Retinomorphic Vision Systems: Reverse Engineering the Vertebrate Retina

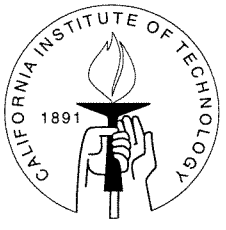
Thesis by

Kwabena A. Boahen

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

1997

(Submitted November 22, 1996)

Acknowledgements

It gives me great pleasure to thank all the wonderful people I got to know on this journey.

Thanks to Carver Mead, who showed me the light.

Thanks to Andreas Andreou, who showed me the possibilities.

Thanks to Peter Sterling, Christof Koch, Alain Martin, Al Barr, John Almann, Richard Andersen, Frank Werblin, and Terry Sejnowski who guided me.

Thanks to Misha Mahowald, Tobi Delbrück, and John Lazzaro who cleared the path.

Thanks to Bhusan Gupta, Ron Benson, Lloyd Watts, Andrew Moore, Carlos Brody, and Marcus Mitchell who encouraged me.

Thanks to Frank Eeckman, Martin Lades, Joachim Buhman, and Thomas Wachtler who collaborated with me.

Thanks to Rahul Sarpeshkar, Lena Peterson, Chris Diorio, Jeff Dickson, Shii-Chii Liu, Brad Minch, and Paul Hassler who accompanied me.

Thanks to William Ceasarotti, Sarah Laxton, and Dazhi Chen who built the gadgets I needed, and to Gary Stupian who did chip surgery.

Thanks to Helen Derevan, Candace Schmidt, Donna Fox, Sandra Renya, Laura Rodriguez, Calvin Jackson, and Jim Campbell who helped in ways too numerous to mention.

And thanks to Lyn Dupré who helped me to describe this journey to you in style.

Abstract

This thesis seeks to explain how the retina satisifies both top-down constraints (functional) and the bottom-up constraints (structural) by analyzing simple physical models of the retina and mimicking its structure and function in silicon. In particular, I examine spatiotemporal filtering in the outer plexiform layer of the vertebrate retina, and show how outer retina processing is augmented by further processing in the inner plexiform layer, creating an efficient implementation that encodes moving stimuli efficiently over a wide range of speeds.

My working hypothesis is that biological sensory systems seek to optimize both functional and structural constraints. On the functional side, they must maximize information uptake from the environment while they minimize redundancy in their outputs. On the structural side, they must maximize resolving power in space and time, by making the processing elements small and fast, while they minimize wiring and energy consumption. If structure and function did indeed coevolve, as I assume, studying how structural and functional contraints are optimized simultaneously is our only hope of understanding why nature picks the solutions that we observe.

Addressing both structural and functional contraints requires combining science and engineering. Scientists study an existing structure, and seek to understand how it functions in an optimal or near-optimal fashion, based on theoretical grounds. Rarely does a scientist ask: Will the structure be more cost effective, more reliable, or more reproducible if a less-than-optimum function is chosen? Engineers, on the other hand, design an optimal implementation for some desired function, based on an exisiting set of standard primitives. Rarely does an engineer ask: Is this the most natural set of primitives to use for this particular function? Thus, neither discipline attempts to optimize both function and structure globally. In contrast, evolution, operating in a purely opportuinistic fashion, continuously seeks increasingly elegant solutions that meet these constraints.

For these reasons, I have adopted a multidisciplinary engineering–science approach that combines analysis with synthesis. When tailored synergestically, this approach can shed light on questions about which neurobiologists care, while advancing the state of the art in sensory-system design.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

The retina is an exquisitely evolved piece of biological wetware. The human retina—as well as other vertebrate retinae—is sensitive to light intensities ranging from dim starlight to direct sunlight: a dynamic range of at least 10 decades. This remarkable ability to adapt to changes in intensity larger than those handled by any other known sensory system is mediated by a variety of gain-control mechanisms that operate over disparate spatial and temporal scales. The vertebrate retina has evolved specialized pathways and elaborate network-control mechanisms that fine tune the degree of pooling and the integration time of these pathways and share elements between them. The existence of all these specialized pathways channels makes the retina a complex, multifaceted structure.

I review the anatomy of the retina in Chapter 2, and, in particular, I describe five specialized channels: a milliphoton-sensitivity channel for night vision, a minute-of-arc acuity channel for luminance, a millisecond-acuity channel for motion, and two channels for chrominance. In my review, the emphasis is placed on those facets of the retina that shed light on how spatiotemporal signals are processed and how motion is encoded.

Even this extremely truncated and oversimplified review of retinal neurobiology makes it abundantly clear that the retina is much more complex than any sensory system currently built by engineers. The retina's parallel dedicated channels make it akin to several specialized cameras coexisting on the same chip. Even if we try to get around this multifaceted character by focusing on just one of these cameras, we are still bewildered because the elements of the cameras are richly interconnected, and the same element may serve several purposes at the same time, or it may be coopted by different cameras at different times. This nonmodularity, which is a defining characteristic of the retina—and of the rest of the brain—results in a efficient implemetation but it makes it extremely difficult for us to understand how the system

operates by using traditional reductionist approaches.

Having studied retinal structure, I turn my attention to retinal function in Chapter 3.

The spike trains produced by the retina are converted back into continuous signals by dendritic integration of excitatory postsynaptic potentials in the lateral geniculate nucleus of the thalamus. For human vision, contrast thresholds of less than 1%, processing speeds of about 20 ms per stage, and temporal resolution in the millisecond range are achieved with spike rates as low as a few hundred per second. No more than 10 spikes, per input, are available during this time. The retina must maximize the amount of information carried by these spikes.

For optimum performance, the retina must efficiently encode stimuli generated by all kinds of events, over a large range of lighting conditions and stimulus velocities. These events fall into three broad classes: static events, punctuated events, and dynamic events. In the absence of any preprocessing, the output activity mirrors the input directly. Changes in lighting, which influence large areas, are reflected directly in the output of every single pixel in the region affected. Static events, such as a stable background, generate persistent activity in a large fraction of the output cells, which transmit the same information repeatedly. Punctuated events generate little activity and are transmitted without any urgency. Dynamic events generate activity over areas far out of proportion to informative features in the stimulus, when the stimulus sweeps rapidly across a large region of the retina. Clearly, these output signals are highly correlated, over time and space, resulting in a high degree of redundancy. Hence, reporting the raw intensity values makes poor use of the limited throughput of the optic nerve.

The retina has evolved sophisticated filtering and adaptation mechanisms to reduce redundancy and to improve coding efficiency. These mechanisms include: Local automatic gain control at the receptors, bandpass spatiotemporal filtering in the outer retina, highpass temporal and spatial filtering in the inner retina, half-wave rectification, spike frequency adaptation, and a foveated architecture. As a result activity in the ganglion cells, which convert these preprocessed signals to spikes and transmit

the spikes over the optic nerve, is different from the stimulus pattern. The activity in the optic nerve is clustered in space *and* time (whitened spectrum): It consists of sporadic short bursts of rapid firing, triggered by punctuated and dynamic events, overlaid on a low, steady background firing rate driven by static events.

To unify retinal structure and function, I duplicate the retina's spatiotemporal dynamics with a simple physical model in Chapter 4.

My goal is to synthesize the minimal amount of machinery required to reproduce the observed qualitative behavior, rather than to provide detailed quantitative predictions of retinal responses. This approach is part of an overarching layered-complexity strategy that I have adopted, where we reverse-engineer the retina by peeling away one level of complexity at a time. Once we know the tradeoffs inherent in the design of a piece of neurocircuitry, we can see how to introduce an additional layer of complexity to improve its performance. Although a linear model cannot include adaptation mechanisms, such as gain control, we can often achieve the desired result by varying the parameters of the linear circuit, such as its gain or its time and space constants, appropriately. Layering adaptation on top of filtering in this fashion is valid, as these two mechanisms act on disparate spatial and temporal scales.

Models of the retina similar to the one that I study here have been proposed and analyzed. However, none of the previous studies analyzed the effect of the model's spatiotemporal inseparability on motion. By studying a minimal model, and treating space as a continuum—using the continuous approximation—just like time, I was able to obtain closed-form analytic solutions, and to develop a clear intuitive picture of the spatiotemporal behavior of the retina. I show that the model's spatiotemporal inseparability has serious consequences for how information about contrast and speed is encoded by the retina. It also results in suboptimal filtering, as the model's spatiotemporal behavior deviates from the optimal filter for the ensemble of natural images.

I show how spatiotemporal inseparability goes hand in hand with local connectivity. As a consequence, nature must choose between a costly spatiotemporally separable optimal filter or a cheap spatiotemporally inseparable suboptimal filter,

weighing coding efficiency against implementation efficiency. By unifying structural bottom-up constraints and functional top-down constraints in this way, I provide an explanation for two key aspects of retinal organization, which are preserved across a large variety of species:

- The retina encodes several parallel information streams in its output that emphasize different aspects of a scene, such as color, edges, and movement. In particular, it has one channel with high spatial resolution and low temporal resolution, and another channel with low spatial resolution and high temporal resolution.

- To a good first approximation, the retinae of all vertebrate species can be described as a locally connected feedforward neural network with three cellular layers that are connected by two layers of processing: the outer plexiform layer (OPL) and the inner plexiform layer (IPL).

The unification of retinal structure and function through theoretical analysis of a physical model concludes the first part of these thesis.

In the second part of the thesis, I switch gears and describe how to replicate neural systems in silicon by exploiting similarities between the biophysics of nerve cells and the physics of MOS transistors.

I begin by comparing and contrasting electrodiffusion in nerve membranes and in MOS transistors in Chapter 5. This comparative study—which, to my surprise, has not yet been done—shows us how best to exploit the native physics of the transistor to model the biophysics of the nerve membrane. The similarities between these two structures are most evident at the microscopic level, since the physics that governs their behavior is the same. Balancing drift and diffusion results in equilibrium concentration profiles that decrease exponentially with potential in both devices.

At the macroscopic level, these devices are qualitatively similar. A pMOS device reproduces the qualitative behavior of a cation channel that sees a higher concentration inside the cell, or of an anion channel that sees a higher concentration outside the cell. And an nMOS transistor reprodces the qualitative behavior of a cation channel

that sees a higher concentration outside the cell, or of an anion channel that sees a higher concentration inside the cell.

However, the membrane's current–voltage relationship has linear asymtotic behavior whereas the transistor's asymtotic behavior is exponential. This difference arises because the concentrations of holes in the the drain–source regions of a pMOS transistor are millions of times larger than the concentration of holes in the n-type bulk. A similar situation holds for electrons in the nMOS transistor. In contrast, the ions that are primarily responsible for the electrical properties of the cell—namely, K^+ and Na^+—have concentration ratios of 1 or 2 decades. We can match the ion-channel's current-voltage curve quantitatively by reducing the doping of the source–drain regions by four or more decades.

I show that when all the ion channels see the same voltage difference—as they do when they are part of the same cell—the relative differences between the currents in different ion-channel populations may be reproduced fairly well using transistors. Thus, we can build a fairly decent single-cell model in a standard CMOS process by using a single transistor to model each population of ion channels. In particular, the model reproduces the behavior of the cell at equilibrium (i.e., the dependence of the resting membrane potential on channel permeability, which is described by the Goldman–Hodgkin–Katz equation [1, 2]).

In Chapter 6, I go beyond the single cell and present implementations for multiple-cell networks. In particular, I propose transistor-based models for gap-junction–coupled cell syncytia. Such syncytia are common in the retina, and they occur in other parts of the brain as well.

I extend the device-level charge-based formulation of the MOS transistor to the circuit level by introducing the concepts of terminal and node charges, and the equivalence principle. With this formalism, we can exploit the linear current–charge relationship of the MOS transistor at the circuit level, enabling us to simulate the diffusion of ions in cell syncytia, or the spread of current in resistive netorks, extremely efficiently. When node charges stand in for membrane voltages, we may model the linear current–voltage relationship of the gap junction with the linear current–charge

relationship of transistors in the subthreshold regime. This analogy enables us to simulate the spread of ions in cell syncytia extremely efficiently.

We can use these single-transistor diffusors to model the lateral spread of these ions, as well as the loss of ions through leakage into the extracellular fluid. These two mechanisms define a local neighborhood over which signals summate, and we can control the size of this region by the relative strengths of the lateral coupling between nodes in the network and the leakage path from these nodes to ground. When we use transistors acting as diffusors, we can control the size of this region electronically, and thereby we can actively regulate local aggregation. The extent of local aggregation determines the extent of collective computation. Cell syncytia can regulate the extent of local aggregation as well. The retina exploits this ability to trade off signal-to-noise ratio for bandwidth.

In Chapter 7, I show how we can model excitatory and inhibitory chemical synapses with single transistors. Together with the single-transistor model of gap junctions, I use these neural analogs to morph the neurocircuitry in the outer retina into silicon. The result is a CMOS circuit that models bandpass spatiotemporal filtering in the outer retina—at the same level of abstraction as the linear electrical circuit model that we studied in Chapter 4. In contrast to the linear physical model in Chapter 4, the CMOS circuit includes a local gain-control mechanism. This nonlinear mechanism models the effect of shunting inhibition from the horizontal cells to the cones.

Unlike the abstract theoretical circuit model, the actual parameters of nominally identical circuit elements on the chip vary from location to location, due to the vagaries of the fabrication process. Consequently, building the model in silicon helps us to understand the effects of structural peturbations and quantum fluctuations on performance, as well as the effects of local gain control on bandpass filtering. It also forces us to address structural constraints, such as the energy and area costs of communication versus computation, which I discussed briefly in the first part of this thesis (Chapter 4).

I analyze the performance tradeoffs that must be made to get spatiotemporal

bandpass filtering and local gain control to coexist in this minimal circuit design. In particular, the high loop gain required to attenuate low-frequency temporal and spatial signals in a negative-feedback circuit results in temporal instability. And controlling the gain by modulating the intercone coupling conductance in proportion to the local intensity causes the receptive field to expand alarmingly. These shortcomings of the simple circuit model of the outer retina that I built forced me to review the retina literature in search of mechanisms that decouple spatiotemporal filtering and local gain control. I found that autofeedback in horizontal cells could provide an elegant solution to this dilemma.

To transmit the graded signals produced by the outer plexiform circuit, I follow the retinal model and develop an adaptive spiking neuron circuit in Chapter 8.

To impement spike frequency adaptation and membrane time-constant adaptation, I introduce three simple circuit elements that model the biophysics of voltage- and calcium-dependent potassium channels. A diode-capacitor integrator models the accumulation and buffering of intracellular calcium. Capacitive coupling between the membrane-voltage node and the calcium-integration node models the fast voltage dependence of the potassium channels. A single transistor, with its gate tied to the calcium-integration node, models the potassium-channel population.

I analyze the effects of these mechanisms, with emphasis on spike timing, and compare my theoretical predictions with experimental measurements. I characterize spike-timing precision by measuring how much time the neuron takes to respond to a step change in its input by firing a spike. I measure the distribution of these firing times over several trials, and define the latency as the position of the peak in the distribution and the synchronicity as the height of the peak, normalized by the height of the uniform distribution. For the same average steady-state firing rate, the calcium dependence and the voltage dependence of the potassium channels improved the adaptive neuron's latency and synchroncity, compared with a simple integrate-and-fire model.

These results call into question several common notions about how neurons encode information. Neurobiologists generally believe that the mean firing rate is a valid

measure of the efficacy of a neuron in producing a response in its target. Furthermore, if the target neuron listens to several neurons, they obtain the net effect by summing their mean firing rates. For such linear summation to be valid, the postsynaptic neuron must smooth out fluctuations in firing rates, or the presynaptic neurons must fire at uniform rates and in an uncorrelated fashion. My measurements invalidate both assumptions, and are in agreement with more recent physiological studies [3].

Neurons are exquisitely sensitive to small changes in their input, and can generate a spike in response to these changes in less than 1 millisecond. Consequently, instead of smoothing out variations in their inputs, they amplify these variations. Second, the latencies are much shorter than the interspike interval, and so the instantaneous firing rate that the target neuron observes when several spike trains converge in its dendritic tree may be much higher than you would expect from simply summing the individual rates. Neurons can use this synchronicity to amplifiy their firing rates. We overlook this mechanism completely when we use mean firing rates and ignore spike timing.

In the final chapter, I describe a retinomorphic vision chip that uses neurobiological principles to perform all four major operations found in biological retinae: continuous sensing for detection; local automatic gain control for amplification; spatiotemporal bandpass filtering for preprocessing; and adaptive sampling for quantization. All four operations are performed right on the focal plane, at the pixel level.

The first—and only—attempt to integrate these four operations was made by Mahowald. The pixel that she designed, which is described in her monograph [4], used continuous sensing for detection, logarithmic compression for amplification, temporal highpass filtering for preprocessing, and a simple integrate-and-fire neuron for quantization. My work improves on, and extends, Mahowald's pioneering research in three ways:

1. By using local gain control for amplification, I extend the dynamic range without sacrificing sensitivity; logarithmic compression, in contrast, trades sensitivity for dynamic range.

2. By using a spatiotemporal bandpass for preprocessing, I cut out wideband spatial and temporal noise; highpass filtering, in contrast, amplifies high-frequency signals with poor signal-to-noise ratios.

3. By using an adaptive neuron for quantization, I increase the sampling rate—and reduce the latency—without increasing the average firing rate; a simple integrate-and-fire neuron, in contrast, must maintain a high steady-state firing rate to sample high-frequency signals.

Like Mahowald's chip, my retinomorphic chip includes a random-access time-division multiplexed communication channel that reads out asynchronous pulse trains from a 64- $\times$ 64-pixel array in the imager chip. The communication channel transmits these spike trains to corresponding locations on a second chip that has a 64 $\times$ 64 array of integrators. Both chips are fully functional. This VLSI chip embodies four principles of retinal operation.

First, the imager adapts its gain locally to extend its input dynamic range without decreasing its sensitivity and without increasing its output dynamic range. The gain is set to be inversely proportional to the local intensity, discounting gradual changes in intensity and producing an output that is proportional to contrast. This adaptation is effective because lighting intensity varies by six decades from high noon to twilight, whereas contrast varies by at most a factor of 20 [6].

Second, the imager bandpass filters the spatiotemporal visual signal to attenuate low-frequency spatial and temporal signals, and to reject wideband noise. The increase in gain with frequency, for frequencies below the peak, matches the $1/f^2$ decrease in power with frequency for natural image spectra, resulting in a flat output power spectrum. This filtering improves information coding efficiency by reducing correlations between neighboring samples in space and time. It also results in a unimodal distribution of pixel amplitudes which is centered on the middle of the output range, and typically decays exponentially in either direction.

Third, the imager adapts its sampling rate locally to minimize redundant sampling of low-frequency temporal signals. In the face of limited communication resources

and energy, this sampling-rate adaptation has the additional benefit of freeing up the bandwidth of the communication channel, which is dynamically reallocated to active pixels, allowing higher peak sampling rates and shorter latencies to be achieved.

Fourth, the imager adapts its step size locally to trade resolution at high contrast levels, which rarely occur, for resolution at low contrast levels, which are much more common. The proportional step size in the adaptive neuron, which results in a logarithmic transfer function, matches an exponentially decaying amplitude probability density, making all quantization intervals equiprobable. Hence, it maximizes the expected number of signals that can be discriminated, given their probability of occurrence.

For independent samples, information is linearly proportional to bandwidth, and is logarithmically proportional to the signal-to-noise ratio [8]. We increase bandwidth by making the receptors smaller and faster, so that they can sample more frequently in space and time. As an unavoidable consequence, they integrate over a smaller volume of space–time, and therefore the signal-to-noise ratio degrades. There is therefore a reciprocal relationship between bandwidth and noise power (variance) [9]. Since their goal is to maximize information, biological sensory systems aggressively trade off signal-to-noise ratio for bandwidth, operating at ratios close to unity [10, 9].

With this optimization principle in mind, I developed compact circuit designs that realize local AGC, bandpass filtering, and adaptive quantization at the pixel level. The overriding design constraints are to whiten the signal, thus making samples independent; to minimize the pixel size, and capacitance, thus making sampling more dense and more rapid; and to minimize power consumption, thus making it possible to acheive very large-scale integration. Hence, all circuits use minimal-area devices and operate in subthreshold, where the transconductance per unit current is maximum. My work demonstrates that extremely efficient and robust information processing systems may be realized by modeling the structure *and* function of neural systems.

Chapter 2 Retinal Structure: Parallel Pathways

The **retina** is an exquisitely evolved piece of biological wetware. It contains about 100 million photoreceptors. Its output—a million or so axonal fibers that make up the optic nerve—conveys visual information to the rest of the nervous system using an all-or-none pulse code.

The human retina—as well as other vertebrate retinae—is sensitive to light intensities ranging from dim starlight to direct sunlight: a dynamic range of at least 10 decades.[1] This range is parceled out between the rods and the cones. **Rods** operate in dim light and can sense the absorption of a single photon, but they saturate at a 100 photons per integration time [12, 13]. **Cones** are 70 times less sensitive than are rods, so their range extends almost 2 decades higher [14].

To deal with the remaining 6 decades, the cones shift their 2-decade output sensitivity range to match the input intensity. This remarkable ability to adapt to changes in intensity larger than those handled by any other known sensory system is mediated by a variety of gain-control mechanisms, ranging from adaptation by the photoreceptors themselves over 4 or 5 decades to adaptation at the neuronal-network level (for reviews, see [15, 16, 17]), that operate over disparate time scales, ranging from less than a second to several minutes.

Separate channels have evolved to handle rod and cone vision, since these photoreceptors operate under such drastically different conditions and have vastly different requirements. The lower limit of light sensitivity is set by the **dark light level**: the signal produced by spontaneous isomerizations of the rhodopsin molecule that mediates the phototransduction process. These spurious events produce responses in the

[1] The energy-flux density at the earth's surface for a dark night sky—the lower limit of scotopic vision—is about 10^{-14}W/cm^2. At high noon—the upper limit of photopic vision—it is set by the solar constant of 0.14W/cm^2.

rod that are identical to a photon response at a rate of one every 160 seconds [13]—equivalent to a light flux of 0.003 photons per integration time per rod. By pooling signals to detect the coincident absorption of several photons in nearby rods, the retina achieves subthreshold sensitivity, or **hypersensitivity**. Thus, it is able to detect the concurrent absorption of 10 photons in a pool of 5000 rods reliably [18]—a light flux of only 0.002 photons per integration time per rod!

