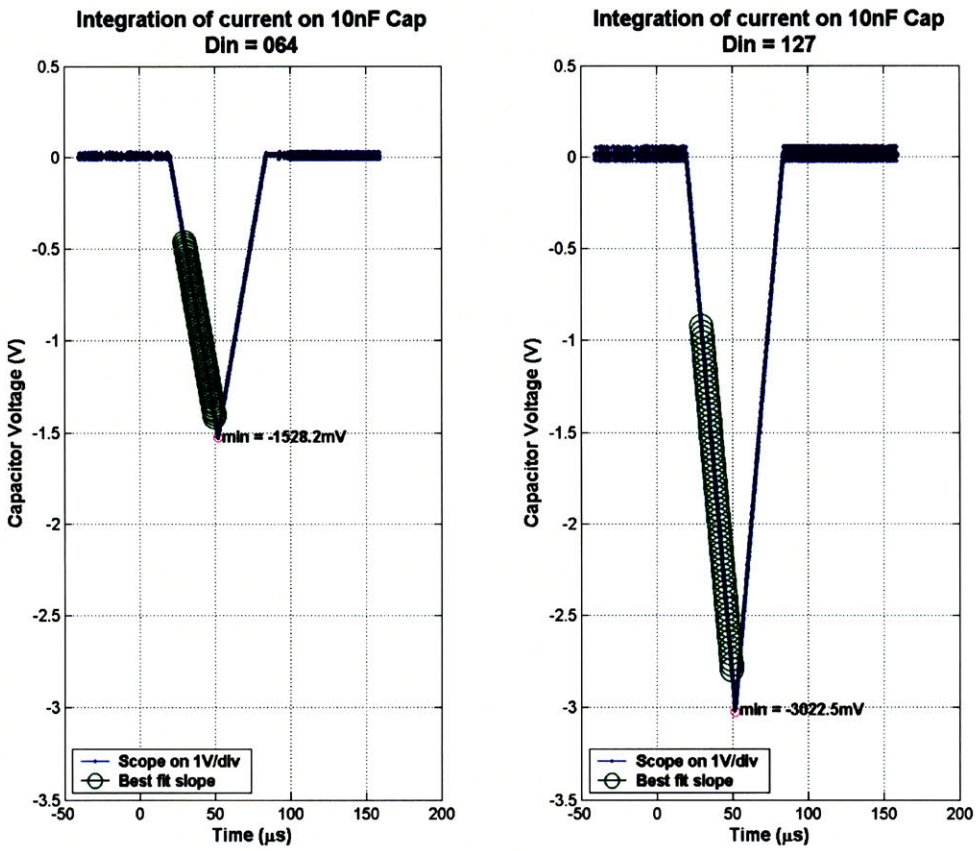


Figure 71. Capacitor ramps for two current levels.

There was a significant amount of clock feedthrough that made the capacitor ramp look bumpy (when you look closely), and so the minimum point was not a precise measure of the output current. I thought a better measure of the output current was to take the best fit slope of the negative ramp from 30µs to 50µs, shown above.

At each current level, I also zoomed in to 10mV/div on the scope to extract the residual integrated charge error given by the voltage at the end of the pulse. Sample waveforms again at Din=64 and Din=127 are shown in (Figure 72) below:

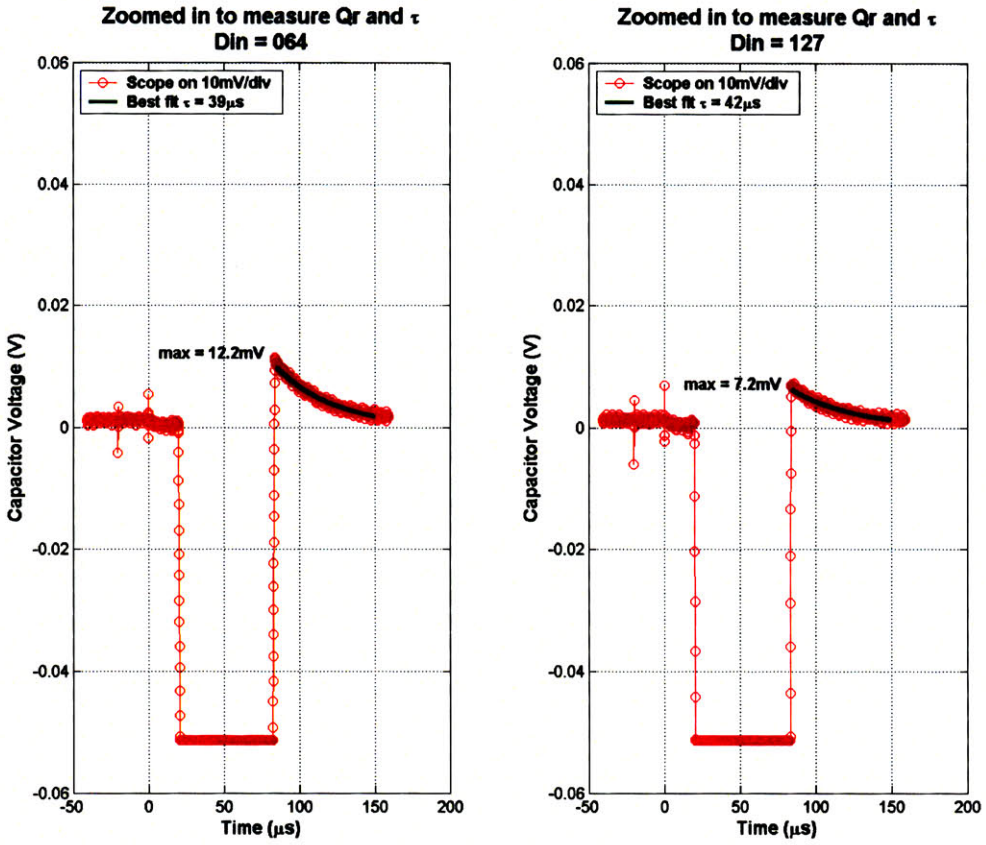


Figure 72. Captured capacitor voltage waveforms at higher magnification.