Cones, on the other hand, operate with flux levels of kilophotons per integration time per receptor, so there is no need for the retina to sacrifice spatial acuity for detectability by pooling the cone signals. The threshold of the cone is set by quantum fluctuations in the photon flux, or **shot noise**: the change in mean light level must be large enough to exceed these fluctuations. Under ideal conditions, humans can detect an 0.5-percent change in mean intensity [19]. The absence of pooling in the cones also explains why the highest spatial acuity is acheived for cones. Humans can resolve 1 arc minute (1/60 degree) at the **fovea**, a small specialized region in the center of the retina where the cones are extremely small and are packed densely—rods are completely excluded from this region.

Needless to say, the vertebrate retina has evolved elaborate network-control mechanisms to fine tune the degree of pooling and the integration time, and to share elements between the rod and cone systems [20].

Our eyes and our brains are also sensitive to temporal changes in the image. The integration time of the primate visual system—which limits the temporal acuity of perception—is on the order of 0.1 second [21]. Thus, images flashed at rates of 50 frames per second or higher appear stable—the basis for our perceiving movies and television as changing smoothly. Yet we can discriminate differences in timing of much less than 0.1 second, because the retina displays **temporal hyperacuity**. Psychophysicists have shown that humans can discriminate reliably the order of onset of two small lines at the 3- to 5-msec level [22, 23]—20 to 30 times shorter than the integration time! Temporal hyperacuity is acheived only if the spatial separation between the two lines is in the range of 2 to 6 minutes. The lower spatial acuity achieved for this task reveals the presence of another channel—other than the one

that subserves spatial acuity—specialized for temporal resolution.

The presence of two channels specialized for temporal and spatial resolution has been confirmed both physiologically and anatomically [24, 25, 26]. Specialization is necessary due to physical laws which dictate that the retina—and any sensory system, for that matter—must integrate over a fixed volume of space–time to achieve a certain signal-to-noise level. When second- and third-order neurons pool receptor signals over a large area, they average out quantum fluctuations and can operate with a shorter integration time without sacrificing contrast sensitivity. [2] This specialization is carried still further in the retina; the signal space is also divided up along the spectral dimension.

Our eyes are sensitive to the wavelength of light, due to the presence of three different types of cone pigments, with peak absorbances at 420 nm (appears violet), 530 nm (appears yellowish green), and 560 nm (yellow). In comparison, rods are tuned to 500 nm (blue–green). These three cone types give us the ability to discriminate wavelengths ranging from 400 nm (violet) to 670 nm (red). In vertebrates, signals from these three cones—incorrectly called blue (instead of violet), green (instead of yellowish green), and red (instead of yellow)—are transmitted by two channels that carry R − G (red minus green) and B − (R + G) (blue minus yellow). This transformation produces a more efficient encoding since the R and G signals are similar, and hence are largely redundant: their difference is close to 0 over most of the retina, and their sum can be sampled less frequently.

All this specialization makes for a total of five specialized channels: one milliphoton-sensitivity channel for night vision, one minute-acuity channel for luminance, one millisecond-acuity channel for motion, and two nanometer-acuity channels for chrominance. In addition, there are several other more specialized pathways. Among other functions, they mediate our closed-loop optokinetic response reflex, which minimizes image slippage on the retina; our pupillary response, which regulates the amount of light entering the eye; our lens accommodation, which focuses the image on the retina; the gain of our open-loop vestibular-occular reflex, which moves the eyes to

[2] It makes no difference whether the quanta are photons, vesicles, or ion channels.

compensate for head movements; and the rate of our biological clocks, to keep them in phase with the day–night cycle.

Complicating matters even more, activity in each channel is encoded with two complementary streams: one that signals increases in amplitude by increasing the amount of neurotransmitter released and the spike-discharge rate (ON pathway), and another that signals decreases in amplitude in a similar fashion (OFF pathway). Thus, both polarities of change are signaled by high neurotransmitter-release rates and high spike-discharge rates, but there is little or no activity in steady state. Complementary signaling is used throughout the retina, except for at the very first synapses—found in the rod and cone terminals. These sites are the only places where the retina maintains elevated neurotransmitter-release rates to signal both increases and decreases using a single stream.

Complementary signaling using the ON and OFF pathways compensates for the inherent shortcomings of neural systems in three ways: (1) noise due to fluctuations in vesicular neurotransmitter-release rates and spike rates decreases in inverse proportion to the square root of the number of vesicles or spikes used to signal; (2) slow repolarization of the synaptic membrane and removal of neurotransmitter from the synaptic cleft does not limit transmission speed; and (3) the energy required to replenish neurotransmitter supplies or to generate a spike is conserved in steady state, when nothing is happening.

The existence of all these dedicated channels and complementary signaling makes the retina a complex, multifaceted structure. In this review, I emphasis those facets of the retina that shed light on how spatiotemporal signals are processed and how motion is encoded. Retinal neurons are sensitive to spatiotemporal changes, in general, and to the speed and the direction of motion, in particular. Nature has evolved several different forms of eyes (e.g., compound versus camera eyes); I limit my discussion to the literature for vertebrate eyes. Given the vagaries of experimental work and history, the majority of the relevant work has been carried out in a few species—in particular, in the tiger salamander, mudpuppy, catfish, skate, rabbit, and cat.

I begin this review by summarizing several salient points concerning retinal neu-

roanatomy.

2.1 Cell Classes

Synaptic interactions in the vertebrate retina occur within two plexuses.[3] The more peripheral one is called the **outer plexiform layer** (OPL), and the more central one is called the **inner plexiform layer** (IPL). The neurons that interact in these plexuses fall into two distinct groups: relay neurons and interneurons. A **relay neuron** receives input in one plexus and delivers its output to another. An **interneuron** remains entirely within a single plexus. The input and output neurons of the retina are specialized relay neurons that sense incident light and that send signals to the rest of the brain, respectively.

There is one class of interneuron for each plexiform layer in the retina, and there are three classes of relay neurons to feed signals from the input to the OPL, from the OPL to the IPL, and from the IPL to the rest of the brain. That makes five topologically defined cell classes.[4] However, this feedforward cascade is interrupted by the presence of a sixth class of cells; these cells pick up signals in the IPL, and transmit these signals back to the OPL.

These six cell classes form a highly regular and densely connected topological network, as shown in the vertical cross-section through the human retina in Figure 2.1. The names and roles of the cell classes follow:

- **Photoreceptors** are the input neurons; they transduce incident light into electrical signals that drive the OPL. In almost all vertebrates, they come in two functionally and morphologically distinct types, called **rods** and **cones** due to their shape. Their cell bodies lie in the **outer nuclear layer** (ONL), above the OPL.

[3]The Latin *plexus* means braid; it is used to describe a complexly interconnected arrangement of parts.

[4]My convention is to use *class* to refer categories that are based on topology, and to use *type* to refer to subcategories that are based on morphology, neurochemistry, or physiology.

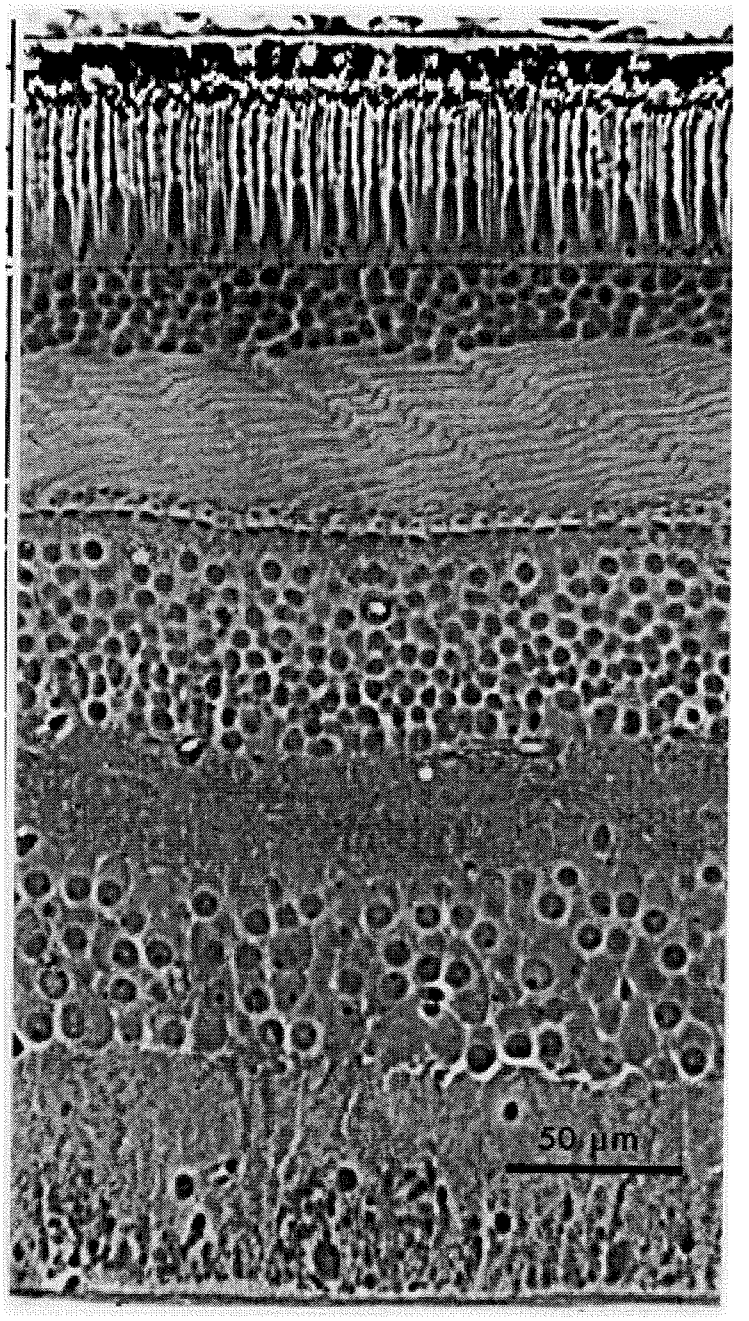


Figure 2.1: Vertical Section Through the Human Retina
The retina is a thin sheet of brain tissue, less than 0.5mm thick, that lines the inside of the orb of the eye. The photoreceptors sit against the eyewall, and light must travel upward, through the entire thickness of the retina, to strike them. Visual information flows downward, passing through at least three different cell types. This gross anatomy is preserved in all vertebrate retinae. In the region shown—which is about 1.25mm away from the center of the fovea—the cone, rod, and ganglion cell densisties are 15,000mm^{-2}, 75,000mm^{-2}, and 40,000mm^{-2}, respectively. Reproduced from [27].

- **Horizontal cells** are the interneurons in the OPL; they play an inhibitory role. Their cell bodies lie in the **inner nuclear layer** (INL), just beneath the OPL. In cold-blooded vertebrates, such as the turtle and the carp, there are three different types with different color selectivities, but such is not the case in primates. Primates have only two horizontal-cell types with no color selectivity whatsoever.

- **Bipolar cells** are the relay neurons between the OPL and the IPL. Their cell bodies lie in the middle of the INL; a single dendritic shaft emerges on the OPL side, and an axonal one emerges on the IPL side, giving them a distinct bipolar structure.

- **Amacrine cells** are the interneurons in the IPL. Amacrine cells are generally believed to be inhibitory, but one type has been shown to use two types of neurotransmitter, one excitatory and the other inhibitory [28]. Amacrine cells are found in the INL, just above the IPL, and in the **ganglion cell layer** (GCL), just below the IPL.

- **Interplexiform cells** form a second class of relay neurons that provide a feedback path from the IPL to the OPL. These cells are found among the amacrine cells in the INL; indeed, certain authors consider them a subgroup of amacrine cells. They modulate synaptic interactions in the OPL and IPL (for reviews, see [29, 30]), and thereby reorganize the retinal microcircuitry to optimize performance in sunlight, moonlight, and starlight [20]. I will not consider them further.

- **Retinal ganglion cells** are the sole output channel for the retina. In primates, about 1 million ganglion-cell axons make up the optic nerve that projects to the brain proper. These cells communicate by sending trains of impulses down their axons. In contrast, the majority of other retinal neurons signal using graded changes in the membrane potential, rather than all-or-none pulses.

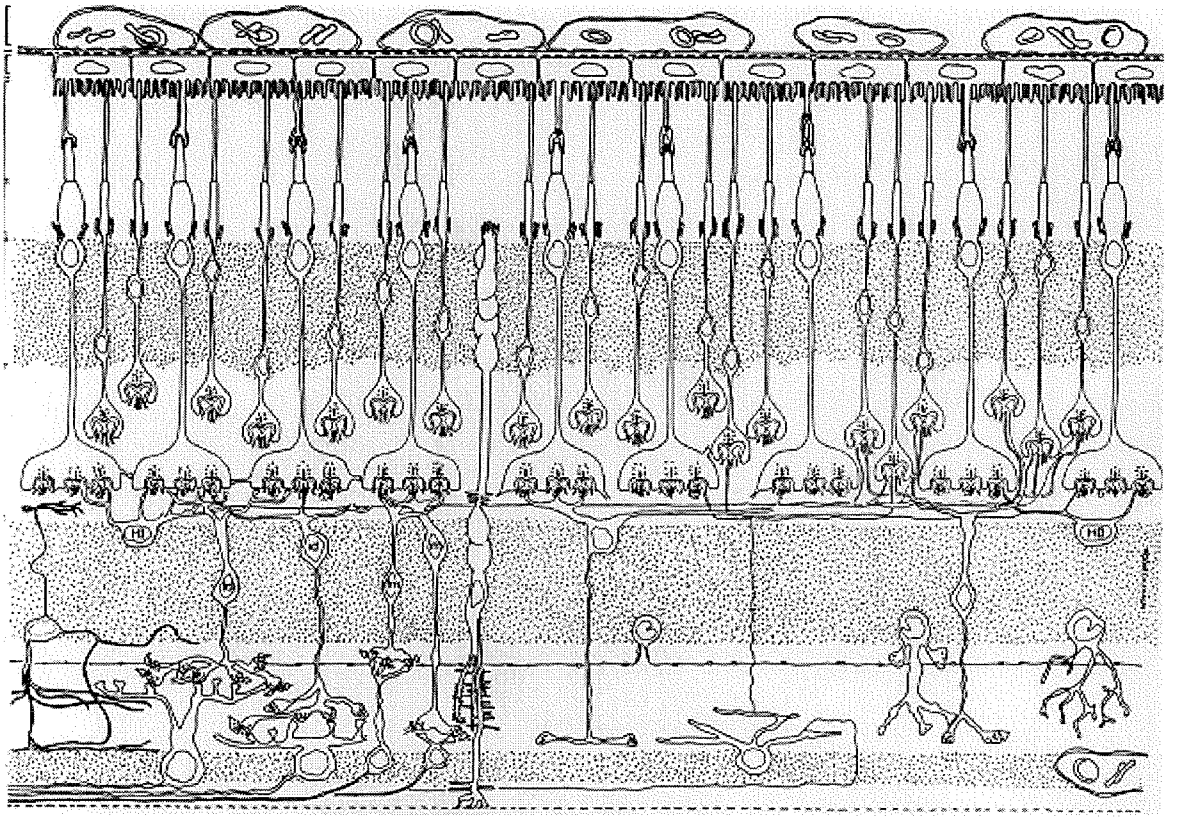


Figure 2.2: SCHEMATIC DIAGRAM OF A TYPICAL VERTEBRATE RETINA
The photoreceptor terminals are enlarged to show the details of the invaginations. The retina can be described to first order as a three-layer, feedforward neural network, with the first level of interconnections making up the outer plexiform layer (**OPL**), and the second level of connections making up the inner plexiform layer (**IPL**). Reproduced from [31].

2.2 The Outer Plexiform Layer

Two morphological types of horizontal cells—the interneurons of the OPL—are readily distinguished in all vertebrate retinae: one has a short axon and spatially segregated dendritic and axonal fields, whereas the other is axonless. In humans and our warm-blooded relatives, the short-axon cells occur in just one variety, labeled **HI** in Figure 2.2, and their axons branch heavily and contact rods, exclusively, whereas their dendrites branch heavily and contact cones, exclusively [32]. The axon connecting these two fields is so long and thin that the fields do not interact electrically [33]. The dendritic field of the axonless cell, labeled **HII** in Figure 2.2, also

contacts cones exclusively. In cold-blooded vertebrates, such as fish, the short-axon cells occur in three varieties, called **H1, H2,** and **H3** [34], that make selective contact with cones [35]. Unlike the primate cells, these cells show color-selective physiological responses [36], and their axons dive down through the INL and terminate among the amacrine cells at the INL–IPL border [34]. Another difference is that the axonless cells of cold-blooded species contact only rods [37].

At least five morphological types of bipolar cells occur in the vertebrate retina, if we do not distinguish the color selectivity of the cones that they contact. The most prominent distinguishing feature of bipolar cells is the sizes of their dendritic or axonal fields. There are two distinct clusters: cells with large dendritic and axonal fields, called **diffuse,** and cells with small fields, called **midgets** [38, 31]. If we take into account whether they contact rods or cones, we find that there is class of large-field cells that contacts only rods [37]. Now, paying close attention to the type of contacts made with receptors, we find that some bipolars enter an invagination in the rod and cone terminals, whereas others make contact at the base of the terminal [39, 32]. An example of each type is shown in Figure 2.2. Note that basal contacts never occur on rods; thus, rods have only one type of bipolar (diffuse and invaginating), whereas cones have four types (midget or diffuse, and flat or invaginating).

Horizontal cells also enter the invaginations in the rod and cone terminals, forming a synaptic complex called a **triad** [32]. Bipolars terminals are always the central element in the triad, with a horizontal-cell process on either side, for a total of three postsynaptic processes. The horizontal-cell processes reach deeper into the cone invaginations and sometimes block the bipolar terminal. On top of the invagination, there is a flattened oblong structure, called a **ribbon**, surrounded by small round pellets that contain neurotransmitter, called **vesicles**. This arrangement is called a **ribbon synapse** and is always associated with the triad (see Figure 2.2). The exact function of this synaptic specialization is still a mystery; the ribbon could facilitate vesicular transport [40] and the invagination might influence the diffusion of neurotransmitter, shaping the concentration profile and controlling the amount of neurotransmitter that reaches the elements of the triad and the basal synapses [41].

The inhibitory action of horizontal cells on cones probably occurs in these invaginations.

In addition to triad synapses and basal junctions, a third type of synaptic structure occurs in the OPL [42]. This structure, which passes ionic currents, is called a **gap junction,** or sometimes an **electrical synapse,** and can be thought of as a low-resistance pathway connecting the two cells. There is an extensive network of gap junctions between all four classes of cells in the OPL: cone to cone, rod to rod, horizontal to horizontal, and bipolar to bipolar (see [43] for a review). In addition, there are gap junctions between rods and cones (see Figure 2.2).

These synaptic interactions in the OPL gives rise to the antagonistic center-surround receptive-field organization first observed by Kuffler in the early fifties, when he recorded spike trains from retinal ganglion cells of the cat [44]. About 1 decade later, Rodieck demonstrated quantitatively that the spatial profiles of the center and surround components of the receptive field are well fitted by Gaussians, and he proposed the highly influential **difference of Gaussians** (DOG) model of spatial filtering in the retina [45].

2.3 The Inner Plexiform Layer

The IPL is over five times thicker than the OPL (see Figure 2.1), and its anatomy and physiology are much less well understood. A plethora of amacrine cell types—the interneurons of the IPL—has been described. So far, 29 types have been identified in the turtle [46, 47], and several dozen have been found in the roach [48, 49]; researchers estimate that there are over 40 types in the mammalian retina [50]. But the purpose of all this diversity remains a mystery, since it is not reflected in the responses that electrophysiologists have recorded from the amacrine cells [46]. Either the physiologists are using stimuli that are too simplistic to discriminate among the morphologically defined amacrine-cell classes, or the anatomists are assigning undue significance to morphological differences that are of little or no consequence. The same situation holds for ganglion cells [51, 52, 53]. There is no doubt, however, that

amacrine cells play a critical role in detecting the speed and direction of motion, since antagonists of known amacrine-cell neurotransmitters abolish direction selectivity and other complex properties of rabbit ganglion cells [54].

Over the past decade, however, researchers using immunohistological markers and improved intracellular recording techiniques have begun to demonstrate a bewildering diversity in functional properties that is correlated with the structural diversity. These findings have forced neurobiologists to reaccess the century-old work of Ramón y Cajal, the preemminent neuroanatomist of his time. Cajal advocated a five-tiered stratification of the IPL—a structural abstraction that has stood the test of time [37]. Each sublayer is demarcated by levels of dendritic arborization. Cajal referred to these levels as simply the first through fifth **strata**, starting at the most peripheral one; they are denoted S1 through S5. Strata are prominent in birds and reptiles, but are difficult to distinguish in fishes and mammals [37, 55], because the latter's amacrine cells have more diffuse arborizations. For this reason, anatomists sometimes simply divide the IPL into five sublayers of equal thickness [56].

Functional correlates underlying the stratification of the IPL have been found recently [49]. Monostratified amacrine cells ramifying exclusively in S1 have sustained OFF responses; that is, lightoff elicits an increase in membrane potential that is maintained for the duration of the stimulus. Similarly, amacrines with arborizations in S4 and S5 have sustained ON responses. Amacrines with small, bistratified, tristratified, or diffuse dendritic fields spanning S2, S3, and S4 have slow-decaying transient ON–OFF responses; that is, both lightoff and lighton initially elicit an increase in membrane potential. Monostratified amacrine cells with large dendritic fields in S2 and S3, most often located close to the S2–S3 border, have fast-decaying, transient ON–OFF responses; the wide-field amacrine cell shown in Figure 2.2 is an example of this type.