From these plots I was also able to extract the best-fit exponential time constant of decay in the charge error, which was around 40μs. I was using R = 2.2kΩ for these measurements, so this confirmed that the on-resistance of the shorting transistor was as designed, <2 kΩ. The largest residual charge error was discovered to be around the current midrange, at Din=64.

5.6.3. DAC error measurements

A plot of the residual charge error over all 128 input levels is shown in (Figure 73) below:

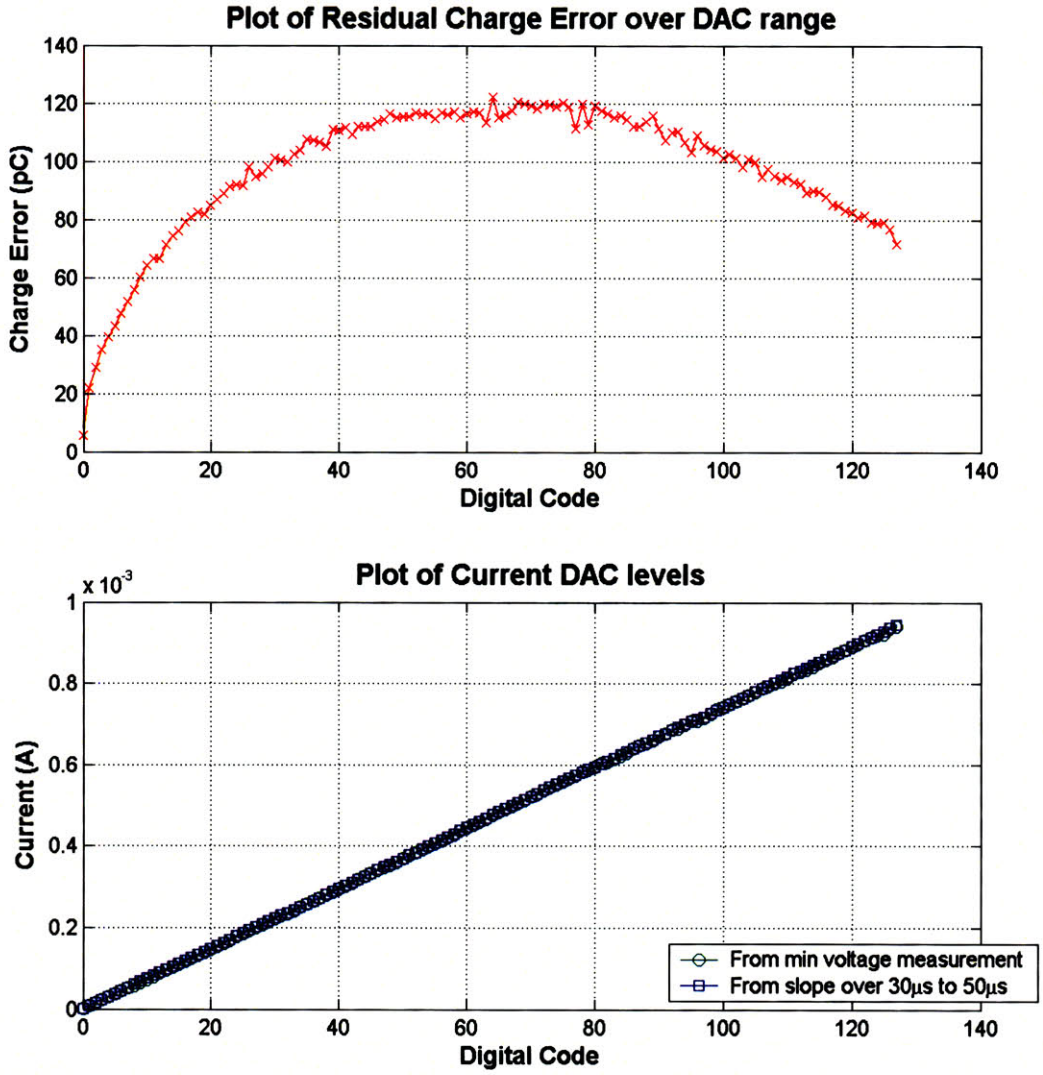


Figure 73. Residual charge error and current DAC output curve.

The maximum residual voltage was converted to a charge residue by simply multiplying by the capacitance of 10nF. As we predicted earlier, the effect of velocity saturation is a likely contributor to the reduction in error residue at higher current levels. The transconductance may even decrease towards the high current of 1mA, and account for the bowed error curve.

Another possible contributor to the reduction in error at high currents is due to a smaller large signal injection onto the hold capacitors from the output of the analog memory, Vhout. At higher currents, since Vp drops, Vhout will swing a smaller voltage

when going from Vref up to Vp.

In practice, we set Vref far below the minimum Vp at ~1V. This is because we do not want the node Vh to spike up and lose charge by turning on the PN junction on the low-leakage switch if Vhout were to make a transient step down. By placing Vref below the minimum Vp, Vhout is guaranteed to only step up, and thus Vh will only spike down. In this way, the PN junction on the low-leakage switch will never be forward biased, but only reverse biased, and thus Vh will lose much less charge in the transient spike. In fact it is likely that this swing in output of the analog memory couples via this mechanism to affect its stored voltage: the error increases slowly to about double the presented values if Vref is raised from ~1V to ~3V.

As mentioned at the end of section 5.4.4, the worst-case charge error of 0.12nC out of 30nC total charge accounts for a mismatch of 0.4% in the current balancer, and relying on >20x further reduction in error from shorting, amounts to <6nA total DC error which is well within the safety limit.

From the plot of DAC output levels against the input code, a plot of the DAC INL and DNL was generated, shown in (Figure 74) below:

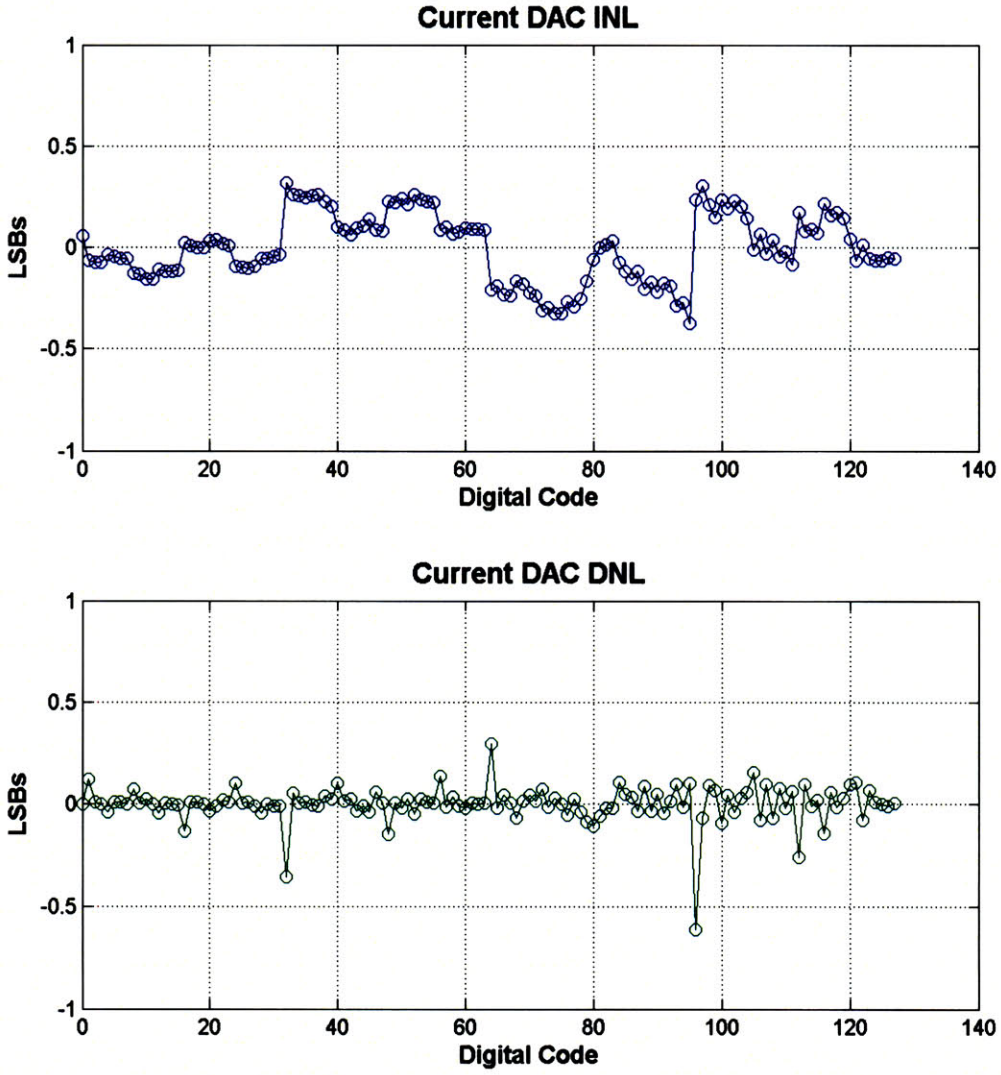


Figure 74. INL and DNL measurements for 7-bit current DAC.

As expected, the major jumps in INL and DNL are at the codes 32, 64 and 96. However, the DAC is clearly 7-bit precise, with well-behaved non-linearity, and is guaranteed to be monotonic since the DNL is <1 LSB for all codes.