Although this elegant organization of structure and function in the IPL—which was first described in fishes—has yet to be demonstrated in mammals, it has been shown that cat ganglion cells with processes confined to S1 and S2 have OFF center responses, whereas those with processes confined to S3, S4, and S5 have ON center

responses [57, 58]; the same is true in fish [59]. This observation is the basis for coarser subdivision of the IPL into just two **sublaminae**, originally named a and b, but now commonly called the OFF sublamina and the ON sublamina, respectively.

Cone and rod bipolar cells fit neatly into this picture, as shown in Figure 2.2. Bipolars that make invaginating contacts with cones terminate in S3 and S4 (ON sublamina), whereas those making flat contacts with cones terminate in S1 and S2 (OFF sublamina) [31]. Rod bipolars, which are of only the invaginating kind, terminate in the ON sublamina, but their processes, which are found in S5, are segregated from those of the invaginating cone bipolars, which are in S3 and S4. Since cones and rods both have an OFF light response, the difference in bipolar cell response polarities implies that the basal and invaginating synaptic contacts act differently.

Both photoreceptors use glutamate, a neurotransmitter that opens sodium channels, causing current to flow into the postsynaptic cell, and thus depolarizing the cell. However, glutamate can also act through a second messenger pathway that closes sodium channels, reducing the current flowing into the postsynaptic cell, and thus hyperpolarizing the cell. The photoreceptors release glutamate when they are depolarized (lightoff), so the former synaptic action is sign preserving (excitatory), and the latter action is sign reversing (inhibitory). The net effect is sodium channels are opened at the invaginating contacts when light increases (ON pathway) and are opened at basal contacts when light decreases (OFF pathway). With a few exceptions (see [60]), this correlation between structure and function is generally true.

Complementary signaling arises in the rod pathway at the IPL, where a narrow-field, bistratified amacrine cell, labeled AII in Figure 2.2, relays the rod signal to the same set of ganglion cells that carry cone signals. Rod bipolars make contact with AII in S5. AII makes a gap junction onto cone bipolar terminals in the ON sublamina, and makes an inhibitory synapse directly onto ganglion cells in the OFF sublamina. This microcircuit has been identified in the cat [61, 62] and the rabbit [63].

Bipolar cells form a synaptic complex in the IPL that is similar to the triad formed by rods and cones, except that there is no invagination present and only two postsynaptic processes occur; hence, it is called a **dyad** [64]. At least one of the

processes always belongs to an amacrine cell; the other may arise from an amacrine cell or a ganglion cell. Usually, one of the postsynaptic processes makes a synapse back onto the bipolar process in close vicinity to the dyad; this process invariably belongs to an amacrine cell [62]. That way, the amacrine cell can inhibit the bipolar-cell terminal, and can terminate the release of neurotransmitter. This **presynaptic inhibition** results in a transient response in the postsynaptic cells, although the bipolar cell receives sustained inputs in the OPL [65, 66].

Another common synaptic structure consists of a chain of two or three conventional synapses, called **serial synapses** [67]. The first synapse is made by an amacrine cell onto another amacrine-cell process, which in turn makes a second synapse onto a nearby ganglion-cell dendrite, onto a bipolar-cell terminal, or onto yet another amacrine-cell process. This arrangement could achieve a net excitatory effect in the third-order cell via two inhibitory synapses.

2.4 Types of Ganglion Cells

The existence of several classes of ganglion cells gives rise to parallel visual pathways [68]. These pathways originate in the OPL, starting with the five types of bipolar cells. The processes of these bipolars segregate into five strata in the IPL, and they drive five types of ganglion cells; the processes of these ganglion cells costratify with the bipolar arborizations. The five types of ganglion cells are called: ON- and OFF-center midgets, ON- and OFF-center parasols, and rod-system ganglion cells. Needless to say, **midget ganglion cells** talk to midget bipolars in S1 and S4; **parasol ganglion cells** talk to diffuse bipolars in S2 and S3; and **rod ganglion cells** talk to rod bipolars in S5 (Figure 2.2). The stylized circuit diagram in Figure 2.3 shows the major synaptic interactions involved in these pathways.

Several different classifications of ganglion cells are in common usage. There is the **beta** (β), **alpha** (α), **gamma** (γ) classification [25] preferred by retinal anatomists; it is based on the cell morphology. Thus, β-cells have medium-sized cell bodies and small dendritic trees, and are synonymous with the aforementioned midgets; α-

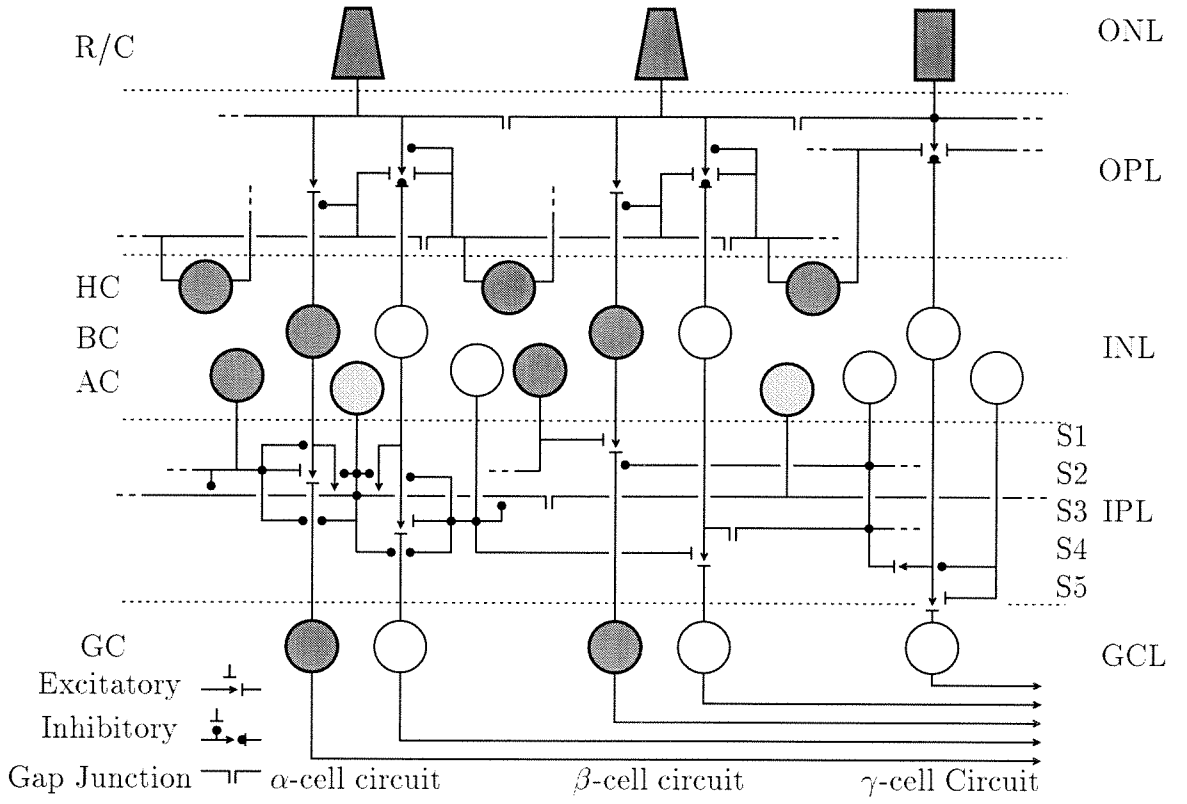


Figure 2.3: Simplified Circuit Diagram of the Retina

Three modules subserving motion (α), spatial color vision (β), and nocturnal vision (γ) are shown. The dark-shaded cells are turned off by light (OFF stream), whereas the unshaded cells are turned on by light (ON stream). The light-shaded cells turn on transiently at both lighton and lightoff. The rods and cones (R/C) are turned off by light; they make an inhibitory synapse onto ON bipolar cells (BC), and an excitatory synapse onto OFF bipolars in the outer plexiform layer (OPL). The horizontal cells (HC) are excited by the receptors, and inhibit the receptors and the OFF bipolars. Rods are coupled by electrical junctions, and cones are coupled as well; there are also gap junctions between rods and cones. The inner plexiform layer (IPL) has five sublayers (S1–S5). S1 and S4 serve the OFF and ON streams of the β circuit, respectively; S2 and S3 serve the OFF and ON streams of the α circuit, respectively; and S5 serves the rod circuit that has only an ON stream. Rod signals are carried by the γ ganglion cells and are also relayed to the β (and α) circuit(s) by a bistratified amacrine cell (type AII) that makes an inhibitory synapse onto the OFF ganglion cell in S1 and a gap junction onto the ON bipolar terminal in S4. The rod bipolar excites another set of amacrine cells (types A13 and A17) that makes inhibitory reciprocal synapses back onto it. The β ganglion cells are driven mainly by bipolars, whereas the αs are driven mainly by amacrines. There are complex interactions with amacrines in the α circuit: The bipolars excite both wide-field (type A19) and narrow-field amacrines (type A2–A3). These amacrine cells make reciprocal synapses onto the bipolars and also feed forward onto the ganglion cells. In addition, the narrow-field amacrine cell inhibits the wide-field cell.

cells have huge cell bodies and large dendritic trees, and are synonymous with the aforementioned parasols; and γ-cells generally have tiny cell bodies and large dendritic trees, and form a functionally heterogenous group that includes the aforementioned rod ganglion cell. The 1 percent of cat ganglion cells that are direction selective fall into this category [52].

There is also the **magno–parvo** classification favored by cortical primate physiologists and anatomists [26]; it is based on the morphology of the target cells in the lateral geniculate nucleus (LGN), where most ganglion-cell axons terminate:[5] Cell bodies in the LGN are arranged in six layers, called **laminae**. Four of these laminae contain small cell bodies, and the remaining two contain large cell bodies—hence, the magnocellular–parvocellular terminology. Naturally, the α cells terminate in the magnocellular layer, whereas the β cells terminate in the parvocellular layer. The γ cells are not included in this classification, since they generally project to the superior colliculus and to the brainstem. The magno–parvo pathways primarily apply to the primate visual system, where they have been traced from the retina deep into cortex [69].

There is yet another classification—**X**, **Y**, **W**—favored by retinal physiologists, and originally elaborated in the cat [24, 70]. This scheme is based on whether a cell responds to spatiotemporal patterns linearly (X) or nonlinearly (Y). To test for linearity, Enroth-Cugell and Robson used an odd symmetric pattern and modulated the luminance of one half with a square wave and the luminance of the other half with the inverse of the square wave [24]. Thus, while the luminance in one half increased, the luminace in the other half decreased at the same rate, keeping the total luminance constant. They demonstrated that some cells did not respond at all when this stimulus was centered perfectly on the receptive field, showing a null response, whereas others responded strongly to each transition in the square wave, showing frequency doubling. They named these cell types X and Y, respectively. Obviously, this linearity test is generally applicable to only those cells with circularly

[5]The LGN is part of the thalamus, a region in the forebrain that serves as a relay station for sensory information bound for cortex.

symmetric receptive fields—hence, the need for a third category (W) to account for the nonconcentric cells.

It later became clear that the X and Y classes, as defined by the linearity test, are not homogeneous [71]. Each category has at least two subgroups characterized by significant differences in axonal conduction velocities. This subspecialization is taken into account by the **brisk** (high- or medium-velocity) and **sluggish** (low-velocity) classification proposed later by Cleland and Levick [51]. All brisk Y cells have high conduction velocity, and all brisk X cells have medium conduction velocity, and all W cells—defined here to be synonymous with the nonconcentric units—have low conduction velocity [71]. Since conduction velocity is correlated with cell-body size, it was not surprising to find that brisk Y cells are synonymous with α cells, and that brisk X cells are synonymous with β cells, whereas sluggish X, sluggish Y, and W cells fall in to the heterogeneuos γ class [53, 51].

All three classifications were developed in mammals (mainly the cat and the monkey) and have not been applied successfully to lower species (e.g., the mudpuppy, salamander, and the frog), since little is known about central structures in these species and their ganglion cells show much richer specializations [72].

Of course, these classification schemes are nonexclusive; actually, they are redundant. In other words, looking at the same cell, an anatomist will identify it as a β cell, a cortex expert will say it is in the parvo pathway, and a retinal physiologist will find that it shows X-type behavior. However, the geniculate-based classification is limited to cells that project there, and hence it can account only for the retinally-labeled α and β cells, which are the predominant input (but see [68, 73]). And the physiological classifications are notorious for lumping different cell types into the same group, because tests for linearity, ON–OFF behavior, or sustained-transient behavior, alone, are too simplistic to discriminate the cells' specializations. For these reasons, I shall use the α, β, γ classification in this dissertation, except where history precludes my doing so.

Physiologically, β cells always respond steadily and have small receptive fields, whereas α cells always respond more transiently and have larger receptive fields.

There is a continuum of receptive-field sizes and temporal characteristics within each class, ranging from small and sustained in the fovea to large and transient in the periphery [51]. Nevertheless, α cells are always two to three times larger and respond more transiently than β cells at the same eccentricity [51]. The anatomy of these cells is well correlated with the physiology. In one study where ON cells in the *area centralis* of the cat retina were reconstructed from electron-microscope pictures, β cells received 72 percent of their inputs from bipolar cells [74], whereas α cells received 85 percent of their inputs from amacrine cells [75]. This distribution fits well with the notion that bipolar cells have a sustained response, whereas amacrine cells are responsible for generating a transient response. Another study looked at OFF cells from the periphery of the cat retina, and found that β cells received 38 percent of their synapses from bipolar cells, whereas α cells received only 20 percent of their synapses from bipolar cells [76]. Again, the cell with the more sustained response receives more bipolar input, although these peripherally located cells are predominantly amacrine driven. The dendritic-field sizes of these ganglion cells are also well correlated with the receptive-field sizes.

In terms of actual numbers and sampling densities, there are about 150,000 to 200,000 ganglion cells in the cat retina [77], versus 1.5 to 1.8 million in the macaque [78]. The α and β cells constitute 50 percent of the cat cells and 90 percent of the monkey cells, and there is about a 9:1 ratio of β cells to α cells in both cases. The sparsity of α cells reflects that nine times fewer α cells are required to tile the retina, since the α dendritic fields are three times larger [68], compared to the β cells. The remaining 50 percent of the cells in the cat, and the remaining 10 percent of the cells in the macaque (these percentages work out to the same absolute amount of about 100,000 in each retina) fall into the heterogeneous γ class. In both species, β cells project to the forebrain. In the cat, α cells project to the forebrain and the midbrain—in particular, to the superior colliculus and the pretectum—whereas α cells in the primate project to only the LGN [79, 68]. Finally, γ cells, which carry more specialized information (e.g., they encode the direction of motion) project predominantly to the midbrain [79].

2.5 Summary

Even this extremely truncated and oversimplified review of retinal neurobiology makes it abundantly clear that the retina is, indeed, much more complex than is any sensory system currently built by engineers. The retina's parallel dedicated channels make it akin to several specialized cameras coexisting on the same chip. Even if we try to get around this multifaceted character by focusing on just one of these cameras, we are still bewildered because the elements of the cameras are richly interconnected, and the same element may serve several purposes at the same time, or it may be coopted by different cameras at different times.

This nonmodularity, which is a defining characteristic of the retina—and of the rest of the brain—makes it extremely difficult for us to understand how the system operates by using traditional reductionist approaches. It is the goal of this thesis to provide a unifying framework that accounts for the following key aspects of retinal organization, which are preserved across a large variety of species:

- The retina encodes several parallel information streams in its output that emphasize different aspects of a scene, such as color, edges, and movement.
- To a good first approximation, the retina in all vertebrate species can be described as a locally connected feedforward neural network with three cellular layers that are connected by two layers of processing: the outer plexiform layer (OPL) and the inner plexiform layer (IPL).

Chapter 3 Retinal Function: Information Encoding

The retina converts continuous spatiotemporal patterns of incident light into spike trains. Transmitted over the optic nerve, these discrete spikes are converted back into continuous signals by dendritic integration of excitatory postsynaptic potentials in the lateral geniculate nucleus of the thalamus. For human vision, contrast thresholds of less than 1%, processing speeds of about 20 ms per stage, and temporal resolution in the submillisecond range are achieved, with spike rates as low as a few hundred per second. No more than 10 spikes, per input, are available during this time. The retina must maximize the amount of information carried by these spikes.

For optimum performance, the retina must efficiently encode stimuli generated by all kinds of events, over a large range of lighting conditions and stimulus velocities.

These events fall into three broad classes, listed in order of decreasing probability of occurrence:

- **Static events:** Generate stable, long-lived stimuli; examples are buildings or trees in the backdrop
- **Punctuated events:** Generate brief, short-lived stimuli; examples are a door opening, a light turning on, or a saccade
- **Dynamic events:** Generate time-varying, ongoing stimuli; examples are a wheel spinning, grass vibrating in the wind, or eyes panning the scene

In the absence of any preprocessing, the output activity mirrors the input directly. Changes in lighting, which influence large areas, are reflected directly in the output of every single pixel in the region affected. Static events, such as a stable background, generate persistent activity in a large fraction of the output cells, which transmit

the same information repeatedly. Punctuated events generate little activity and are transmitted without any urgency. Dynamic events generate activity over areas far out of proportion to informative features in the stimulus, when the stimulus sweeps rapidly across a large region of the retina. Clearly, these output signals are highly correlated, over time and space, resulting in a high degree of redundancy. Hence, reporting the raw intensity values makes poor use of the limited throughput of the optic nerve.

3.1 Optimal Filtering

Barlow observed over 30 years ago that it would be most efficient for the retina to use the fewest spikes to transmit the most commmonly occuring patterns [80]. Since then, vision researchers have succeeded in formalizing this efficient-encoding hypothesis, and have made quantitative predictions that are in good agreement with selected experimental observations [81, 82, 83]. In this section, I adopt this formalism and derive optimal filters for the retina, using results from communication theory.

First, I present abbreviated derivations of results that we shall need from information theory. My main goals are to define terms and to point out the assumptions that underpin these results, rather than to achieve mathematical rigor.

The amount of **information** transmitted by a communication channel is defined as the amount by which that channel's output, Y, reduces our uncertainity about its input, X :

$$I(X;Y) \equiv H(X) - H(X|Y),$$

where $H(X)$ is the entropy of X and $H(X|Y)$ is the entropy of X given Y[8]. The quantity $I(X;Y)$ defined above is also known as the **mutual information** between X and Y. In the special case where the channel simply adds noise, $\mathcal{N}$, the information transmitted is simply

$$I(X;Y) \equiv H(Y) - H(\mathcal{N}). \tag{3.1}$$

The **capacity** of a channel is defined as the maximum rate at which that channel

can transmit information. Equation 3.1 shows that the rate is maximized when the entropy of the transmitted signal is maximized. If the transmitted signal is subject to an average power constraint, P, then the ensemble with maximum entropy is the Gaussian distribution with variance $\sigma^2 = P + N$, when the noise also is assumed to be Gaussian, with variance N. If the signal is bandlimited to W, it can produce only $2W$ nonredundant symbols per second. This conclusion follows from Nyquist's (and Shannon's) sampling theorem. And the information carried by each symbol is

$$I_{\text{symb}} = \frac{1}{2}\log_2(2\pi e(P+N)) - \frac{1}{2}\log_2(2\pi eN) = \frac{1}{2}\log_2\left(1+\frac{P}{N}\right),$$

as a Gaussian distribution with variance σ^2 has entropy $\log_2(2\pi e\sigma^2)/2$. Hence, the channel capacity is

$$C = W\log_2\left(1+\frac{P}{N}\right),$$

in bits per second (baud), a result that was first obtained by Shannon [8]. Notice that the information rate is linearly proportional to the bandwidth, since every sample is independent, but is only logarithmic in the signal-to-noise ratio (SNR), since it takes only b bits to specify one of 2^b possibilities.

We can extend this result easily to obtain the information transmitted through a channel, given the expected power spectra of the transmitted signal and the noise. We chop up the frequency spectrum into small bands, assign the right amount of signal and noise power to each band, and assume that the amplitude distribution in each band is Gaussian. This procedure yields

$$I = \int_0^W \log_2\left(1+\frac{S_0(f)}{N_0(f)}\right) df, \tag{3.2}$$

where $S_0(f)$ and $N_0(f)$ are the expected power spectral densities of the signal and the noise, respectively. In general, the frequency f is a vector in three-dimensional space–time, (i.e., $f = (f_x, f_y, f_t)$). For concreteness, we can think of f as spatial frequency, without loss of generality.

I now address the efficient-encoding problem. Following a strategy similar to that

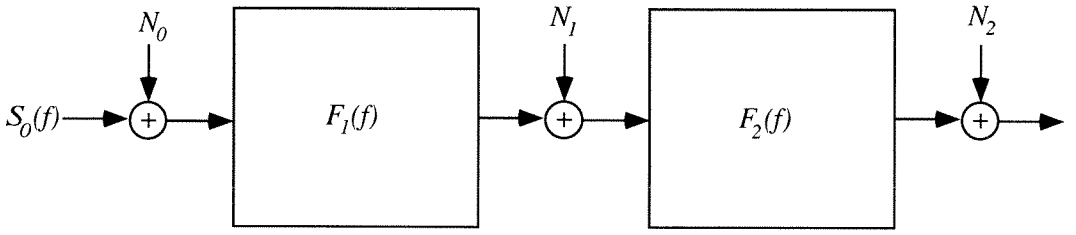


Figure 3.1: Optimal Filter Design
A filter cascade consisting of a noise-suppression filter, $F_1(f)$, and a whitening filter, $F_2(f)$, is used. The noise added by the input and output channels, and by the intermediate channel between the filters, also is included in this model.

of Attick and Redlich [81], I design a filter cascade to transmit the photoreceptor signal optimally; it consists of two filters, as shown in Figure 3.1.