5.6.4. Vp settling time measurements

As the major difficulty in simulation was to achieve a settling time that was well within the 16 μs specification, I took some plots to verify the settling time behavior, shown in (Figure 75) below:

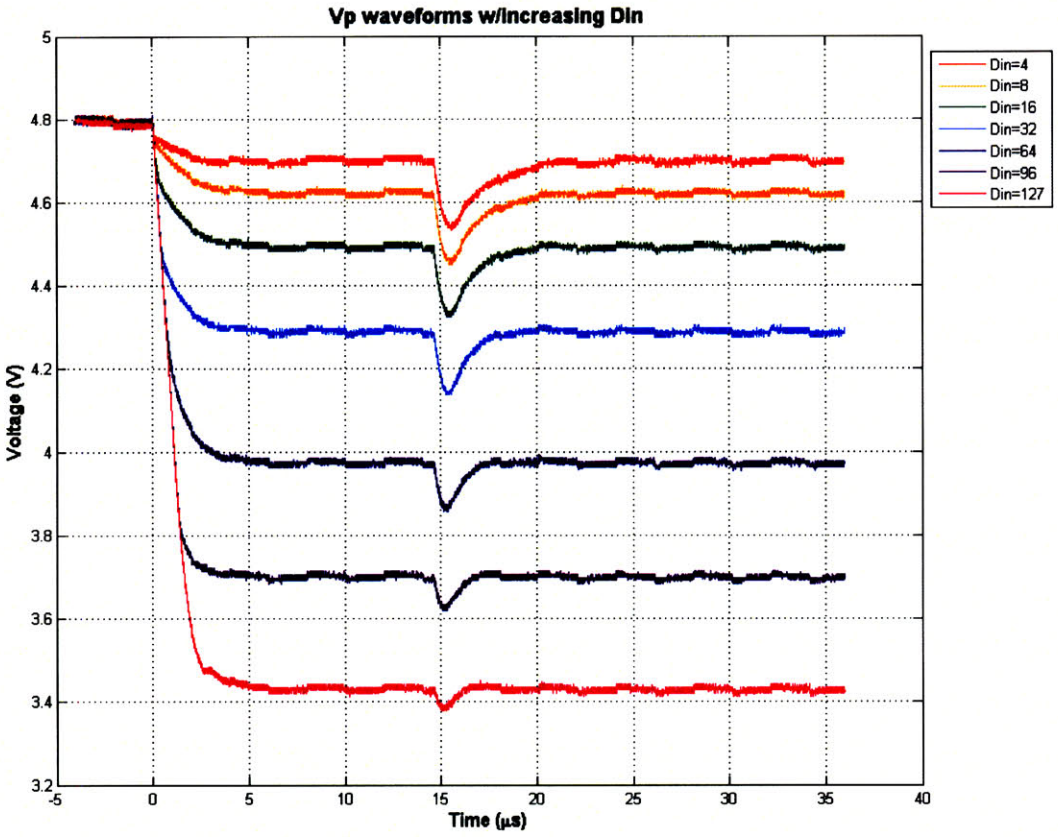


Figure 75. Vp settling time plots.

All input levels show a first-order settling time response, and reaching final value within the 16 µs limit. As mentioned earlier, the HOLD phase must be scheduled to begin before 16 µs (which is when the negative pulse starts), and so the analog timer is shown to set the SAMPLE time just under 15 µs in these plots.

As the current is increased in linear increments from Din=32 to 64 to 96 to 127, we can see that the 3 Vp increments remain constant around 300mV. Hence we can conclude that the g_m of the device has indeed saturated by ~500 µA and maybe even decreased a little, which helps to limit the charge error at higher currents.

A significant source of error is likely to be the feedthrough from the clock, which creates the bumpiness in the curves, although it is not certain that the internal Vp node is similarly affected (shown above is a buffered version to go off chip). Zooming in to look at the Vp waveform, we can see the effect of the clock quite clearly in (Figure 76):

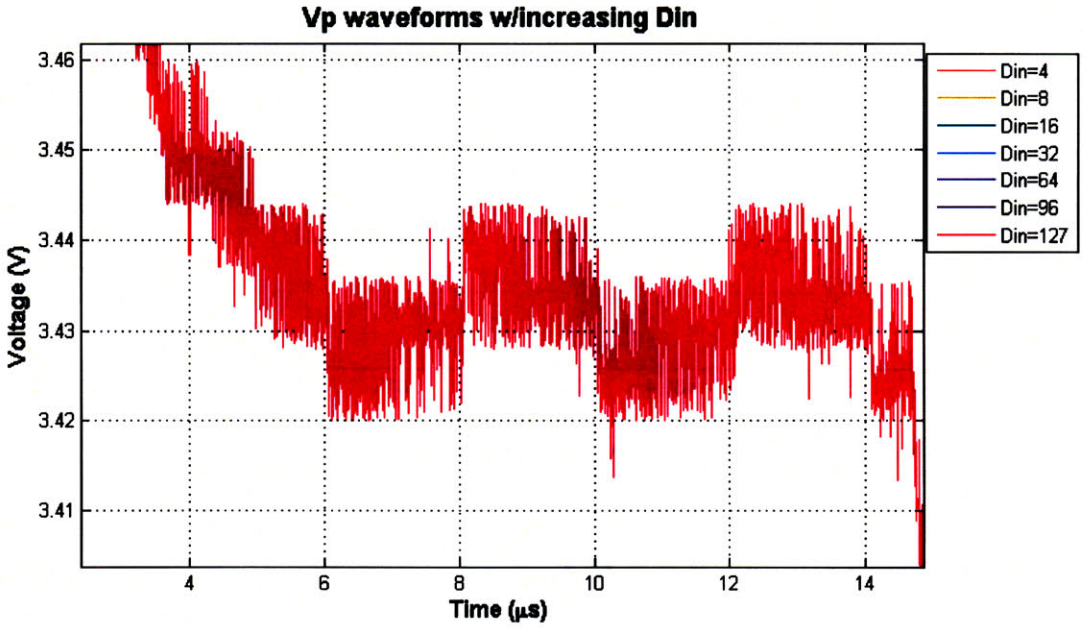


Figure 76. Magnified view of Vp curve to see clock injection.

With almost 10mV injections onto the buffered version of Vp, it is likely that clock injection might be responsible for degrading the expected precision of the charge balance by ~1 bit.

5.6.5. Noise measurements

To verify the level of circuit noise, I used the scope to take multiple readings of the final capacitor voltage after a biphasic pulse. The scope waveforms, together with a probability distribution to show the mean and standard deviation at the end of the pulse (~80.8µs) are shown below in (Figure 77):

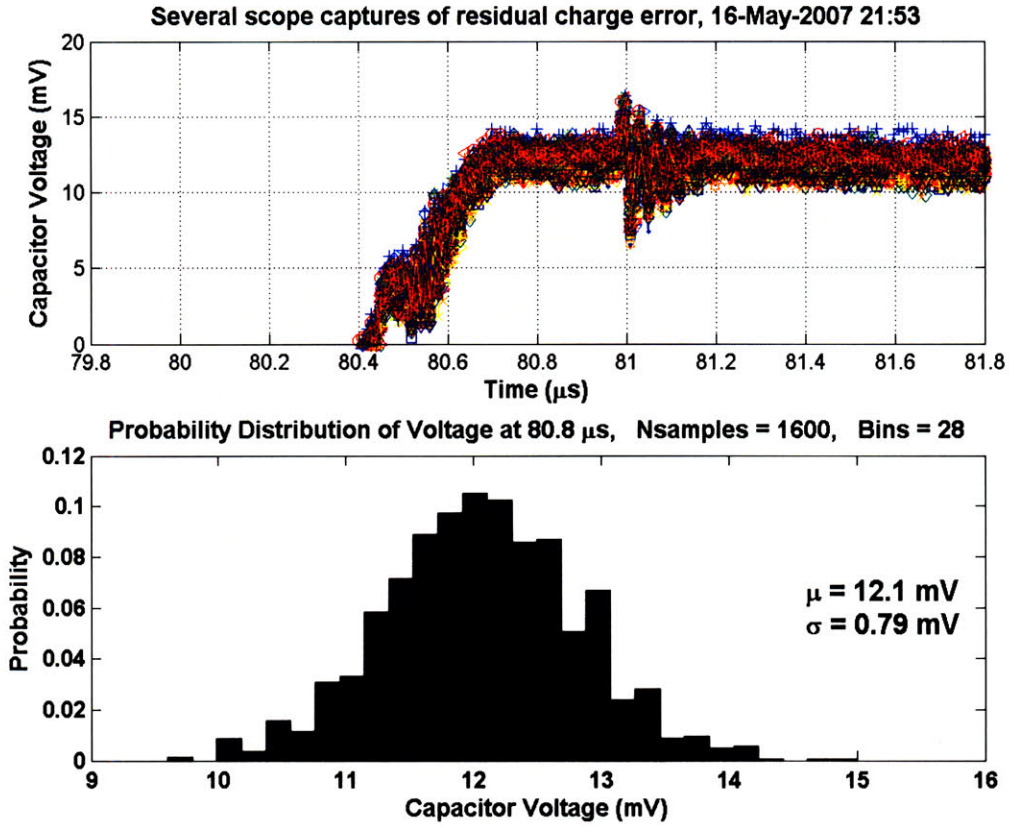


Figure 77. Using multiple scope acquisitions to find the mean and standard deviation of residual error voltage at 80.8μs. The disturbance at ~80.5μs signifies the end of the pulse, and the second disturbance at 81μs is due to the start of shorting.

The standard deviation of ~0.8mV in the residual error voltage translates to an 8pC error on the 10nF capacitor. From the expected noise calculated in section 5.4.3, the measured noise is 3.3× higher than the calculated noise of 2.4pC, but is still a small fraction of the 120pC systematic charge error, and a very small fraction of the 30nC total charge.

Since 8pC of noise is an error on the order of 12 bits smaller than 30nC, the additional noise could come from many sources, including timing jitter in the digital circuits setting the pulsewidth, and 4kTG current noise from the resistor in the SAMP switch connecting the P and N halves of the circuit together. Comparing just the switch resistance contribution against the shot noise contributions:

$$Vp_{,PSD}\ (\text{from transistors}) = \frac{N \cdot 4kT\left(\frac{2}{3}g_m\right)}{g_m^{\ 2}}$$

$$Vp_{,PSD}\ (\text{from resistor, with conductance G}) = \frac{4kT(G)}{g_m^{\ 2}}$$

$$\text{Hence, extra noise factor} = \left(1 + \frac{G}{\frac{2}{3}N \cdot g_m}\right)$$

$$\text{Even at the highest } g_m \simeq 1\,mA/V \text{ and } N = 2,\ \frac{1}{\frac{2}{3}N \cdot g_m} \simeq 750\Omega$$

As R = 1/G may be on the same order as 750Ω, its contribution is not negligible and would account for at least some of the additional noise.

5.6.6. Power consumption

A summary of the power consumption of this chip is shown in (Table 12):

Table 12. Table of power consumption for dynamic current balancer

Current	Not shared (uA)	Supply (V)	Shared (uA)	Supply (V)
Ibref	2.40	6.00	1.20	6.00
Ibs	1.20	6.00	1.20	6.00
Ibh	2.40	6.00	1.20	6.00
Ichg	0.13	3.00	0.13	3.00
Icomp	1.28	3.00	0.64	3.00
Ibcp			0.64	4.50
Ibcn			0.64	10.50
Idelay			0.15	3.00
IdelayF			0.15	3.00
In			7.70	9.00
It			1.00	9.00
Subtotal (uW):	40.23		112.71	
Per channel (uW):	40.23		7.04	

The positive supply is taken as +6V, the negative supply as -9V, and I also assume that we have a low voltage supply for digital logic and other biasing at +3V. Many of the bias distribution currents (i.e. in half of a current mirror) can be shared among 16 channels, for example by distributing 1/16 of the current to each channel and then using 16x mirror gain in the bias mirror to provide the currents that cannot be shared.