The first filter's job is to ensure that as much as possible of the channel capacity is used by the signal, and that as little as possible is used by the input noise. To acheive these goals, I amplify or attenuate each frequency band in accordance with the SNR and with the absolute noise level in that band. The filter should attenuate bands where the noise is larger than the signal; it is wasteful to transmit these bands because the noise takes up most of the channel capacity. As information is logarithmic in SNR, bands with vastly different SNRs contribute similar amounts of information—as long as the SNR is greater than unity. So we should not pick the band with the best SNR and reject the others, as we would do if we wanted to maximize the SNR. Instead, we should amplify bands where the noise is smaller than the channel noise, so as to preserve the SNR—but only if the SNR exceeds unity.

Once we have suppressed the noise, our next optimization is to redistribute the energy of the signal across frequency bands to make the most efficient use of the limited power of the transmitted signal. The second filter achieves this goal by trading SNR for bandwidth. The large signals required to obtain a high SNR contribute linearly to power, but contribute only logaritmically to information rate. Spreading out the energy of the transmitted signal in frequency increases the information rate linearly, because each frequency band provides independent information. Therefore,

if it is about information that we care, we are better off transmitting signals that have a low SNR but a high bandwidth.

By design, the whitening filter does not discriminate between signal and noise; it transmits information about all its inputs equally well. The noise-suppresion filter and the whitening filter exhibit an interesting complementarity: The former makes the signal less noisy, and the latter makes the signal more noiselike.

The compound effect of the noise-suppression filter and the whitening filter is illustrated in Figure 3.2 for a signal with a $1/f^2$ power spectrum. This distribution of spectral energy is typical of both the spatial and temporal frequency composition of natural scenes.

To find the optimal noise-suppression filter, we maximize the functional

$$\begin{aligned} E_1[F_1(f)] &= (1+B)\int_0^\infty \log_2\left(1+\frac{F_1(f)S_0(f)}{F_1(f)N_0+N_1}\right)df \\ &- \int_0^\infty \log_2\left(1+\frac{F_1(f)(S_0(f)+N_0)}{N_1}\right)df, \end{aligned} \tag{3.3}$$

where $F_1(f)$ is the power gain of the filter, and $S_0(f)$, N_0, and N_1 are the power spectral densities of the input signal, the noise in the input signal, and the noise added by the channel that transmits F_1's output signal (See Figure 3.1). The first term measures how much information is carried in the output signal; the second term measures how much capacity is required to transmit the signal as well as the noise. I have included a relative cost factor, B : the excess capacity that we are willing to use to transmit 1 baud of information about the signal. B is dimensionless because it is the ratio.

Taking the functional derivative and equating it to zero, we find that

$$F_1(f) = \frac{N_1}{N_0}\frac{BS_0(f)-N_0}{S_0(f)+N_0}. \tag{3.4}$$

When $BS_0(f) \gg N_0$ (high SNR), the second factor becomes unity, and filter's gain is set to amplify the input noise up to the noise level of the output channel; this boost prevents channel noise from degrading the SNR. For $S_0(f) < N_0/B$ (SNR less than

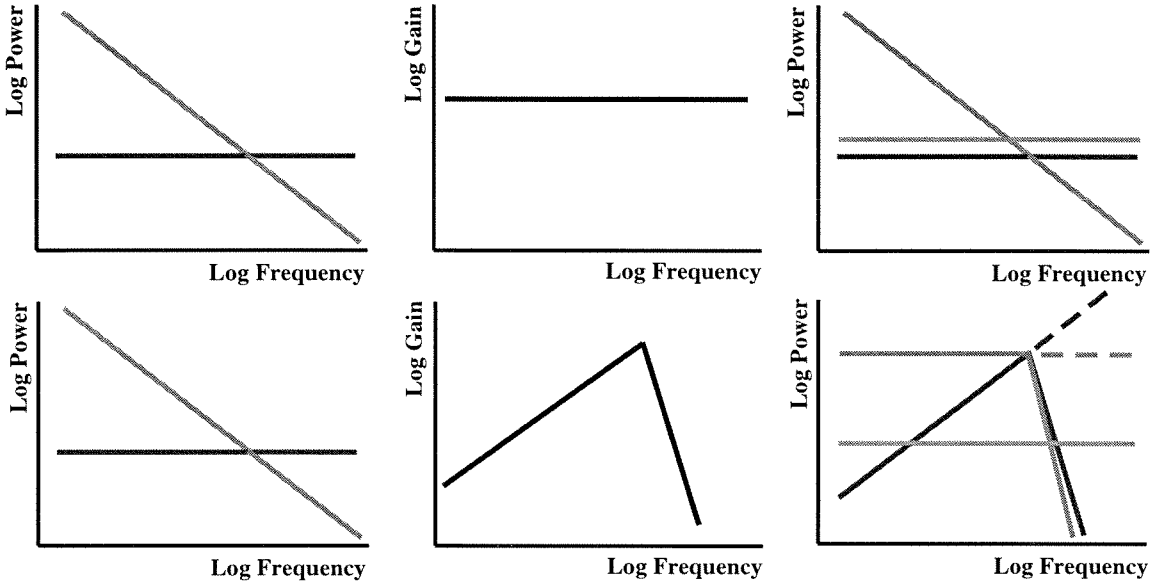


Figure 3.2: Effect of Optimal Filter

Filtering 1/f signals plus white noise (left column) using allpass or bandpass filters (middle column) prior to transmission of output signals (right column) down a communication channel that adds white noise (light grey line). Top Row: Transmitting the raw image signal allocates too much power to low frequencies, which do not carry any more information than do the other frequencies, and too much power to broadband noise, which dominates at high frequencies. Bottom Row: Bandpass filtering optimizes the use of power by attenuating low frequencies and rejecting wideband noise.

$1/B$), the power gain becomes negative. In this regime, more than B bits of noise are transmitted for each bit of signal, and the only way that we can satisfy the cost constraint is by making the power negative. Obviously, such values are not physically possible; the best that we can do is to set the gain to zero in these frequency bands.

To find the optimal whitening filter, we maximize the functional

$$\begin{aligned} E_2[F_2(f)] &= (P_2 + N_2) \int_0^\infty \log_2 \left(1 + \frac{F_2(f)(S_1(f) + N_1(f))}{N_2}\right) df \\ &- \int_0^\infty F_2(f)(S_1(f) + N_1(f)) df, \end{aligned} \tag{3.5}$$

where $F_2(f)$ is the power gain of the filter, and $S_1(f) \equiv F_1(f)S_0(f)$ and $N_1(f) \equiv F_1(f)N_0 + N_1$ are the signal and noise power at the input. The first term is the

information rate for the filtered signal, plus that signal's noise, given the channel noise N_2; the second term is the power of the transmitted signal. I have included a relative cost factor, $P_2 + N_2$: the amount of power, per unit frequency, that we are willing to expend, for signal plus noise, for each baud (bits/sec) of information. Note that P_2 and N_2 are given in units of energy, just like $S_1(f)$, $N_1(f)$, and N_2.

Taking the derivative of this functional and equating it to zero, we obtain

$$F_2(f) = \frac{P_2}{S_1(f) + N_1(f)}. \tag{3.6}$$

With this filter, the transmitted signal, $F_2(f)(S_1(f) + N_1(f))$, is white, with uniform power spectral density of P_2. This signal level is expected, given the cost that we assigned to information, because the optimization procedure equalizes the marginal costs of information and energy. Substituting the expressions for $S_1(f)$ and $N_1(f)$, and using the result for $F_1(f)$, we find that the whitening filter is related to $S_0(f)$ by simply

$$F_2(f) = \frac{N_0}{BN_1} \frac{P_2}{S_0(f)}. \tag{3.7}$$

When $S_0(f) < N_0/B$ and $F_1(f)$ is set to zero, there is no signal and $F_2(f) = P_2/N_1$.

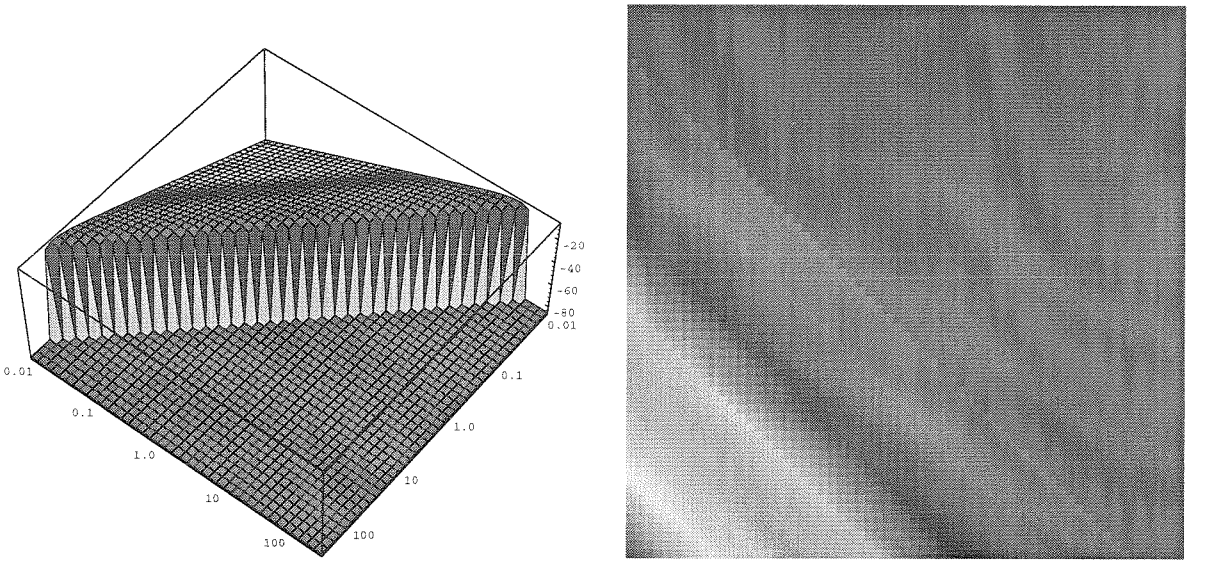
Measurements of the expected power spectral density of natural scenes yield $S_0(f) \simeq K/f^2$, where K has units of power times frequency[84, 85].[1] Noise due to quantum fluctuations—in the photon flux from the light source, or in the ionic flux through the photoreceptor membranes, or in vesicular neurotransmitter supply to the synaptic cleft—is white.

Given these ensemble statistics, the optimum noise-suppression filter is

$$F_1(f) = \frac{N_1}{N_0} \frac{B - (f/f_1)^2}{1 + (f/f_1)^2}, \tag{3.8}$$

where $f_1 = \sqrt{(K/N_0)}$. This filter is lowpass, with corner frequency at the point where

[1] The finite-power constraint requires that this spectral density flatten out at very low frequencies.



(a) (b)

Figure 3.3: Optimal Spatiotemporal Filter

Surface (a) and contour (b) plots of optimal spatiotemporal filter for natural scenes which have power spectra of the form $S(f_r, f_t) = (1/f_x^2)(1/f_t^2)$. The filter is optimized to equalize the energy at all frequencies and to reject bands with signal-to-noise ratio less than unity.

the SNR becomes $1/B$. The optimal whitening filter is

$$F_2(f) = \frac{P_2}{BN_1}\left(\frac{f}{f_1}\right)^2. \tag{3.9}$$

It is highpass. Hence the overall cascade is bandpass. The optimal spatiotemporal filter for signals with expected power spectrum of the form $S_0(f_r, f_t) = (1/f_r^2)(1/f_t^2)$, is plotted in Figure 3.3. The filter has low gain at low frequencies and its gain peaks along the diagonal line (on log–log coordinates) $f_r f_t$ = constant. I compare and contrast the optimal filter with the retinal filter in Section 3.2.3, after I present psychophisical and physiological measurements of the visual system in Sections 3.2.1 and 3.2.2.

3.2 Spatiotemporal Sensitivity

I now review measurements of the spatiotemporal-frequency sensitivity of the retina; these measurements reveal which parts of the visual signal are transmitted down the optic nerve and which parts are filtered out. We will consider the overall performance of the entire visual system, as assessed by behavioral experiments, and the performance of the retina itself, as assessed by spike-train recordings from individual ganglion cells. The engineering-style measurements I review here reveal a great deal about the microarchitecture of the retina, and about the mechanisms that the retina uses to process spatiotemporal visual signals.

Physiologists and psychophysicists have measured the spatiotemporal sensitivity of the retina using a frequency-domain approach, and have proposed fairly detailed biophysical models to account for the data. In contrast with the flashing spots and annulli much loved by physiologists for stimulating cells, this engineering-style approach uses moving (or flickering) sinusoidal gratings. In theory, these two approaches should yield the same information; in practice, the frequency-domain measurements are more robust and more sensitive. The down side of frequency-domain methods is that we must invoke linearity, space invariance, and time invariance to predict responses to

more complex stimuli. These ideal properties hold for only low contrast levels, for a fixed eccentricity, and for a fixed background-intensity level—conditions under which ON–OFF rectification, variations in sampling density, and light adaptation in the retina are insignificant.

3.2.1 Psychophysical Measurements

Over a period of 20 years, Kelly, and other psychophysicists, obtained a complete quantitative description of the spatiotemporal threshold surface of human vision [88, 92]. By compensating for the subject's eye movements [91], Kelly was able to measure responses to moving gratings. These measurements do not invoke any nonlinearities, because the signals used are at the threshold of perception. The experiment is repeated at different levels of background intensity, spanning several decades, to characterize nonlinear light-adaptation effects. These psychophysical experiments provide an input–output description of the entire visual system—including the optics, vitreous humour, spatial sampling, and all the parallel pathways—up to the observer, which is somewhere deep inside the cortex at an unknown location! Therefore, they provide only a lower bound for the performance of the individual stages. Also, only the magnitude of the response is measurable; the phase of the response cannot be obtained with these methods. In the absence of more powerful noninvasive measurement techniques, these data are all that are currently available for humans. Two sets of these psychophysical data are shown in Figure 3.4 and in Figure 3.5.

The first data set characterizes the dependence of temporal- and spatial-frequency selectivity on ambient light level, over 6 decades of intensity (Figure 3.4). Both spatial and temporal responses are bandpass under brightly lit conditions, but the sensitivity to high frequencies decreases as the lights dim, and the peak shifts to lower frequencies. At the lowest intensity levels, the response becomes lowpass. Notice that the transition occurs 1.5 decades earlier for the temporal response, at 3.75td versus 0.09td for the spatial response.

The second data set reveals the dependence of spatial filtering on temporal fre-

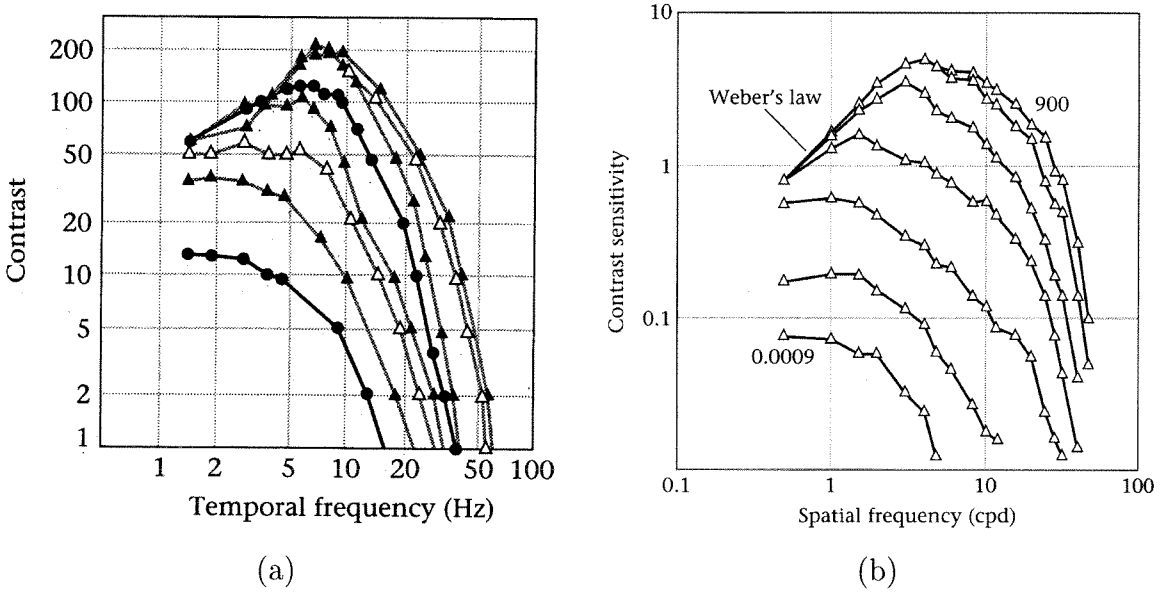


Figure 3.4: INTENSITY ADAPTATION AND FREQUENCY SENSITIVITY OF HUMANS
In these psychophysical experiments, researchers choose a background intensity level and modulate its amplitude sinusoidally, either in space or in time. Then they measure the modulation level, expressed as a fraction of the background intensity, required just to exceed the perceptual threshold of the subject. The contrast sensitivity, which is defined as the reciprocal of the threshold modulation, is plotted here. (a) Flicker contrast sensitivity versus temporal frequency for six different background-intensity levels: 0.375, 1, 3.75, 10, 37.5, 100, 1000, 10,000 td (from lowest curve to highest curve). The response changes from lowpass to bandpass and shifts to higher frequencies as intensity increases. (b) Grating contrast sensitivity versus spatial frequency for seven different background intensity levels: 0.0009 to 900td, increasing in steps of 1 decade (from lowest curve to highest curve). For intensities above 900td, the curves are identical to the one for 900td. Again, the response changes from lowpass to bandpass and shifts to higher frequencies as intensity increases. We can convert the troland units (td) used for intensity to photons absorbed per second per cone by multiplying by 10, or to photons absorbed per second per rod by multiplying by 4. Reproduced from [86]. Original sources: a [87];b [88].

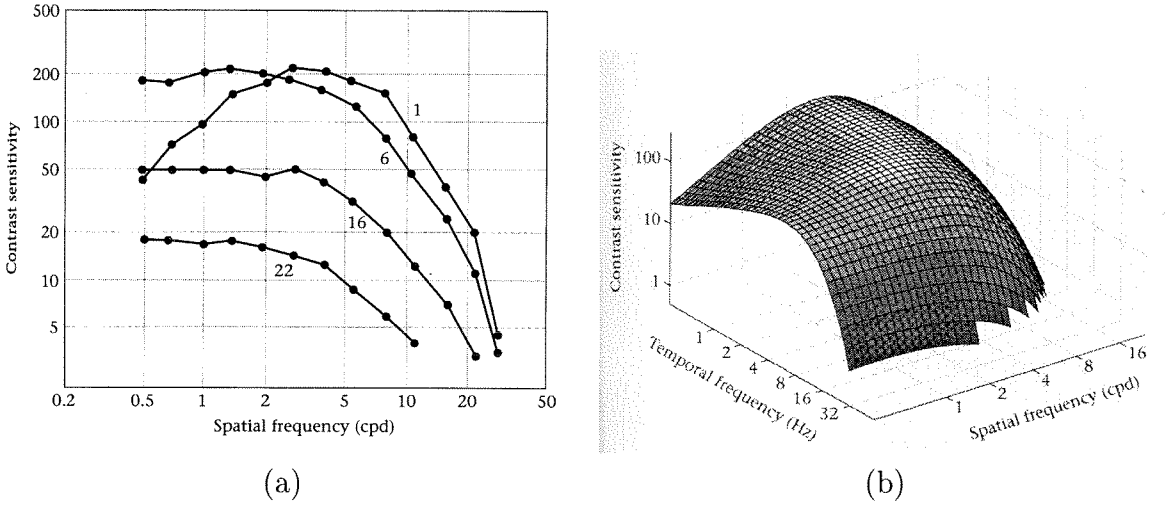


Figure 3.5: SPATIOTEMPORAL CONTRAST SENSITIVITY OF HUMANS
Sinusoidal gratings, superimposed on a mean background level, were used and their amplitudes were modulated sinusoidally in time to produce a contrast-reversing pattern.(a) Contrast sensitivity versus spatial frequency at four different temporal frequencies (in units of cps). The spatial-frequency response is bandpass at low temporal frequencies, but becomes lowpass at high temporal frequencies. (b) Three-dimensional plot of spatiotemporal contrast-sensitivity function. The mean intensity was 1000 td. The curves in (a) correspond to cross-sections of this surface taken parallel to the spatial frequency axis, at different points on the temporal frequency axis. Plotting the measurements in three-dimensions makes it evident that the temporal frequency sensitivity also is bandpass at low spatial frequencies and becomes lowpass at high spatial frequencies. Reproduced from [86]. Original sources: a [89]; b [90, 91, 92].

quency, and vice versa, at high intensity levels (Figure 3.5). Although the spatial-frequency response is bandpass at low temporal frequencies, it does not remain so as temporal frequency increases. Sensitivity to low spatial frequencies increases with temporal frequency, and the spatial filter transitions gradually from bandpass to low-pass. Sensitivity to low spatial frequencies does not increase indefinitely; it starts decreasing for temporal frequencies above 6cps. Similarly, the temporal-frequency response is bandpass at low spatial frequencies, and becomes lowpass at high spatial frequencies.

The dependence of filtering on the frequency in the other dimension is uncannily similar to the dependence on intensity. There is an important quantitative difference, however. As the frequency in the other space–time dimension increases, the filter changes from bandpass to lowpass, but the cutoff point does not shift to lower frequencies as it does with decreasing intensity—all the curves approach the same point at high frequencies.

3.2.2 Physiological Measurements

In a remarkable series of physiological experiments, Enroth-Cugell and her coworkers characterized the spatiotemporal properties of X and Y retinal ganglion cells in the cat [94, 95, 93]. Unlike psychophysicists, physiologists usually use high-contrast stimuli that maximally excite the cell, and, no doubt, drive it into the nonlinear regime. Enroth-Cugell and coworkers were careful to collect data for spike rates below 10 spikes/sec, because their measurements indicated that nonlinearities became significant outside this range (this spike rate corresponds to 10 percent or lower contrast) [94, 93]. They also characterized the nonlinear spatial interactions in the Y cell's receptive field [95]; the data are shown in Figure 3.6. Most physiological studies to date provide data at only one level of adaptation (but see [93, 96]), whereas psychophysical studies may span up to 6 decades of light intensity.

Despite species differences and measurement techniques, the physiological responses of the X cell parallel the psychophysical ones: The temporal filter is bandpass

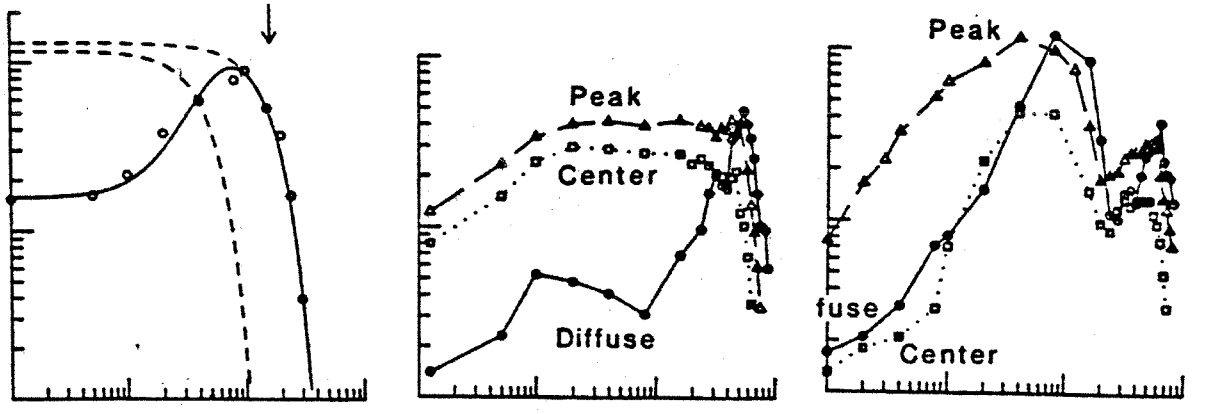


Figure 3.6: SPATIOTEMPORAL CONTRAST SENSITIVITY OF CAT GANGLION CELLS
In these physiological experiments, researchers use sinusoidal spatiotemporal patterns—which they generate either by moving a grating or by modulating the light intensity with time—to stimulate the cell; they then record the cell's spike train. To obtain a measure of the linear component of the response, they compute the magnitude and phase of the Fourier component of the average spike rate at the temporal frequency of the stimulus. The responsitivity, which is defined as the ratio between the amplitude of this Fourier component and the contrast of the stimulus, is plotted here. (a) Grating-contrast responsitivity versus spatial frequency for an X cell. For the high intensities used, the response is bandpass and is well fitted by the difference-of-Gaussians (DOG) model. (The solid curve is obtained from the difference of the two dashed Gaussian curves.) The experimenter measured the temporal-frequency responses shown in the other panels at three carefully chosen spatial frequencies, labeled diffuse (low frequency), peak (frequency at which spatial bandpass peaks), and center (bandwidth of center Gaussian used in DOG model). (b) Contrast sensitivity versus temporal frequency for a X cell (diffuse = 0.01 cycles per degree, (cpd), peak = 1.4 cpd, and center = 2.4 cpd). A bandpass response is obtained at low spatial frequencies, but the response becomes lowpass at high spatial frequencies. (c). Contrast sensitivity versus temporal frequency for a Y cell (diffuse = 0.01 cpd, peak = 0.2 cpd, and center = 0.42 cpd). The response is always bandpass, although the peak becomes broader at the peak spatial frequency. The phase measurements, which have been omitted for brevity, show that, at low frequencies, the Y cells lead the X cells by almost 90°; at higher frequencies, however, the phase changes linearly with frequency for both cells—a characteristic of a pure delay element. The corresponding delay—that is, the delay between the peaks in the input sinusoid and in the cell's firing rate—was 24 msec for X cells, and 20 msec for Y cells. Reproduced from [93].

at low spatial frequencies, and becomes lowpass at high spatial frequencies, but the high-frequency cutoff remains the unchanged. The same description applies to spatial filtering vis-à-vis temporal frequency. However, the Y cell's responses are strikingly different; they are bandpass for all spatial frequencies.

This difference between X and Y cells is interesting as the primary difference between these pathways is in the amount of processing that occurs in the IPL. Y (α) cells receive a lot of input from amacrine cells, whereas X (β) receive little input from amacrines. Hence, the amacrine cells are probably responsible for removing the low spatial frequencies present in the X cells' responses at high temporal frequencies, and the low temporal frequencies present in the X cells' responses at high spatial frequencies.

3.2.3 Theory and Experiment

The theory of optimal filtering qualitatively accounts for the bandpass spatial filtering and the bandpass temporal filtering, since the power spectrum of natural scenes is given by f^{-2} for both temporal and spatial frequency. However, it does not account for the effect of temporal frequency on spatial filtering, or vice versa.

The full spatiotemporal frequency power spectrum of natural scenes is given by $f_r^{-2} f_t^{-2}$ [84, 85]; it can be factored into a spatial-frequency component and a temporal-frequency component. Therefore, increasing the frequency in the other dimension should have the same effect as reducing the overall signal level. Hence, we would expect the behavior to be the same as that we saw for intensity. And we do indeed observe the same qualitative behavior. However, the high-frequency cutoff does not shift to lower frequencies as the frequency in the other dimension increases, as we expect from the theory (Compare Figure 3.3 and Figure 3.5b).

Bandpass behavior is optimal at high light levels, where the SNR is high. However, we do not expect it to be optimal when the SNR is low, simply because there is little point in redistributing the signal energy over the spectrum when the noise is dominant everywhere.

SNR will decrease with intensity because the variance of the noise decreases linearly as the mean number of quanta integrated over a given volume of space–time decreases. Indeed, the variance is equal to the mean for a Possion-like point process; hence, the SNR is simply $\sqrt{I}/I$, when the light intensity is expressed as quanta per space–time volume. Therefore, the SNR decreases as $1/\sqrt{I}$.

This dependence of SNR on intensity explains the shift of the cut-off frequency, and the peak frequency, to lower frequencies. The response becomes lowpass when the unity-SNR frequency approaches 0. The fact that the transition occurs first in the temporal-frequency response may indicate that the pathway that carries temporal information integrates over a smaller volume of space–time than does the pathway that carries spatial information.

A very striking aspect of the psychophysical measurements is that the contrast thresholds all fall within less than 2 decades, although the signals that are producing this contrast change by 6 decades. That is, the visual system acts as though the light has changed by only 1.5 decades, when in fact it has changed a millionfold! We know from recording from retinal ganglion cells that this intensity normalization happens in the retina. The retina cares less about absolute intensity, and more about how much signals change relative to the ambient intensity level. Hence, we see the importance of contrast in characterizing the behavior of the visual system.

The dependence of visual sensitivity on light intensity is characterized by **Weber's law**, which states that the threshold is proportional to the mean intensity level. Weber's law works fantastically well at high intensities, as is evident the asymtotic behavior of the curves in Figure 3.4. However, it breaks down at low light levels, as the curves move apart with decreasing intensity.

In the low-intensity region, the dependence of visual sensitivity on light intensity is characterized by the **de Vries–Rose law**, which states that the threshold is proportional to the square root of the mean intensity level. Since the noise dominates in this regime, and the noise level is given by the square root of the intensity, the visual system sets its gain inversely to the input level in both cases. As a result, the output-signal falls in a limited range.

Operation	Standard	Retinal
Detection	Integrating	Continuous
Gain control	Global	Local
Filtering	Allpass	Bandpass
Quantization	Fixed	Adaptive
Architecture	Serial	Parallel

Table 3.1: STANDARD VERSUS RETINAL DESIGN PRINCIPLES

3.3 Biology Versus Engineering

The functional and structural organization of the retina is radically different from that of standard human-engineered imagers. The design principles employed by standard imager technology are outlined in Table 3.1; the design principles of the retina are also listed for comparison. These principles are compared and contrasted in Section 3.3.1 through Section 3.3.5.

3.3.1 Sensing: Continuous Versus Integrating

Integrating detectors (e.g., charge-coupled devices (CCDs) [97] and photogates [98]) suffer from blooming at high intensity levels and require a destructive readout (reset) operation. Continuous-sensing detectors (e.g., photodiodes or phototransistors) do not bloom, and can therefore operate over a much larger dynamic range [99]. In addition, redundant readout operations can be eliminated, with considerable power savings, because charge does not accumulate.

Continuous-sensing detectors have been shunned, however, because they suffer from gain and offset mismatches that give rise to salt-and-pepper noise in the image. However, Buhman and colleagues have shown that the powerful learning capabilities of image-recognition systems can compensate easily for this fixed pattern noise [100].

The real benefit of using continuous sensors lies in the latter's ability to perform analog preprocessing before quantizing the signal. A signal that takes on a discrete set of values at a discrete set of times (quantized in amplitude and time) carries less information than does a signal that takes on the full continuous spectrum of

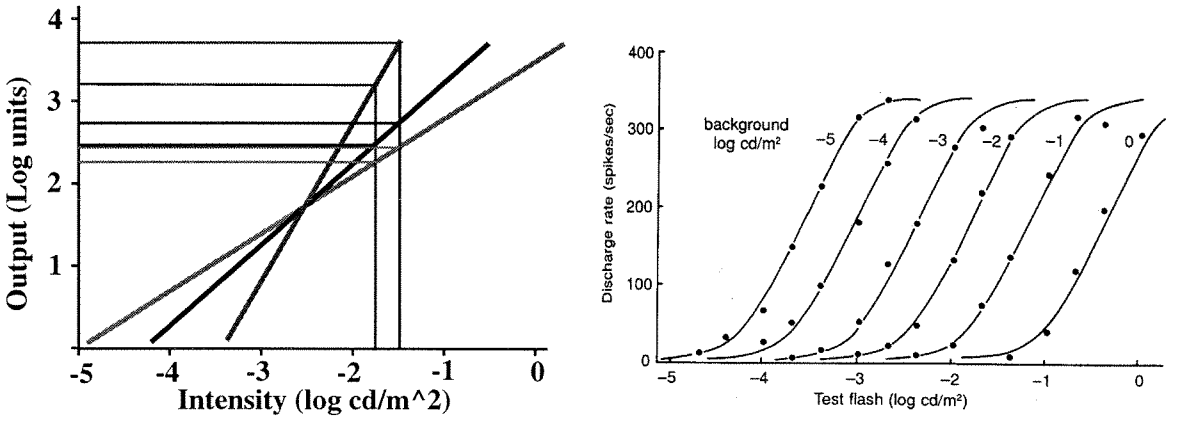


Figure 3.7: INPUT–OUTPUT TRANSFER CURVES FOR LIGHT SENSORS
(a) As larger and larger input ranges are spanned, the slope decreases, and finer resolution is required to detect the same percentage change in the input signal. (b) Using transfer curves that can be centered at the local intensity level decouples dynamic range and resolution. Each curve spans only a 20-fold input range, since local variations in intensity are due primarily to changes in reflectivity: A black sheet has a reflectivity of 0.05, and a white sheet has a reflectivity of 0.95. These transfer curves were measured for the cat retina, and were reproduced from [101].

amplitudes and times. For instance, graded potentials in the nervous system can transmit information at the rate of 1650 bits per second—over four times the highest rate measured for spike trains [10].

The analog operations described in and Section 3.3.2 Section 3.3.3 reshape the spectral and amplitude distribution of the analog signal, to transmit information efficiently through this bottleneck.

3.3.2 Amplification: Local Versus Global Control

Imagers that use global automatic gain control (AGC) can operate under only uniform lighting because the 1000-fold variation of intensity in a scene with shadows exceeds their 8-bit dynamic range.[2] A charge-coupled device or photogate can achieve 12 bits (almost 4 decades) [98], and a photodiode or phototransistor can achieve 20 bits

[2]I am assuming a linear encoding—a practice that is the standard. This assumption limits the dynamic range to 2^b for a b-bit encoding.

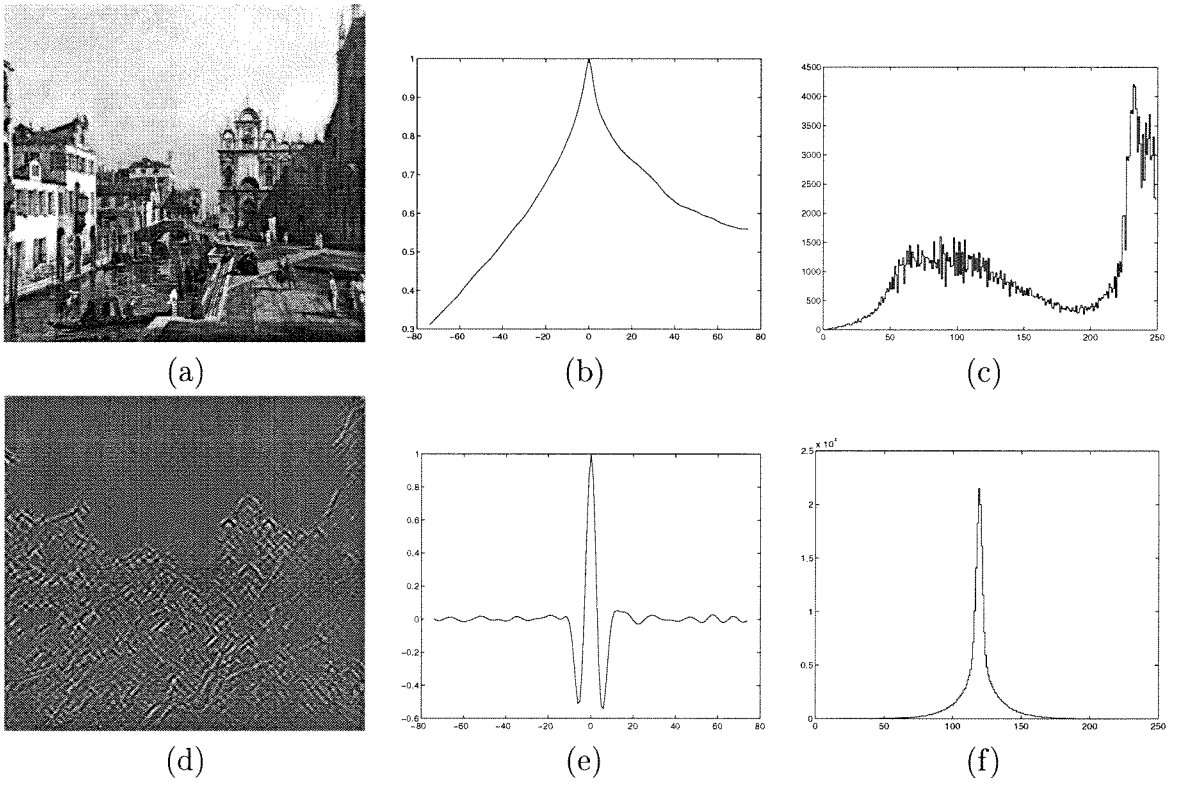


Figure 3.8: BANDPASS FILTERING

The top row shows the original 512×479×8-bit image (a), that image's autocorrelation (b), and its amplitude histogram (c). The bottom row shows the bandpass-filtered image (d), that image's autocorrelation (e), and its amplitude histogram (f). In the original image, pixels are highly correlated, and the correlation falls off slowly with distance. Whereas, the correlation is a lot less in the bandpass filtered image, and falls off rapidly. The distribution of amplitudes in the original image is broad and bimodal, due to the relatively bright overcast sky and the dark foreground objects. In contrast, the amplitude distribution for the filtered image is clustered around 0 (119), and decays rapidly.

(6 decades) [99, 102]—but the phototransistor's performance in the lowest 2 decades is plagued by slow temporal response. The dynamic range of the system's output, however, is limited by the cost of precision analog read-out electronics and A/D converters, and by video standards.

When AGC acts globally, the input dynamic range matches the output dynamic range, and the only way to extend the input range is to extend the output dynamic range. In practice, we must reduce the noise floor to improve resolution.

As shown in Figure 3.7, local AGC decouples dynamic range and resolution, extending the input dynamic range by mapping different parts of the input range to the limited output range, depending on the local intensity level. This solution is beneficial if the resolution required to discriminate various shades of gray (1 in 100 for the human visual system) is poorer than the resolution required to span the range of all possible input levels (at least 1 in 100,000 for the photopic range of human vision).

3.3.3 Filtering: Bandpass Versus Allpass

On average, natural images have a $1/f^2$ power spectrum for both spatial and temporal frequency [103, 85], whereas noise, due to quantum fluctuations, has a flat spectrum. Consequently, imagers that transmit the full range of frequencies present pass on mainly noise at high frequencies, where the signal-to-noise is poor, and pass on redundant information at low frequencies, where the signal-to-noise is good. Bandpass spatiotemporal filtering rejects the wideband noise, and attenuates the redundant low-frequency signals; this strategy is the optimal one for removing redundancy in the presence of white noise [81, 83, 82].

Figure 3.8b and d illustrate the redundancy reduction that I achieved using bandpass filtering, by computing the correlation between pixel values.[3] The correlation is over 40% for pixels that are 60 pixels apart in the raw image. In the filtered image,

[3] I performed bandpass filtering by convolving the input image with the Laplacian of a Gaussian with $\sigma = 2.5$ pixels. I calculated the autocorrelation of the images by subtracting out the mean, shifting a copy of the image up or right by 1 to 75 pixels, multiplying corresponding pixels, and summing; I normalized the results to yield a maximum of unity. Rightward shifts are plotted on the positive axis (0 to 75), and upward shifts are plotted on the negative axis (0 to -75).

pixels tha are more than 10 pixels apart have less than 5% correlation. Comparison of the amplitude histograms before and after filtering (Figure 3.8c,f) demonstrates that bandpass filtering has two additional benefits.

First, bandpass filtering results in a sparse output representation. For our sample image, 24.4% of the pixels fall within ±0.39% of the full-scale range (i.e., ±1LSB at 8-bit resolution); 77.5% of them fall within ±5% (i.e., ±13 at −127 to +127 amplitude range). Hence, if we choose to ignore amplitudes smaller than 5%, we need to transmit only 22.5% of the pixels. In practice, the degree of sparseness will depend on the cut-off frequency of the bandpass filter. Although rejection of high frequencies introduces some redundancy, this rejection is necessary to protect the signal from noise that is introduced by the signal source or by the circuit elements.

And second, bandpass filtering results in a unimodal amplitude distribution that falls off exponentially. For our sample image, the distribution is fit by a sum of two exponentials that change by a factor of $e = 2.72$ whenever the amplitude changes by 2.5 and by 14.0, on a ±128 scale; the rapidly decaying exponential starts out 4.5 times larger. Empirical observations confirm that this simple model holds for a wide range of images.

In contrast, the distribution of raw intensity values is difficult to predict, because gross variations occur from scene to scene, due to variations in illumination, image-formation geometry (surface and light-source orientation), and shadows [6]. These slowly changing components of the image are removed by local AGC and bandpass filtering. When the bandpass characteristics are fixed and the intensity is normalized, the parameters of the amplitude distribution are determined mainly by reflectivity and therefore vary much less; the quantizer can exploit this invariance to distribute its codes more effectively.

3.3.4 Quantization: Adaptive Versus Fixed

The quantization intervals of traditional A/D converters are set to match the maximum rate of change and the smallest amplitude, as shown in Figure 3.9. This uniform

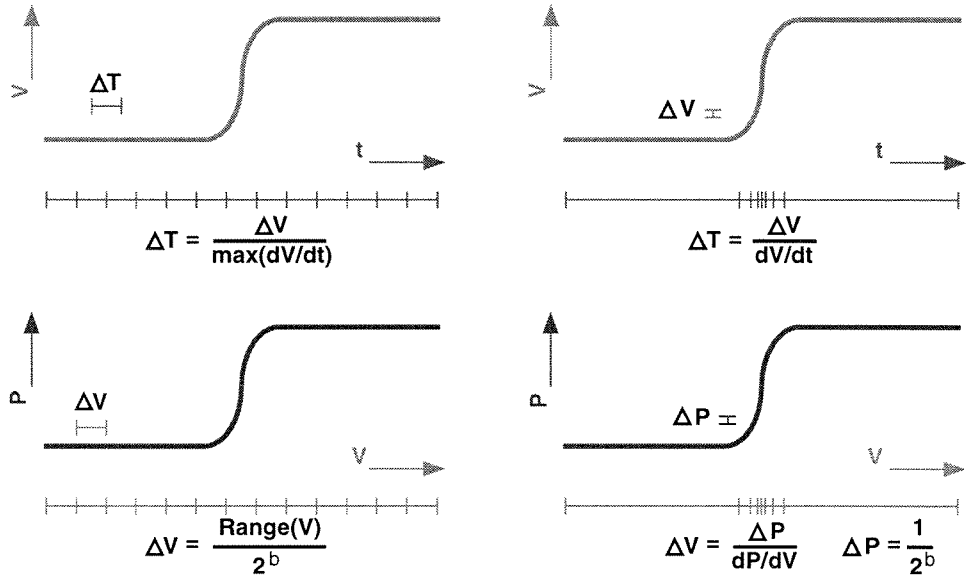


Figure 3.9: Quantization in Time and Amplitude

Top row: Time intervals (ΔT) are set to match the maximum rate of change (left column). The signal is sampled repeatedly, even when $dV/dt \approx 0$—that is, when the change is insignificant (oversampling). Instead of fixing the time step, it is more efficient to fix the voltage step (ΔV), and to adapt the time intervals dynamically to achieve this change in voltage, as shown on the right. Bottom row: Amplitude intervals (ΔV) are also uniformly distributed. The signal is sampled repeatedly, even though $dP/dV \approx 0$—that is, although the probability that the input amplitude falls in this interval is negligible. Instead of fixing the voltage step, it is more efficient to target a certain change in the cumulative probability ($\Delta P = 2^{-b}$, where b is the number of bits per sample), and to choose voltage intervals statistically to achieve this change in probability, as shown on the right.

quantization is optimum only when high frequencies dominate and all amplitudes are equally likely. As we have seen, neither case applies to natural scenes: the power spectrum decays with frequency, as in $1/f^2$, and the amplitude probability density decays exponentially—after local gain control and bandpass filtering remove variations in illumination. Therefore, uniform quantizers produce numerous redundant samples, because changes in the signal are relatively rare [85], and underutilize their large amplitude codes, because these signal amplitudes occur rarely in natural scenes [6].