Ibcp and Ibcn are cascode currents which by design are not on except during stimulation, hence they are counted as being shared, but require the full positive or

negative supply rail. Hence their supply voltage is a weighted sum of the lower biasing rail and a full supply rail.

Exact values of Idelay and IdelayF are not critical, and in practice, 150nA for both Idelay and IdelayF set up delays < 1μs, which give robust performance over any variation which is < 1μs. Thus IdelayF can be combined with Idelay to create a single gate voltage which is biased with 300nA (for robustness), that is then distributed to all channels.

Finally, In and It can be large and should also be distributed as gate voltages, using tight layout to minimize mismatch between channels, in order to conserve power.

When all the shared powers are divided by 16 channels, the quiescent power consumption per channel currently comes up to 47.3 μW. The power has not been optimized, however, and power could be reduced in the analog memory (Ibh), source follower (Ibs) and buffer (Ibref) since we have met the settling time spec by a significant margin. With perhaps a 50% reduction in the 40μW of unshared power, at the cost of some of the excess bandwidth, the power should not be a big fraction of the total stimulation power, which on high-voltage rails would range from 1mW to even 10mW.

5.7. Summary of contributions

An analysis of electrode impedance and its effect on the problem of charge-balanced stimulation was discussed. We present a two-stage technique which first reduces the mismatch error to 0.4% by dynamic current balancing, followed by a period of shorting that gets the average DC error below 6nA. Shorting is not without its caveats, which are presented, but is relied upon in our approach since it handles a much smaller error than if operating on its own.

The fabricated stimulator is 7-bit accurate, and fully compatible with the asynchronous stimulation protocol in AIS. There are a few refinements that can be explored to improve the design, but as it stands proves that electronic charge balancing can be a viable solution to eliminate DC blocking capacitors.

My particular contributions are in:

1. Designing an electrode driver which is compatible with asynchronous stimulation triggers (e.g. from the AIS ABEP).
2. Achieving 0.4% mismatch in a dynamic current mirror (max of 120pC out of 30nC full scale) with >12V output swing.

3. Dealing with the issues of achieving full settling within a 16μs sample time on 47μW per channel (for a 64μs biphasic pulse).
4. Obtaining low-noise (σ_Q = 8pC out of 30nC full scale, or 11-12 effective bits).
5. Analyzing the limitations of shorting as a technique to reduce residual charge error.

6. CONCLUSION

A complete electrode stimulation architecture for implanted bioelectronics was developed in this thesis. We started with a theoretical underpinning which motivates a shift from the paradigm of synchronous stimulation towards asynchronous stimulation, and developed a bio-inspired strategy for encoding asynchronous stimulation times, which was proven in simulation and perceptual tests to show promise for improving cochlear implant performance. Our strategy was termed Asynchronous Interleaved Stimulation, or AIS.

We went on to show that a practical implementation of asynchronous signal processing for encoding stimulation pulses was possible in low-power analog electronics. This was necessary as a proof-of-concept since asynchronous hardware is unconventional and asynchronous signal processing is expensive when using a synchronous processor. By providing custom hardware which can be made compatible with implantation, our system also is a step towards testing AIS on patients, which is absolutely essential to determine the true merits of the strategy.

Finally, we developed an asynchronous electrode stimulator with current-balanced output drivers, which when combined with shorting as a two-part process, has the potential to allow stimulation to be performed safely without DC blocking capacitors. The understanding of electrode shorting as a technique to achieve charge-balance is also advanced based on consideration of electrode impedances. More work, however, has to be done to establish the biological safety of using electronic charge balancing, and as in any implanted electronic system, failsafes have to be implemented as precautions against device failure.

These results should be useful in the design of next-generation cochlear implants and electrode stimulation systems in general.