Assuming that temporal changes are due primarily to motion, we can estimate the amount of redundancy from the spatial-frequency power spectrum and from the velocity distribution. The velocity distribution, measured for movies and amateur videos, is dominated by low velocities and falls off with a power law of 3.7 [85]. High velocities will be even more drastically attenuated in an active vision system that compensates for global motion, and that tracks objects [84]. After bandpass filtering, signals that change gradually over space are eliminated and rapid changes occur only rarely and over much more restricted areas.

Due to the absence of high speeds and of nonlocal intensity variations, the imager's output signals rarely change rapidly. Consequently, adapting the sampling rate to the rate of change of the signal greatly reduces the number of samples produced. Alternatively, this adaptation allows higher temporal bandwidths to be achieved for a given mean sampling rate.

Using the amplitude distribution of our bandpass-filtered sample image, we can calculate the probability of failing to discriminate between a pair of samples drawn from that distribution: It is 0.0384 when the 2^8 quantization levels are uniformly distributed—an order of magnitude bigger than the minimum confusion rate of $1/256 = 0.0039$, which occurs when we choose the quantization levels to make it equally likely that we will draw a sample from each interval. In fact, the confusion rate of 0.0384 can be achieved with just $\log_2(1/0.0384) = 4.7$ bits per sample if the quantization levels are optimally distributed.

A quantizer that assigns its codes to probable amplitudes, rather than to improbable ones, maximizes the probability of discriminating between any two amplitude

levels drawn from the input distribution; thus, information is maximized when all codes are equiprobable [8].

3.3.5 Architecture: Parallel Versus Serial

In addition to differing in the aforementioned design principles, biological and human-made vision systems also use radically different architectures. The retina performs the four operations listed in Table 3.1 in a pixel-parallel fashion, whereas most synthetic imagers perform only detection in the pixel. The few synthetic imagers that also amplify and quantize the signal, perform these operations pixel serially, and set the gain, sampling rate, and step size to be the same for all pixels [98, 104, 105]. In sharp contrast to human-engineered imagers, the retina adapts its gain, sampling rate, and step size locally, to minimize redundancy; the retina also whitens the signal in space and time, to make its output samples independent.

Since the work on television in the fifties, engineers have known that images, and other naturally occuring signals such as speech, are highly redundant. They realized that sending the raw intensity values is not the most efficient way to transmit information about these signals; the data can be encoded much more efficiently. Indeed, they perform such encoding routinely, after acquiring and quantizing the image, using digital computer. However, we are now learning that the retina knows about efficient encoding, as well, and the lesson that it teaches us is that we can make major gains by performing these operations right up front in the pixel.

There is, however, a stiff price to pay to get pixel-parallel operation. We must add several transistors to the pixel to perform the computations, and these transistors take up room, increasing the size of the pixel. The wires needed to communicate between pixels take up even more room than that alloted to the transistors! Since silicon-based VLSI technology is two-dimensional, the pixel area must increase, and we end up sacrificing the sampling density.

Similar structural constraints are faced by neurobiology: It can fit into a given volume a limited number of synapses and a limited length of dendrites and axons.

Neurites used for communication also take up much more room than do the synapses that do the computation. These contraints are less severe for biology, but not by much. Judging from the number of cell layers, gray matter is thick enough to stack only two to five layers of processing. Also, the dimensions of wires and transistors are now in the submicron range, approaching the sizes of the finests dendrites and the smallest synapses. Thus, silicon-based VLSI is encroaching on the territory of carbon-based VLSI.

Matching the level of integration is a necessary first step. The real challenge, however, is figuring out how to use wires and energy efficiently, so that we can harness the awesome computational power available from gigantic numbers of synapses or transistors. Together with optimizing the functional contraints discussed in this chapter, the retina optimizes these structural constraints as well. As we shall see in Chapter 3, this global optimization explains the descrepancy between the optimal theoretical spatiotemporal filter we derived and the retinal filter.

3.4 Summary

The retina has evolved sophisticated filtering and adaptation mechanisms to reduce redundancy and to improve coding efficiency. Six such mechanisms follow:

1. *Local automatic gain control* at the receptor level eliminates the dependence on lighting intensity—the receptors respond to only contrast—extending the sensor's dynamic range.

2. *Bandpass spatiotemporal filtering* in the first stage of the retina (OPL) attenuates signals that do not occur at a fine spatial *or* temporal scale, ameliorating redundant transmission of low-frequency signals and eliminating noisy high-frequency signals.

3. *Highpass temporal and spatial filtering* in the second stage of the retina (IPL) attenuates signals that do not occur at a fine spatial scale *and* temporal scale, eliminating the redundant signals passed by the OPL, which responds strongly

to low temporal frequencies that occur at high spatial frequencies (sustained response to static edge) or to low spatial frequencies that occur at high temporal frequencies (blurring of rapidly moving edge).

4. *Half-wave rectification*, together with dual-channel encoding (ON and OFF output cell types), in the relay cells between the OPL and the IPL (bipolar cells), and between the retina and the rest of the brain (ganglion cells), eliminates the elevated quiescent neurotransmitter release rates and the elevated firing rates required to signal both positive and negative excursions using a single channel.

5. *Phasic transient–sustained response* in the ganglion cells avoids temporal aliasing by transmitting rapid changes in the signal using a brief, high-frequency burst of spikes, and, at the same time, avoids redundant sampling by transmitting slow changes in the signal using modulation of a low, sustained firing rate.

6. *Foveated architecture*, with active directing of the gaze, eliminates the need to sample all points in the scene at the highest spatial and temporal resolution, while providing the illusion of doing so everywhere. The cells' spatiotemporal receptive fields are optimized: smaller and more sustained at the fovea (parvocellular or X-cell type), where the image is stabilized by tracking, and larger and more transient in the periphery (magnocellular or Y-cell type), where motion occurs.

The resulting activity in the ganglion cells, which convert these preprocessed signals to spikes and transmit the spikes over the optic nerve, is different from the stimulus pattern.

For relatively long periods, the scene captured by the retina is stable. These static events produce sparse activity in the OPL's output, since the OPL does not respond to low spatial frequencies, and produce virtually no activity in the IPL's output, since the IPL is selective for temporal frequency as well as for spatial frequency. The

OPL's sustained responses drive the 50,000 or so ganglion cells in the fovea, allowing the fine details of an object stabilized by tracking to be analyzed. The vast majority of the ganglion cells—about 1 million in all—is driven predominantly by the IPL, and fires at extremely low quiescent rates of 10 spikes per second, or less, in response to the static event.

When a localized punctuated event—like a small light flash—occurs, the OPL and the IPL respond strongly, since both high temporal frequencies and high spatial frequencies are present. Thus, a minute subpopulation of OPL- and IPL-driven ganglion cells raises its firing rates to a few hundred spikes per second. On the other hand, if the punctuated event lights up a large area, the OPL-driven ganglion cells still respond strongly, for a short time, due to the presence of high temporal frequencies, whereas the response of the IPL-driven ganglion cells is attenuated, due to the presence of low spatial frequencies. Consequently, the number of ganglion cells that respond is miniscule.

A dynamic event—such as a spinning windmill or panning the eyes—produces puntuated events at adjacent locations in rapid succesion. In the limit, a dynamic event is equivalent to a punctuated event that lights up a large area. Thus, a dynamic event can activate a large number of OPL-driven ganglion cells. However, IPL-driven ganglion cells, which cover most of the retina, are not activated, because the low spatial frequencies produced in the OPL's output by dynamic stimuli are suppressed by amacrine cells, attenuating the IPL's response.

In effect, the activity in the optic nerve is clustered in space *and* time (whitened spectrum): It consists of sporadic short bursts of rapid firing, triggered by punctuated and dynamic events, overlaid on a low, steady background firing rate driven by static events.

Chapter 4 Retinal Spatiotemporal Dynamics: A Physical Model

To discover how the retina implements bandpass spatiotemporal filtering, and to understand the tradeoffs that it makes in the face of severe wiring limitations, I analyze the spatiotemporal behavior of a simple dynamic model of the retina. This model is a physical one: It is built out of resistors, capacitors, and transconductances. It is based on the neurocircuitry of the vertebrate retina; it includes several major synaptic interactions in the outer plexiform layer (OPL). My goal is synthesize the minimal amount of machinery required to reproduce the observed qualitative behavior, rather than to provide detailed quantitative predictions of retinal responses.

In particular, I seek the simplest linear physical model that reproduces the salient features of retinal spatiotemporal dynamics, and I employ circuit theory and Fourier methods to obtain closed-form analytical descriptions of its behavior. These analytical expressions are indispensable to understanding the tradeoffs inherent in this simplified retina model. To the extent that these tradeoffs arise from fundamental physical limitations—such as the inseparability of spatial and temporal processing—they carry over to the real retina, or at least to those parts of the retinal structure that the model includes.

This approach is part of an overarching layered-complexity strategy that I have adopted, where we reverse-engineer the retina by peeling away one level of complexity at a time. Once we know the tradeoffs inherent in the design of a piece of neurocircuitry, we can see how to introduce an additional layer of complexity to improve its performance. Although a linear model cannot include adaptation mechanisms, such as gain control, we can often achieve the desired result by varying the parameters of the linear circuit, such as its gain or its time and space constants, appropriately.

Adaptation matches the gain of the filter to the mean signal level, and matches

the tuning of the filter to the signal-to-noise ratio. Since the linear filter's tradeoffs are stated in terms of these very same parameters, studying the linear case helps us understand how adaptation affects system performance. By relating these parameters to the values of resistors, capacitors, and transconductances in the model, the linear analysis can guide the design of these adaptation mechanisms.

Layering adaptation on top of filtering in this fashion is valid, since these two mechanisms act on disparate spatial and temporal scales. Filtering occurs over tens of milliseconds of time and tens of minutes of visual angle, whereas adaptation occurs over hundreds of milliseconds of time and degrees of visual angle.

4.1 Assumptions of the Model

I construct linear electrical-circuit models of the retinal neurocircuitry by simplifying the latter's biophysical elements in three ways:

1. *Gap-junction–coupled cell syncytia are isotropic resistive grids.* I abstract the fine physical structure of these cells into a **characteristic lateral resistance** and a **characteristic vertical conductance**. The former models the gap junctions, and the latter models the parallel combination of synaptic and leakage conductances; voltage dependencies, calcium dependencies, and nonlinearities of the membrane channels are ignored.

2. *Synaptic inputs are variable current sources.* I treat chemical synapses, which are usually modeled by conductance changes, as variable current sources. These model synapses are characterized by a **transconductance:** the additional current injected across the postsynaptic membrane per unit change in the presynaptic voltage. [1]

3. *Synaptic transmission is instantaneous.* I ignore the time dependencies of neurotransmitter release and diffusion, and those of the channel-gating mechanisms.

[1] Synapse models based on conductance changes are characterized by a conductance per unit voltage. Multiplication of this parameter by the voltage across the channel gives the equivalent transconductance.

Hence, the model's temporal dynamics arise solely from the membrane capacitances, and are characterized by the time constants of the cells.

Ignoring the fine details of cell morphologies an treating syncitia as isotropic resistive networks is justified by virtue of the dense, strong, local electrical connectivity in these cell syncytia. As the receptive fields are larger than the extent of the cells' dendritic arbors, the relay of signals from cell to cell across gap junctions appears to play a dominant role in shaping the cells' receptive fields—not the fine details of the dendritic arbor.

Ignoring voltage and calcium dependencies, and other nonlinearities, and treating synapses as current sources, is justified because the retina responds linearly for contrasts less than 10% [94]. Given that the threshold is 0.5% contrast, the retina is linear over a 20-fold range. For these small signal changes, the nonlinear voltage–current relationships of the ion channels, and of the gap junctions, can be replaced by their slope conductances, and the conductance changes due to activating more ion channels are negligible compared to the conductance of the cell.

Ignoring the time-course of synaptic transmission is justified because synaptic transmission occurs much faster than the cell responds, due to the large capacitance of the cell membrane.

Several researchers have used resistive networks to model gap-junction–coupled syncytia, going back to the work of Torre and Owen on rod coupling [106]. Chemical synapses have also been modeled previously as transconductances by Yagi and his colleagues [107]. Yagi and colleagues included time dependencies in their synapse model by using complex transadmittances, instead of real transconductances [107].

The model that they obtained by making these simplifications is discrete in space, but continuous in time; it is described by a difference equation in space and a differential equation in time. We can analyze such discrete–continuous systems by taking the Laplace transform in time, and obtaining a solution to the difference equation in space in terms of geometrically weighted Laplace transforms terms, as Yagi and his colleagues did [107]. Another approach is to work with discrete spatial frequencies and continous temporal frequencies, using the z-transform and the Fourier transform,

respectively, as Beaudot has shown [108]. Both of these approaches work, but they produce unweildly solutions that are difficult to grasp intuitively.

To obtain simple and intuitive results, I analyze the model in the continuum limit, where second-order spatial differences become second-order spatial derivatives. As

$$V_{i+1} - 2V_i + V_{i-1} = \varepsilon^2 \frac{d^2V}{dx^2} + \frac{1}{12}\varepsilon^4 \frac{d^4V}{dx^4} + \cdots$$

where $V(\varepsilon i) \equiv V_i$.[2] The error that we incurr by taking the continuous approximation is

$$\xi_{xx} \approx \frac{\pi^2}{12}\left(\frac{f_x}{f_{\text{Nyq}}}\right)^2,$$

when expressed as a fraction, where f_x is the spatial frequency and $f_{\text{Nyq}} \equiv (2\varepsilon)^{-1}$ is the Nyquist limit. It is negligible for spatial frequencies $f_x^2 \ll (12/\pi^2) f_{\text{Nyq}}^2$. We can use this expression to calculate the total error, if we know the power spectrum of the input signal. When most of the signal energy is at low frequencies—as it is for a step edge—the error is small. Hence, we do not lose much precision by taking the continuous approximation, and we gain much clarity by treating space and time uniformly.

Another concern that we have to address when we simulate a discrete network with a continuous one is the frequency limitations imposed by Nyquist's sampling theorem. To prevent aliasing, the discrete network is prohibited from seeing any frequencies higher than the Nyquist limit (f_{Nyq}). The continuous network, on the other hand, has no such restriction, and may produce frequencies higher than the Nyquist limit. We must filter out these frequencies before we can make valid predictions about the discrete network that we are simulating.

The continuous-space approximation has been used previously by Chen and Freeman [109]. They drew the analogy between gap-junction–coupled syncytia and a cable; this insight enabled them to apply results obtained for cables by Jack and others [110] to analyze the spatiotemporal dynamics of their retina model [109]. However,

[2] I obtained this result by using the Taylor series expansion for $V(x)$ at $x = \epsilon i$ to obtain expressions for V_{i-1} and V_{i+1}

their analysis focused on the overall spatiotemporal behavior of the retina, from the cornea to the ganglion cells. My analysis is restricted to the outer retina, and reveals more about the contribution of the cone–horizontal-cell circuit to the retina's response to spatiotemporal signals.

4.2 Linear Model of the Outer Plexiform Layer

The OPL circuit model is shown in Figure 4.1. Models more or less identical to this one have been proposed previously by Chen and Freeman, and by Yagi and colleagues [107]. As stated in Section 4.1, I use an analytical approach that is similar to that of Chen and Freeman by taking the continuous approximation, whereas Yagi analyzed the discrete case.

In the continuum limit, we have

$$I_o + \nabla^2 V_c / r_{cc} = g_{c0} V_c + c_{c0} \dot{V}_c + g_{ch} V_h, \tag{4.1}$$

$$g_{hc} V_c + \nabla^2 V_h / r_{hh} = g_{h0} V_h + c_{h0} \dot{V}_h, \tag{4.2}$$

where current per unit area, sheet resistance, conductance per unit area, and capacitance per unit area are used. The voltages V_c and V_h are continuous functions of space, (x, y), and time, t; $\nabla^2 f$ is the spatial Laplacian of f (i.e. $\partial^2 f/\partial x^2 + \partial^2 f/\partial y^2$), and $\dot{f}$ is the temporal derivative of f (i.e. $\partial f/\partial t$).

Assuming infinite spatial extent and homogeneous initial conditions, we can take Fourier transforms in space and time. Transforming the equations and solving, we obtain the following transfer functions between inputs and outputs.

$$\tilde{H}_c(\rho, \omega) \equiv \frac{\tilde{V}_c}{\tilde{I}_o} = \frac{1}{g_{ch}} \frac{\ell_h^2 \rho^2 + i\tau_h \omega + \epsilon_h}{(\ell_c^2 \rho^2 + i\tau_c \omega + \epsilon_c)(\ell_h^2 \rho^2 + i\tau_h \omega + \epsilon_h) + 1}, \tag{4.3}$$

$$\tilde{H}_h(\rho, \omega) \equiv \frac{\tilde{V}_h}{\tilde{I}_o} = \frac{1}{g_{ch}} \frac{1}{(\ell_c^2 \rho^2 + i\tau_c \omega + \epsilon_c)(\ell_h^2 \rho^2 + i\tau_h \omega + \epsilon_h) + 1}, \tag{4.4}$$

where $\tilde{f}(\rho, \omega)$ denotes the Fourier transform of $f(x, y, t)$; $\rho = \sqrt{(\rho_x^2 + \rho_y^2)}$ is the magnitude of the spatial frequency, and ω is temporal frequency (both are in radians). Here,

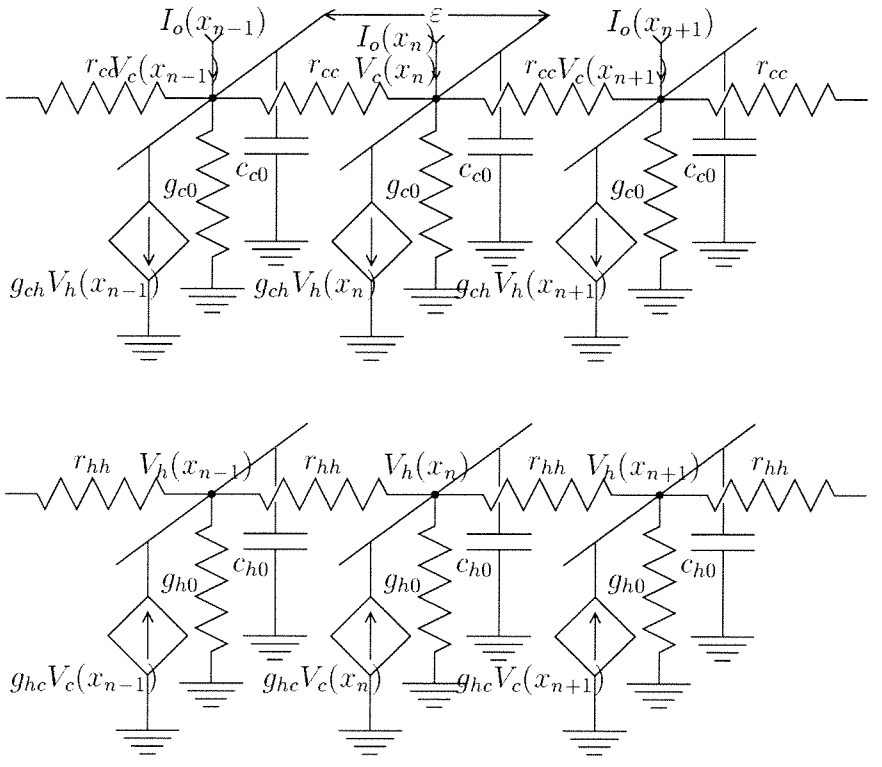


Figure 4.1: PHYSICAL MODEL OF THE OUTER RETINA

The two resistive networks model the cone and horizontal-cell syncytia. The voltages V_c and V_h represent the membrane potentials, and the current I_o represents inputs from the cone outer segment. The diamonds are symbols for current sources controlled by voltages in another part of the circuit; they model chemical synapses. The direction of current flow, indicated by the arrow, is into the network for an excitatory synapse. The membrane capacitances of the cells are included to model dynamic behavior. The parameter ε is a measure of cell sizes; it links the modeled quantities that are in current per unit area, sheet resistance, conductance per unit area, and capacitance per unit area to the physiological ones.

$\tau_c = c_{c0}/g_{ch}$ and $\tau_h = c_{h0}/g_{hc}$ are the time constants of the cells; $\ell_c = (r_{cc}g_{ch})^{-1/2}$ and $\ell_h = (r_{hh}g_{hc})^{-1/2}$ are the space constants of the decoupled syncytia, with transconductances replaced by conductances to ground; and $\epsilon_c = g_{c0}/g_{ch}$; and $\epsilon_h = g_{h0}/g_{hc}$ are the ratios of membrane-leakage conductance to synaptic transconductance. The reciprocal of ϵ_c is equal to the change in voltage that occurs in the cone for a unit change in voltage in the horizontal cell. I call this ratio the voltage gain from the horizontal cell to the cone; the voltage gain from cone to horizontal is defined similarly.

The inputs to the model are currents per unit area, and the responses of the cell are voltages, so the transfer functions have units of resistance times area, or the reciprocal of transcondutance per unit area. To obtain a dimensionless measure of frequency sensitivity, I shall multiply the transfer function by g_{ch}. I define this dimensionless measure as the gain: $\tilde{A}(\rho,\omega) \equiv g_{ch}\tilde{V}/\tilde{I}_o$. That is, the gain is the ratio between the voltage response and the input current when the transconductance g_{ch} is 1 unit.