7. APPENDICES

7.1. Sample MATLAB code to shift bits into AIS ABEP

```
if exist('dio')
    delete(dio)
    clear dio
end

dio = digitalio('parallel','LPT1');
addline(dio,0,0,'Out','srClk');          % red wire, pin 2
addline(dio,2,0,'Out','srIn');           % yellow wire, pin 4
addline(dio,3,0,'Out','VisSCLK');        % green wire, pin 5
addline(dio,1,1,'In','Ch1_VisSel');      % purple wire, pin 13 (at right edge of port)

if (~exist('g1'))
    g1 = 7;
end
if (~exist('imax'))
    imax = 15;
end
if (~exist('iattack'))
    iattack = 7;
end
if (~exist('irelease'))
    irelease = 7;
end

chan= [1 2 3 4 5 6  7  8  9 10 11 12 13 14 15  16];
taus= [0 2 2 4 7 10 9 12 19 22 30 40 50 78 94 127]; % Chip A 0-127(n)
iref= [4 4 4 3 2 2  2  3  2  3  3  2  2  3  4   2]; % Chip A 0-7(n) minI=0 maxI=7
iref= 7 - iref;                                      % iref is actually 0-7(p)
qs  = repmat(40,1,16);                               % Q=4
atk = [3 3 3 3 3 3  3  3  3  3  3  3  3  3  3   3]; % 0-3(n) minI=0 maxI=3
rel = [7 7 7 7 7 7  7  7  7  7  7  7  7  7  7   7]; % 0-7(p) minI=7 maxI=0
xp5 = [0 0 0 0 0 0  0  0  0  0  0  0  0  0  0   0]; % binary halve Iedout = 0
x2  = [1 1 1 1 1 1  1  1  1  1  1  1  1  1  1   1]; % binary double Iedout = 1
Ia  = [1 1 1 1 1 1  1  1  1  1  1  1  1  1  1   1]; % 0-7(p) minI=7 maxI=0
I3  = [4 3 3 3 3 3  4  3  3  3  6  7  7  7  7   3]; % 0-7(p) minI=7 maxI=0
Ir  = [5 5 5 5 5 5  5  5  5  5  5  5  5  5  5   5]; % 0-7(n) minI=0 maxI=7

% 2kHz/SR=0 1.5kHz/SR=1 1kHz/SR=2 0.5kHz/SR=3
% 0-3(p) minI=3 maxI=0
if (~exist('SR'))
    SR = 0;     % 2kHz by default
end

clear vec;
vec = dec2binvec(iattack,4);
vec = [vec dec2binvec(irelease,4)];
vec = [vec dec2binvec(g1,4)];
vec = [vec dec2binvec(imax,4)];

for ch = 1:16
    vec = [vec dec2binvec(taus(ch),7)];
    vec = [vec dec2binvec(qs(ch),7)];
    vec = [vec dec2binvec(rel(ch),3)];
    vec = [vec dec2binvec(atk(ch),2)];
    vec = [vec dec2binvec(xp5(ch),1)];
    vec = [vec dec2binvec(x2(ch),1)];
    vec = [vec dec2binvec(iref(ch),3)];
    vec = [vec fliplr( dec2binvec(Ia(ch),3) )];     % Ia MSB first, fliplr() needed
    vec = [vec fliplr( dec2binvec(I3(ch),3) )];     % I3 MSB first, fliplr() needed
    vec = [vec fliplr( dec2binvec(Ir(ch),3) )];     % Ir MSB first, fliplr() needed
end

vec = [vec dec2binvec(SR,2)];
vec = fliplr(vec);      % push in the last bit first
```

```
putvalue( dio.srClk, 0 );
for i = 1:length(vec)
    putvalue( dio.srIn, vec(i) );
    putvalue( dio.srClk, 1 );
    putvalue( dio.srClk, 0 );
end
```

8. REFERENCES

8.1. List of Tables

8.2. List of Figures

8.3. Bibliography

Bhatti, P. T. and K. D. Wise (2006). "A 32-Site 4-Channel High-Density Electrode Array for a Cochlear Prosthesis." Solid-State Circuits, IEEE Journal of **41**(12): 2965-2973.

Cauwenberghs, G. and V. A. Pedroni (1994). "A charge-based CMOS parallel analog vector quantizer." Adv. Neural Inform. Proc. Syst (NIPS) **7**: 779-786.

Chen, H. and F.-G. Zeng (2004). "Frequency modulation detection in cochlear implant subjects." The Journal of the Acoustical Society of America **116**(4): 2269-2277.

Collins, L. M., G. H. Wakefield, et al. (1994). "Temporal pattern discrimination and speech recognition under electrical stimulation." The Journal of the Acoustical Society of America **96**(5): 2731-2737.

Dayan, P. and L. F. Abbott (2001). Theoretical neuroscience : computational and mathematical modeling of neural systems. Cambridge, Mass., MIT Press.

deBoer, E. and H. R. deJongh (1978). "On cochlear encoding: Potentialities and limitations of the reverse-correlation technique." The Journal of the Acoustical Society of America **63**(1): 115-135.

Dorman, M. F., P. C. Loizou, et al. (1998). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6--20 channels." The Journal of the Acoustical Society of America **104**(6): 3583-3585.

Dorman, M. F., P. C. Loizou, et al. (1997). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs." The Journal of the Acoustical Society of America **102**(4): 2403-2411.

Friesen, L. M., R. V. Shannon, et al. (2001). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants." The Journal of the Acoustical Society of America **110**(2): 1150-1163.

Fu, Q.-J., R. V. Shannon, et al. (1998). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing." The Journal of the Acoustical Society of America **104**(6): 3586-3596.

Grayden, D. B., A. N. Burkitt, et al. (2004). A cochlear implant speech processing strategy based on an auditory model. Intelligent Sensors, Sensor Networks and Information Processing Conference, Melbourne, Australia.

Kiang, N. Y. S. and E. C. Moxon (1972). "Physiological considerations in artificial stimulation of the inner ear." Ann. Otol. Rhinol. Laryngol. **91**: 714-730.

Lan, N., K. B. Nie, et al. (2004). "A novel speech-processing strategy incorporating tonal information for cochlear implants." Biomedical Engineering, IEEE Transactions on **51**(5): 752-760.

Litvak, L. M., B. Delgutte, et al. (2003). "Improved neural representation of vowels in electric stimulation using desynchronizing pulse trains." The Journal of the Acoustical Society of America **114**(4): 2099-2111.

Litvak, L. M., B. Delgutte, et al. (2003). "Improved temporal coding of sinusoids in electric stimulation of the auditory nerve using desynchronizing pulse trains." The Journal of the Acoustical Society of America **114**(4): 2079-2098.