The transfer functions $\tilde{H}_c(\rho,\omega)$ and $\tilde{H}_h(\rho,\omega)$ give the responses of the cones and the horizontal cells to sinusoidal spatiotemporal patterns, like the one shown in Figure 4.2. The voltage response of the model is given by $\tilde{H}(\rho,\omega)I\sin(\rho_x x + \rho_y y + \omega t)$; it is simply a scaled and shifted version of the signal. The scaling is given by the magnitude of $\tilde{H}$, and the phase shift is given by the argument of $\tilde{H}$. Since these quantities do not depend on the orientation of the grating, the model does not have orientation or direction selectivity.

Any moving image can be expressed as a sum of sinusoidal spatiotemporal patterns. Hence, by using the frequency-response function $\tilde{H}$ to shift and scale each frequency component, we can obtain the model's response to motion. This is our primary motivation for studying the model's spatiotemporal-frequency response.

For illustrative purposes, we use the following set of parameters: $\ell_c = 0.05°; \ell_h = 0.2°; \tau_c = 30\text{ms}; \tau_h = 200\text{ms}; \epsilon_c = 0.3; \epsilon_h = 0.1; g_{ch} = 0.2\text{pA/mV}$. Unless otherwise stated, all model responses shown were obtained with these parameters.

In presenting the results from the model, I shall use typographical conventions to distinguish between model and reality. For example, a cone is a node in the circuit model; whereas a cone is a biological photoreceptor. The cone's response is given by

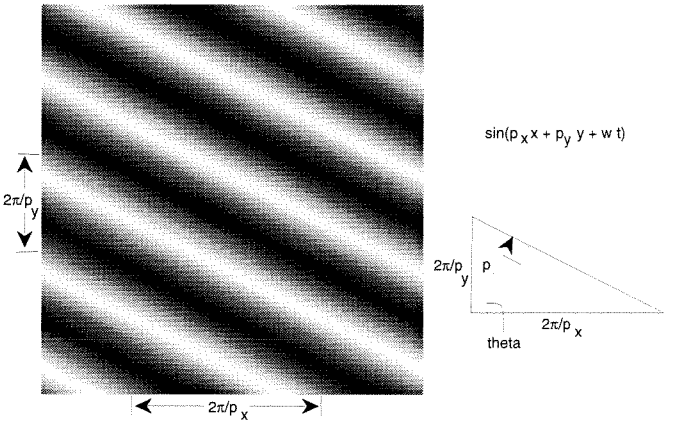
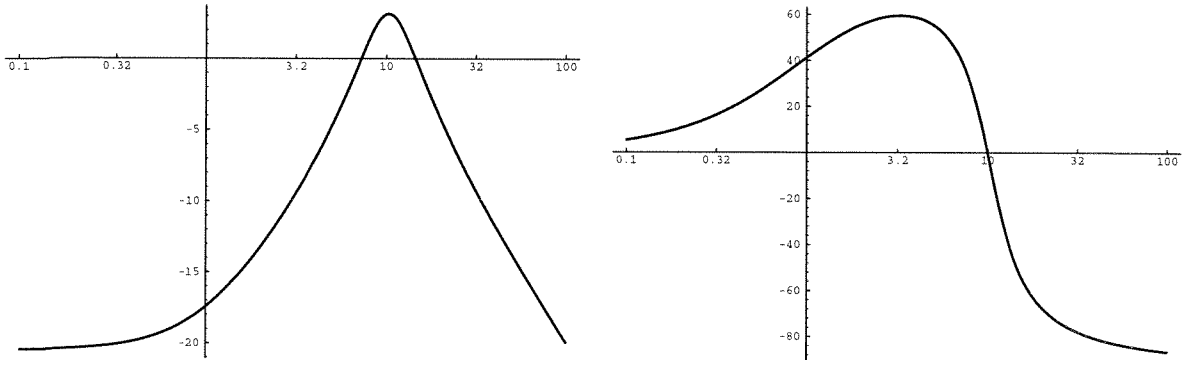


Figure 4.2: SINUSOIDAL SPATIOTEMPORAL SIGNALS

The values of this signal change over time and space according to the expression $I\sin(\rho_x x + \rho_y y + \omega t)$, where I is the peak amplitude. When time is frozen, the signal is just a sinusoidal grating, like the one shown here. The grating's orientation, θ, is given by the direction of the spatial-frequency vector: $\theta = \tan^{-1}(\rho_y/\rho_x)$. The grating's wavelength, λ, is given by the magnitude of the spatial-frequency vector: $\lambda = 2\pi/\sqrt{(\rho_x^2 + \rho_y^2)}$. When time is running, we can track a particular point, with intensity I_p, and find that it appears to move with speed $v = \omega/\sqrt{(\rho_x^2 + \rho_y^2)}$, in a direction opposite to the spatial frequency vector, due to the constraint that $\rho_x x + \rho_y y + \omega t = \sin^{-1}(I_\mathrm{p}/I)$. Because this contraint applies to all points, the whole grating moves with the same velocity. Actually, the motion of such a pattern is ambiguous; for example, moving the grating in the x direction at a higher speed ω/ρ_x will produce the same spatiotemporal pattern. The model's response to such patterns is derived in the text.



(a) (b)

Figure 4.3: SENSITIVITY OF CONES TO FULL-FIELD FLICKER
Amplitude (a) and phase (b) of cone responses to temporal frequency from OPL circuit model. For the parameters values chosen, the cone's response peaks at 10cps, and levels off below 0.33cps.

the voltage of that node, and is the analog of the membrane potential of the cone. I will also plot frequency responses on a logarithmic scale, in dB[3]; spatial frequency in is units of cpd (cycles per degree), and temporal frequency is in units of cps (cycles per second).

4.3 Responses to Flicker and Gratings

Full-field flicker and stationary sinusoidal gratings are used by physiologists and psychophysicists to characterize the temporal and spatial responses of the visual system. In the same vein, I present analytical expressions that describe the model's response to these classic stimuli. I describe the salient features of these responses, and relate them, quantitatively, to the model's parameters. I validate the model by comparing its responses to biological measurements.

[3]20dB is equivalent to a tenfold increase in amplitude.

4.3.1 Full-Field Flicker

If the spatial frequency is sufficiently low (i.e. $\rho \ll \sqrt{\epsilon_c/\ell_c}, \sqrt{\epsilon_h/\ell_h},$) we can drop the spatial-frequency terms and obtain

$$\begin{aligned}\tilde{H}_c(0,\omega) &= \frac{1}{g_{ch}} \frac{i\tau_h\omega + \epsilon_h}{\tau_c\tau_h(i\omega)^2 + (\epsilon_h\tau_c + \epsilon_c\tau_h)i\omega + \epsilon_c\epsilon_h + 1}, \\ \tilde{H}_h(0,\omega) &= \frac{1}{g_{ch}} \frac{1}{\tau_c\tau_h(i\omega)^2 + (\epsilon_h\tau_c + \epsilon_c\tau_h)i\omega + \epsilon_c\epsilon_h + 1}.\end{aligned}$$

These expressions give the sensitivity of cones and horizontal cells to full-field flicker; the magnitude and phase of $\tilde{H}_c(0,\omega)$ are plotted in Figure 4.3.

The cones have a bandpass temporal-frequency response with a distinct peak at $\hat{\omega} \simeq 1/\sqrt{\tau_c\tau_h}$; the response rolls off at 20dB per decade beyond this point. Surprisingly, the peak frequency of the cone is determined not by the cone's individual time constant alone, but rather by the geometric mean of the time constants of the cone and the horizontal cell. The gain reaches a maximum value of $\sqrt{(\tau_h/\tau_c)}Q_t$ at the peak, where

$$Q_t \simeq \left(\epsilon_c\sqrt{\frac{\tau_h}{\tau_c}} + \epsilon_h\sqrt{\frac{\tau_c}{\tau_h}}\right)^{-1}.$$

These expressions for the peak frequencies and the peak gain are based on the approximation $\epsilon_c\epsilon_h \ll 1$. The cone's response levels off for frequencies below ϵ_h/τ_h. Decreasing ϵ_h, which increases the voltage gain from the cone to the horizontal cell, moves this point to lower frequencies, attenuating the low spatial frequencies further.

The phase is initially 0, rises gradually as the temporal frequency increases, reaches a maximum, and then decreases rapidly around the peak temporal frequency. The phase passes through 0 at the peak frequency $\hat{\omega}$, becomes negative, and approaches $-90°$.

Compare the flicker response of the model to the flicker-sensitivity curves of humans and cats, obtained from psychophysical and physiological measurements, shown in Figure 3.4a and Figure 3.6b,c, respectively. The model shows the same bandpass characteristic observed for high intensities and diffuse patterns in the human measurements and in the X-cell and Y-cell cat measurements.

The main differences between the model and the biology is that the model reproduces neither the steepness of the high-frequency cutoff, nor the rapid phase changes that occur there—cat X cells go out to 600° [93]. Kelly fitted this steep cutoff with a model of the long thin process that carries signals from the outer segment of the cone to the cone terminal; he used a continuous, passive, lossy cable model that produced an exponential rolloff [111]. Chen and Freeman fitted the sharp cutoff in Frishman's X-cell data with a model of the phototransduction process that consisted of a cascade of eight first-order reactions [109]. These high-order cascades produced a steep cutoff and introduced a large delay, which resulted in rapid phase changes. In fact, Frishman and her colleagues showed that their phase measurements were approximated quite well with a pure delay of 24ms for the X cell and 20ms for the Y cell.

The simple model that I analyzed includes neither the phototransduction process nor the cable properties of the cone. Such a simple model cannot be expected to reproduce the biological responses exactly. My goal is to capture the bandpass character of the biological responses, and the model meets that goal. When the contribution of the phototransduction cascade to the cutoff and the phase is removed, the residual gain and phase shift look much like those of this simple model (see plots of contributions of individual stages in Chen and Freeman's more elaborate model [109]).

The model reproduces the rise in sensitivity before the peak, as do the models proposed by Kelly and by Chen and colleagues. These models produce this behavior by placing lowpass filters in negative-feedback loops around lowpass feedforward stages (the feedback filters must have a frequency cutoff that is lower than that of the feedforward filters), just like this model does. Alternatively, the rise in sensitivity before the peak may be obtained by introducing a parallel pathway with a lower-frequency cutoff and subtracting that pathway's output from the main feedforward pathway. These two architectures are called **feedback** and **feedforward**, respectively.

Kelly used two to four feedback loops to fit the data in the region where the gain is increasing; the number of loops needed increased with intensity [111]. On the other hand, Chen and Freeman tried to fit both the feedforward and the feedback models to the cat X-cell data, and found that the latter provided a much closer fit [109].

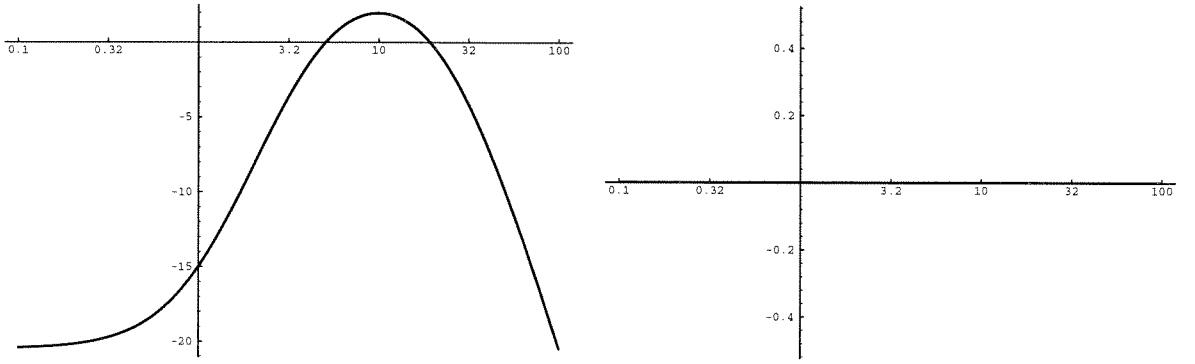


Figure 4.4: SENSITIVITY OF CONES TO STATIONARY GRATINGS
Amplitude (a) and phase (b) responses of cones to spatial frequency from OPL circuit model. For the parameter values chosen, the cone's response peaks at 10cpd, and levels off below 1.6cpd.

Only the feedback configuration could produce the sharp resonant peak evident in the full-field flicker responses; the gain at the peak is five or six times higher than a feedforward scheme predicts. However, physiological measurements in monkeys [96], and psychophysical measurements in humans [111], do not show such high resonances and are fitted by feedforward models well. The feedback model was also the only one of eight configurations studied by Chen and Freeman that satisfied the restrictions placed on the relative strengths and the relative delays between the center and the surround by the experimental measurements [109].

In summary, a simple linear two-layer feedback model:

- Accounts for the bandpass responses to temporal frequency observed for high intensities and diffuse patterns in human psychophysics and cat physiology.
- But it does not reproduce the steepness of the high-frequency cutoff, nor does it reproduce the large phase accumulation.
- This shortcoming is, most likely, because the model does not include the cable properties of the photoreceptor neurites, nor does it include the cascade of chemical reactions involved in phototransduction.

4.3.2 Stationary Gratings

If the temporal frequency is sufficiently low (i.e. $\omega \ll \epsilon_c/\tau_c,\ \epsilon_h/\tau_h$), we can drop the temporal frequency terms and obtain

$$\begin{aligned}\tilde{H}_c(\rho,0) &= \frac{1}{g_{ch}} \frac{\ell_h^2\rho^2 + \epsilon_h}{\ell_c^2\ell_h^2\rho^4 + (\epsilon_c\ell_h^2 + \epsilon_h\ell_c^2)\rho^2 + \epsilon_c\epsilon_h + 1}, \\ \tilde{H}_h(\rho,0) &= \frac{1}{g_{ch}} \frac{1}{\ell_c^2\ell_h^2\rho^4 + (\epsilon_c\ell_h^2 + \epsilon_h\ell_c^2)\rho^2 + \epsilon_c\epsilon_h + 1}.\end{aligned}$$

These expressions give the sensitivities to stationary gratings; the magnitude and phase of $\tilde{H}_c(\rho,0)$ are plotted in Figure 4.4.

The spatial-frequency responses parallel the temporal ones; the cones also have a bandpass spatial response. However, the amplitude of the spatial responses rises twice as fast as does the temporal response, on a log–log plot, and rolls off twice as fast as well. Another difference is that the phase of the spatial response never deviates from 0.

The cone's response peaks at $\hat{\rho} \simeq \sqrt{(1 - \epsilon_h\ell_c/\ell_h)}\sqrt{(\ell_c\ell_h)}$, attaining a maximum gain of $(\ell_h/\ell_c)Q_x$, where

$$Q_x \simeq \left(2 + \epsilon_c\frac{\ell_h}{\ell_c} - \epsilon_h\frac{\ell_c}{\ell_h}\right)^{-1}.$$

Again, I made the approximation $\epsilon_c\epsilon_h \ll 1$. In close analogy to the temporal behavior, the peak frequency is determined by the geometric mean of the space constants of the decoupled syncytia. The cone response levels off for frequencies below $\sqrt{\epsilon_h/\ell_h}$.

Compare the grating response of the model to the grating-sensitivity curves of humans and cats, obtained from psychophysical and physiological measurements, shown in Figure 3.4b and Figure 3.6a, respectively. The OPL model shows the same bandpass characteristic observed for high intensities and low temporal modulation in the human measurements and the cat X-cell measurements.

Again, the main difference between the biological responses and the model's are the model's inability to produce steep rolloff. Kelly and other researchers used an

exponentially weighted function of the form $\rho^2 e^{-\rho}$ to fit the steep rolloff found in psychophysical measurements [112, 113, 114]. This choice is consistent with Rodieck's difference-of-Gaussians (DoG) model for the receptive field center and surround, which also results in an exponential rolloff with frequency. The DoG model fits Frishman's cat ganglion-cell measurements perfectly. More detailed models, based on retinal anatomy, have shown that the gaussian-like spatial profile of the receptive-field center arises from spatial summation by the bipolar-cell dendrites [115]. Smith showed that the gaussian-like spatial profile of the receptive-field surround arises from the presence of two types of horizontal cells [115].

The simple model that I analyzed does not include bipolar convergence, and it has only one type of horizontal cell. Nevertheless, the model captures the bandpass character of the biological responses, and reproduces the increase in sensitivity before the peak, much like the other models.

For the spatial behavior, Kelly found that a feedforward model with a single stage in the parallel inhibitory pathway could account for increasing gain; the contribution from the inhibitory pathway falls off as intensity decreases [112]. At extremely low intensity levels, the response becomes lowpass and can be fitted with the exponential cut-off function over the entire frequency range [113, 114]. Thus, the trends for spatial and temporal frequencies are identical: The sensitivity to higher frequencies, and the peak contrast gain, both increase with intensity. On the other hand, Chen and Freeman used a single feedback stage to fit the spatial responses of cat X cells measured at high intensity, just as this model does.

In summary, a simple linear two-layer feedback model:

- Accounts for the bandpass responses to spatial frequency observed for high intensities and slow temporal modulation in human psychophysics and cat physiology.

- But it does not reproduce the steepness of the high-frequency cutoff,

- This shortcming is, most likely, because the model does not include spatial summation in the bipolar cell dendrites, nor does it include a second horizontal

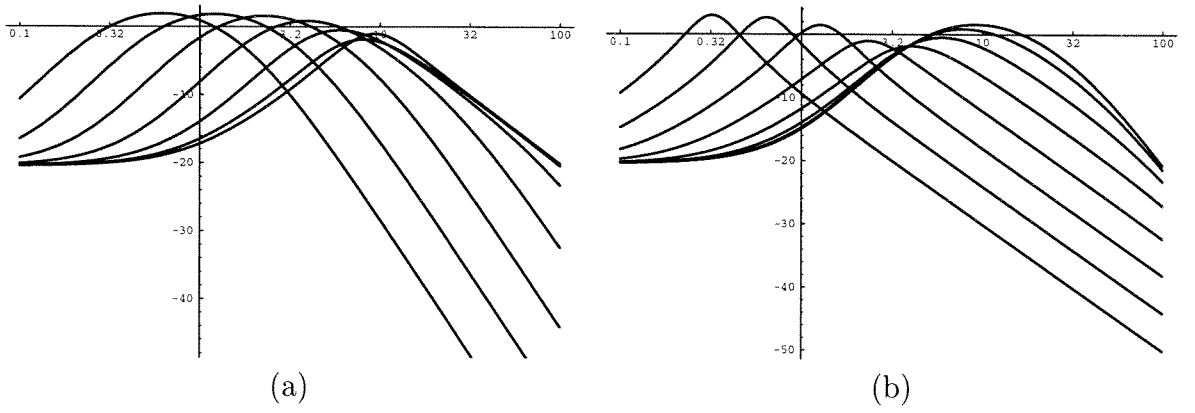


Figure 4.5: SENSITIVITY OF CONES TO MOVING GRATINGS
Response to gratings moving at various speeds, plotted versus (a) temporal frequency or (b) spatial frequency. The low-speed temporal-frequency response in (a) looks like the stationary-grating frequency response, and the high-speed spatial-frequency response in (b) looks like the flicker-frequency response. At speeds below $\hat{\omega}/\hat{\rho} = 1$dps, the peak in the temporal-frequency response tracks the speed; at speeds below 1dps, the peak in the spatial-frequency response tracks the speed. As the curves shift, their shape remain the same.

cell network.

4.3.3 Moving Gratings

A grating with spatial frequency ρ produces the temporal frequency $\omega = v\rho$ when it moves with velocity v in the direction of its orientation. Therefore, it is easy to predict the response to such a stimulus. We simply substitute $v\rho$ for ω, and evaluate $\tilde{H}(\rho, v\rho)$; or, we substitute ω/v for ρ, and evaluate $\tilde{H}(\omega/v, \omega)$. The resulting expression tells us how spatial filtering and temporal filtering, respectively, depend on speed. In fact, we can draw salient conclusions without doing any algebra.

For slow speeds, the temporal frequencies, $v\rho$, produced by the motion are low. Hence, the temporal terms drop out, and the response is identical to that for stationary gratings, $\tilde{H}_c(\rho, 0)$, and does not depend on speed. However, if we plot the response versus temporal frequency (i.e., $\tilde{H}_c(\omega/v, 0)$), we find that the response has the same shape as the grating-sensitivity curve, but shifts to higher temporal frequen-

cies as the speed increases. In particular, at each speed, v, the peak response occurs at the temporal frequency $v\hat{\rho}$.

On the other hand, for fast speeds, the spatial frequencies, ω/v, produced by the motion are small. Hence, the spatial terms drop out, and the response has the same shape as the flicker curve, $\tilde{H}_c(0,\omega)$, and does not depend on speed. However, if we plot the response versus spatial frequency (i.e. $\tilde{H}_c(0,\rho v)$), we find that the response has the same shape as the flicker sensitivity curve, but shifts to lower spatial frequencies as the speed increases. In particular, at each speed, v, the peak response occurs at the spatial frequency $\hat{\omega}/v$.

This behavior holds for a whole class of the spatiotemporal filters, since my argument does not depend on the detailed form of the transfer function $\tilde{H}_c(\rho,\omega)$. The argument works whenever the spatial- and temporal-frequency terms become negligible at low frequencies. Consequently, for all spatiotemporal filters that satisfy this requirement, we can state the following general results:

- As speed decreases, the shape of the temporal frequency sensitivity curve asymptotically approaches that for spatial frequency, but it shifts to proportionately lower temporal frequencies.

- As speed increases, the shape of the spatial frequency sensitivity curve asymptotically approaches that for temporal frequency, but it shifts to proportionately lower spatial frequencies.

We confirm these conclusions by computing and plotting the spatial- and temporal-frequency responses for gratings moving at various speeds, as shown in Figure 4.5.