Litvak, L. M., Z. M. Smith, et al. (2003). "Desynchronization of electrically evoked auditory-nerve activity by high-frequency pulse trains of long duration." The Journal of the Acoustical Society of America **114**(4): 2066-2078.

Lobo, A., F. Toledos, et al. (2002). The effect of envelope low-pass filtering on melody recognition. 33rd Neural Prosthesis Workshop, Bethesda, MD.

Loeb, G. E. (2005). "Are Cochlear Implant Patients Suffering From Perceptual Dissonance?" Ear & Hearing **26**(5): 435-450.

Matsuoka, A. J., J. T. Rubinstein, et al. (2001). "The effects of interpulse interval on stochastic properties of electrical stimulation: models and measurements." Biomedical Engineering, IEEE Transactions on **48**(4): 416-424.

Miller, C. A., P. J. Abbas, et al. (2001). "Response Properties of the Refractory Auditory Nerve Fiber." JARO - Journal of the Association for Research in Otolaryngology **2**(3): 216-232.

Nie, K., G. Stickney, et al. (2005). "Encoding frequency Modulation to improve cochlear implant performance in noise." Biomedical Engineering, IEEE Transactions on **52**(1): 64-73.

Nilsson, M., S. D. Soli, et al. (1994). "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise." The Journal of the Acoustical Society of America **95**(2): 1085-1099.

O'Halloran, M. and R. Sarpeshkar (2004). "A 10-nW 12-bit accurate analog storage cell with 10-aA leakage." Solid-State Circuits, IEEE Journal of **39**(11): 1985-1996.

Oxenham, A. J., J. G. W. Bernstein, et al. (2004). "Correct tonotopic representation is necessary for complex pitch perception." PNAS **101**(5): 1421-1425.

Poissant, S. F., N. A. Whitmal, et al. (2006). "Effects of reverberation and masking on speech intelligibility in cochlear implant simulations." The Journal of the Acoustical Society of America **119**(3): 1606-1615.

Rubinstein, J. T., B. S. Wilson, et al. (1999). "Pseudospontaneous activity: stochastic independence of auditory nerve fibers with electrical stimulation." Hear. Res. **127**: 108-118.

Sarpeshkar, R., M. W. Baker, et al. (2005). "An analog bionic ear processor with zero-crossing detection." Solid-State Circuits Conference, ISSCC '05. Digest of Technical Papers, IEEE International: 78-79.

Sarpeshkar, R., C. Salthouse, et al. (2005). "An ultra-low-power programmable analog bionic ear processor." Biomedical Engineering, IEEE Transactions on **52**(4): 711-727.

Shannon, R. V. (1983). "Multichannel electrical stimulation of the auditory nerve in man. I. Basic psychophysics." Hear. Res. **11**: 157-189.

Shannon, R. V., F.-G. Zeng, et al. (1995). "Speech Recognition with Primarily Temporal Cues." Science **270**(5234): 303-304.

Siebert, W. M. (1986). Circuits, signals, and systems. Cambridge, Mass., MIT Press.

Sit, J.-J., A. M. Simonson, et al. (2007). "A low-power asynchronous interleaved sampling algorithm for cochlear implants that encodes envelope and phase information." Biomedical Engineering, IEEE Transactions on **54**(1): 138-149.

Smith, E. C. and M. S. Lewicki (2006). "Efficient auditory coding." Nature **439**(7079): 978-982.

Smith, Z. M., B. Delgutte, et al. (2002). "Chimaeric sounds reveal dichotomies in auditory perception." Nature **416**(6876): 87-90.

Stein, R. B. (1965). "A Theoretical Analysis of Neuronal Variability." Biophysical Journal **5**: 173-194.

Taylor, S. R. and E. Gileadi (1995). "Physical Interpretation of the Warburg Impedance." Corrosion **10**: 664.

Tong, Y. C. and G. M. Clark (1985). "Absolute identification of electric pulse rates and electrode positions by cochlear implant patients." The Journal of the Acoustical Society of America **77**(5): 1881-1888.

Townshend, B., N. Cotter, et al. (1987). "Pitch perception by cochlear implant subjects." The Journal of the Acoustical Society of America **82**(1): 106-115.

van Hoesel, R. J. M. and R. S. Tyler (2003). "Speech perception, localization, and lateralization with bilateral cochlear implants." The Journal of the Acoustical Society of America **113**(3): 1617-1630.

Vandali, A. E., C. Sucher, et al. (2005). "Pitch ranking ability of cochlear implant recipients: A comparison of sound-processing strategies." The Journal of the Acoustical Society of America **117**(5): 3126-3138.

Wessel, R., C. Koch, et al. (1996). "Coding of time-varying electric field amplitude modulations in a wave-type electric fish." J Neurophysiol **75**(6): 2280-2293.

Wilson, B. S., C. C. Finley, et al. (1991). "Better speech recognition with cochlear implants." Nature **352**(6332): 236-238.

Xu, J., R. K. Shepherd, et al. (1997). "Chronic electrical stimulation of the auditory nerve at high stimulus rates: a physiological and histopathological study." Hearing Research **105**(1-2): 1-29.

Zeng, F.-G. (2002). "Temporal pitch in electric hearing." Hearing Res. **174**: 101-106.

Zhak, S. M., M. W. Baker, et al. (2003). "A low-power wide dynamic range envelope detector." Solid-State Circuits, IEEE Journal of **38**(10): 1750-1753.