Compare these moving-grating responses to the psychophysical measurements from humans shown in Figure 4.6a. The model reproduces the dependence of the peak frequency on speed. Kelly fitted the horizontal displacement of the peak with speed, over the range from 0.15dps to 32dps, using the expression $\hat{\rho} = 7.3/(v+2)$. This quantitative relation reveals that the peak spatial frequency is indeed inversely proportional to speed for high speeds, as predicted by the model. This OPL circuit model does not account for the dependence of the peak height on speed; additional

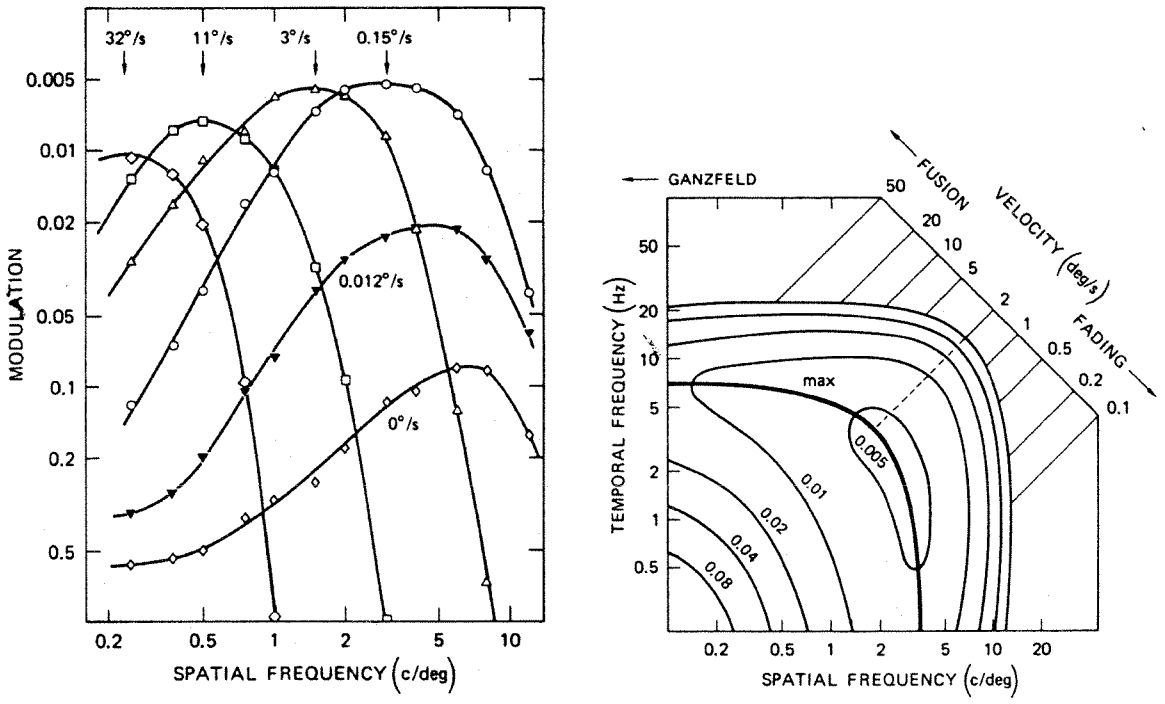


Figure 4.6: HUMAN SENSITIVITY TO MOTION

(a) Response to sinusoidal gratings, moving at six different speeds, versus spatial frequency. The relative amplitude of the modulation of the baseline intensity (i.e., threshold contrast) is plotted; the baseline intensity was 300td. The subject's eye movements where tracked and compensated for in these experiments. The response was always bandpass, and its peak remained at about 5 cpd for speeds below 2 dps. At higher speeds, the peak position shifted to lower spatial frequencies. The peak amplitude also changed, rising rapidly initially, and then falling slowly above 2 dps. Reproduced from [92]. (b) Contour plot of spatiotemporal-contrast–threshold surface (same as in Figure 3.5b). Sensitivity doubles from one contour to the next. The heavy line represents the maximum sensitivity at each velocity; a velocity axis is included on the upper right. The surface is roughly symmetric about the line $v = 2$dps. Reproduced from [92].

spatial and temporal filtering in the inner retina could account for the dependence of the height of the peak on speed.

The model also predicts that the peak temporal frequency is proportional to speed for low speeds. This behavior is also evident in the psychophysical data (Figure 4.6), although Kelly did not plot his data versus temporal frequency. The curves for 0dps, 0.012dps, and 0.15dps peak at about 6cpd, 4cpd, and 3cpd, respectively. We obtain the temporal frequencies produced by these moving gratings by multiplying spatial frequency by speed, which gives 0cps, 0.048cps, and 0.45cps, respectively. Hence, the peak temporal frequency is roughly proportional to speed.

In summary, a simple linear two-layer feedback model:

- Accounts for the dependence of the peak spatial and temporal frequencies on speed observed for high and low speeds, respectively, in human psychophysics.
- But it does not reproduce the dependence of the peak height on speed.
- This shortcoming is, most likely, because the model does not include highpass temporal and spatial filtering in the inner retina.

This description of the locus of the peak position completes my discussion of the model's sensitivity to stimuli used in classic psychophysical and physiological experiments. To understand exactly how these responses arise, and to extend the description to arbitrary dynamic patterns, we must turn to the complete three-dimensional spatiotemporal transfer function.

4.4 Spatiotemporal Sensitivity

Plots of the magnitude and phase of the cone's spatiotemporal-frequency transfer function, $\tilde{H}_c(\rho, \omega)$, are shown in Figure 4.7. The function is more or less symmetric about the 45° degree axis, because interchanging ρ^2 and $i\omega$ in Equations 4.3 and 4.4 produces homomorphic equations. Consequently, everything that we say about spatial frequency with respect to temporal frequency is still true when the two

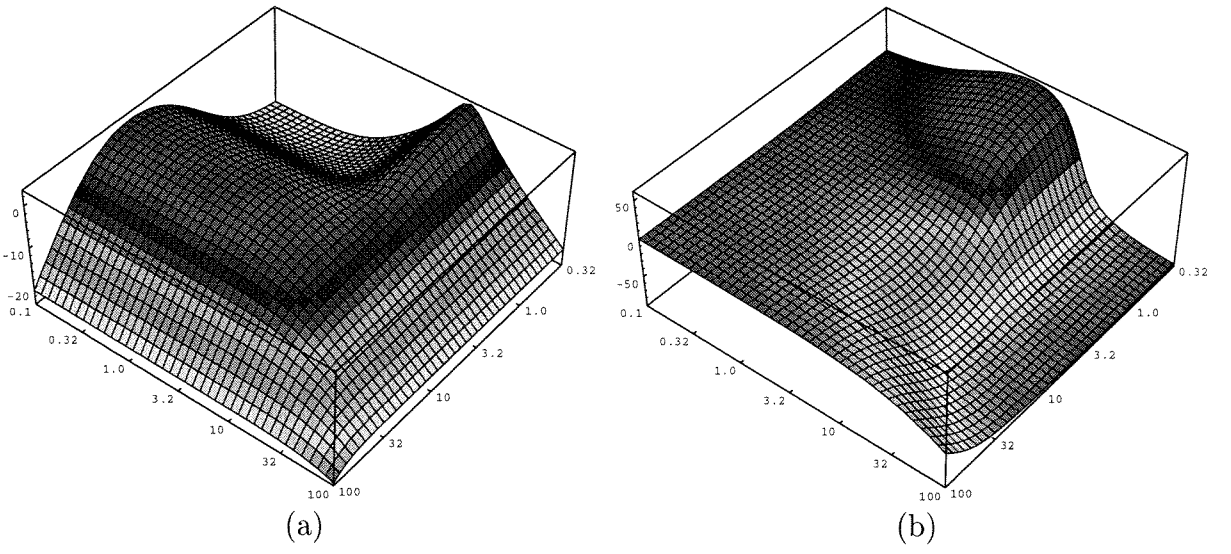


Figure 4.7: SPATIOTEMPORAL SENSITIVITY OF CONES
Three-dimensional plots showing (a) magnitude and (b) phase of cone responses from OPL circuit model versus spatial and temporal frequency. At higher spatial frequencies, the low-frequency temporal sensitivity increases, and vice versa. The phase is positive for low temporal frequencies, negative for high temporal frequencies and tends to 0 at high spatial frequencies.

are interchanged. Bear in mind this duality as we discuss the salient features of the spatiotemporal frequency responses.

I begin the discussion by taking the cross-sections of these surfaces that correspond to the flicker and grating curves presented in Figure 4.3a,b and Figure 4.4a,b. The flicker and grating curves are defined by the intersections of the spatiotemporal surface with the $\rho = 0$ and $\omega = 0$ planes, respectively. As expected, the amplitude response (Figure 4.7a) is bandpass at these planes.

As we move the spatial-frequency plane away from the $\rho = 0$ plane, to higher spatial frequencies, we observe increasingly strong responses to low temporal frequencies. Similarly, as we move the temporal-frequency plane away from the $\omega = 0$ plane, to higher temporal frequencies, we observe increasingly strong responses to low spatial frequencies. Thus, the filters for temporal frequency and for spatial frequency are letting through more low-frequency energy—they are becoming less bandpasslike and more lowpasslike. When the planes reach the peak temporal frequency, $\hat{\omega}$, and the

peak spatial frequency, $\hat{\rho}$, the transition is complete, and the filters become purely lowpass, without any peak whatsoever. Both filters remain lowpass as the planes move beyond the peak frequencies.

The model's spatiotemporal frequency sensitivity mirrors that for humans, shown as a family of curves in Figure 3.5a, as a three-dimensional plot in Figure 3.5b, and as a contour plot in Figure 4.6b; and mirrors that of cats, shown as a family of curves in Figure 3.6b. In particular, this simple linear two-layer feedback model shares four salient features with human psychophysics and cat physiology:

1. Spatial filtering is bandpass at low temporal frequencies, and is tuned to a particular spatial frequency $\hat{\rho}$.

2. Temporal filtering is bandpass at low spatial frequencies, and is tuned to a particular temporal frequency $\hat{\omega}$.

3. Spatial filtering becomes lowpass at high temporal frequencies, and tuning disappears completely when the temporal frequency exceeds $\hat{\omega}$.

4. Temporal filtering becomes lowpass at high spatial frequencies, and tuning disappears completely when the spatial frequency exceeds $\hat{\rho}$

So far, we have discussed how the model's transfer function modulates the amplitude of the input spatiotemporal sinusoid. Let us now consider the phase shift introduced by the transfer function.

As usual, the flicker and grating curves are defined by the intersections of the spatiotemporal surface with the $\rho = 0$ and $\omega = 0$ planes, respectively (see Figure 4.7). At the $\rho = 0$ plane, the response leads for frequencies below the peak temporal frequency, $\hat{\omega}$; it lags for frequencies above $\hat{\omega}$; and the phase decreases rapidly around $\hat{\omega}$, passing through 0 at $\hat{\omega}$. The behavior at the $\omega = 0$ plane also is as expected, with a phase shift of zero.

As we move the spatial-frequency plane away from the $\rho = 0$ plane, to higher spatial frequencies, the phase lead decreases across the entire $\omega < \hat{\omega}$ region, going towards 0—and even changes to a small phase lag over a small subregion near $(\hat{\rho}, \hat{\omega})$

The phase lag also decreases across the entire $\omega > \hat{\omega}$ region, going towards 0. Both transitions appear progressively, occuring at different spatial frequencies for different temporal frequencies—the point at which the leads and lags disappear is roughly proportional to the temporal frequency.

Rapid changes in phase occur roughly along a horizontal line defined by $\omega = \hat{\omega}$, and along a diagonal line defined $\omega = (\hat{\omega}/\hat{\rho})\rho$. (The contour plot in Figure 4.8b shows this clearly.) Taken together, these lines divide the phase plot into three distinct regions. For frequencies below the diagonal line, the phase is close to 0. For frequencies above the diagonal line and below the horizontal line, there is a large phase lead. And for frequencies above both lines, there is a large phase lag.

In summary, the model's transfer function, with respect to phase shift, has three salient features:

1. There is no phase shift when the spatial frequency is above $(\omega/\hat{\omega})\hat{\rho}$ or the temporal frequency is below $(\rho/\hat{\rho})\hat{\omega}$.

2. A large phase lead occurs when the temporal frequency is below $\hat{\omega}$ and the spatial frequency is below $(\omega/\hat{\omega})\hat{\rho}$.

3. A large phase lag occurs when the temporal frequency is above $\hat{\omega}$ and the spatial frequency is below $(\omega/\hat{\omega})\hat{\rho}$.

Unfortunately, the detailed phase characteristics are not available for the biological systems, so we cannot make a comparison.

A theme that unifies the amplitude and phase characteristics is the dependence of spatial filtering on temporal frequency, and vice versa. In this model, spatial filtering cannot be separated from temporal filtering: $\tilde{H}_c(\rho, \omega) \neq H_x(\rho)H_t(\omega)$. That is, we cannot realize the filtering performed by the model by cascading a spatial filter, $H_x(\rho)$, with a temporal filter, $H_t(\omega)$. This **spatiotemporal inseparability** arises because the same elements in the circuit are used to perform both spatial filtering and temporal filtering. A purely spatial filter cannot have any time dependencies in its wires; a purely temporal filter cannot have any crosstalk with its neighbors. The

OPL model adheres to neither of these edicts: Signals are spread out in time as they are spread out in space—and vice versa—as they are passed from place to place by the internode conductances and as they are passed from time to time by the node capacitances.

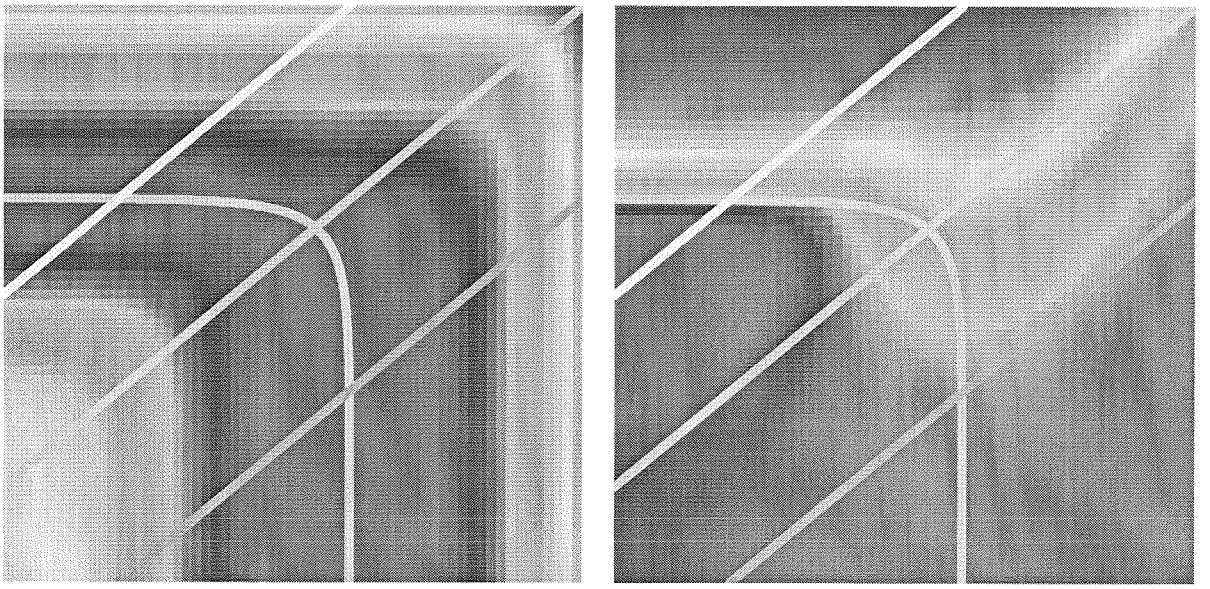
4.5 Responses to Moving Images

To understand how the OPL model responds to motion, we will find it most instructive to display the three-dimensional spatiotemporal-frequency transfer function as a contour plot, and to superimpose the input spectrum ont this plot. The transfer function is replotted in this fashion in Figure 4.8; observe the simliarity between this plot and the contour plot of human spatiotemporal contrast sensitivity in Figure 4.6b.

The speed $\hat{v} \equiv \hat{\omega}/\hat{\rho}$, given by the ratio between the peak temporal frequency and the peak spatial frequency, plays a decisive role in the model. I call this speed the **pivotal speed**, because it demarcates the border between the low-speed region, where motion produces higher temporal frequencies, and the high-speed region, where motion produces lower spatial frequencies. These two distinct behaviors arise because the line $\omega = \rho\hat{v}$ bisects the L-shaped ridge of the amplitude plot into two arms, one running horizontally and the other running vertically; the ridges takes the corner right at the pivotal speed line, $\omega = \rho\hat{v}$.

At speeds below $\hat{v}$, the spectrum intersects the vertical arm, and the locus of the peak remains at the same spatial frequency $\hat{\rho}$, but it moves to higher temporal frequencies $\hat{\rho}v$ with increasing speed, v. At speeds above $\hat{v}$, the spectrum intersects the horizontal arm, and the locus of the peak remains at the same temporal frequency $\hat{\omega}$, but it moves to lower spatial frequencies $\hat{\omega}/v$ with increasing speed, v. This picture explains the responses obtained for moving gratings plotted in Figure 4.5a,b.

We can derive the locus of the peak by differentiating $|\tilde{H}_c(\rho, \rho v)|$ with respect to ρ and setting the derivative to 0 to find the maximum. However, a simple approximate



(a) (b)

Figure 4.8: Motion and Spatiotemporal Sensitivity of Cones
Color-coded contour plots showing amplitude (a) and phase (b) of the cone responses from the OPL circuit model. A cyclic color encoding was used, starting with red at the bottom end of the scale, and coming back to red at the top. (a) The amplitude plot looks like a mountain range with a sharp bend in it; the L-shaped ridge that runs along the top is shown. (b) The phase plot looks like a plain in the area below the $\omega = \rho\hat{v}$ line, where $\hat{v}$ is the ratio between the peak temporal frequency, $\hat{\omega}$, and the peak spatial frequency, $\hat{\rho}$. The area above this line is shared by a mountain and a valley, with the $\omega = \hat{\omega}$ line demarcating the border between them. Each diagonal bar in these plots is the support of the input image's spectrum, for uniform translation at a different speeds. The support is defined by the line $\omega = \rho v$, where v is the speed. On $\log(\omega)$-$\log(\rho)$ scales, this line is always at an angle of 45°; it shifts to higher temporal frequencies as speed increases. The three speeds shown are $0.1\hat{v}$, $\hat{v}$, and $10\hat{v}$. The intersection of these diagonal lines with the spatiotemporal surface gives the sensitivity of the model to moving gratings, and produces the curves plotted in Figure 4.5a,b, when projected onto the temporal-frequency axis or onto the spatial-frequency axis, respectively.

expression describes the locus of the peak, $(\hat{\rho}(v), \hat{\omega}(v))$, quite well.

$$\frac{\hat{\rho}(v)}{\hat{\rho}(0)} + \frac{\hat{\omega}(v)}{\hat{\omega}(\infty)} = 1,$$

where $\hat{\rho}(0) = \hat{\rho}$ is the peak spatial frequency for the stationary-grating sensitivity curve, and $\hat{\omega}(\infty) = \hat{\omega}$ is the peak temporal frequency for the full-field–flicker sensitivity curve. This expression accounts for the peak position in both the high-speed and the low-speed regions. To locate the peak on the temporal-frequency axis, we replace $\hat{\rho}(v)$ with $\hat{\omega}(v)/v$; to locate the peak on the spatial-frequency axis, we replace $\hat{\omega}(v)$ with $\hat{\rho}(v)v$. Making these substitutions gives

$$\begin{aligned} \hat{\omega}(v) &= \frac{v}{\hat{v} + v}\hat{\omega}(\infty), \\ \hat{\rho}(v) &= \frac{\hat{v}}{\hat{v} + v}\hat{\rho}(0), \end{aligned}$$

where $\hat{v} \equiv \hat{\omega}(\infty)/\hat{\rho}(0)$ is the speed at which the behavior crosses over from the low-speed regime to the high-speed regime.

With the aid of these contour plots, it is easy to see how the model's spatiotemporal inseparability shapes the model's response to moving gratings. The amplitude of the response to a grating changes when that grating flickers or moves, because the gain of the spatial filter depends on temporal frequency. For temporal frequencies below the peak temporal frequency, $\hat{\omega}$, this dependence makes the response increase with flicker rate or with speed—except when the spatial frequency of the grating is equal to the peak spatial frequency, $\hat{\rho}$. For this exceptional situation, where the spatial filter is tuned to the spatial frequency of the grating, the response becomes speed invariant for temporal frequencies below $\hat{\omega}$ (i.e., speeds below $\hat{v}$). In general, however, the response of the OPL model is not speed invariant.

The phase of the grating response also changes drastically with increasing speed, and the change is nonmonotonic. The phase starts increasing after the speed exceeds the pivotal speed, $\hat{v}$, reaches a maximum, and then starts to decrease, reaching 0 when the speed is equal to $\hat{\omega}/\rho$, where ρ is the spatial frequency of the grating. However,

when the spatial frequency of the grating is equal to the peak spatial frequency, $\rho = \hat{\rho}$, the phase does not change with speed—except for a small range around $\hat{v}$, where a small lag occurs.

4.5.1 Speed-Invariant Contrast Estimation

The model predicts that the amplitude and phase of the outer retina's response depends on the speed of the moving grating. Speed-dependent responses give rise to the question: Can a speed-invariant estimate of contrast be obtained from the outer retina's output signals?

If the input pattern has a broad spatial-frequency spectrum (e.g. an impulse, an edge, or a random-dot pattern), we can get a flicker- or speed-invariant estimate of contrast by measuring the energy at the spatial frequency to which the spatial filter is tuned. Since the response phase does not change much at this spatial frequency, we can get away with just measuring the peak amplitude. However, this strategy works in only the low-speed regime, where the temporal frequencies generated are below the cutoff point. In the high-speed region, we can obtain a speed invariant estimate of contrast by taking the dual approach.

Due to the dual relationship between spatial filtering and temporal filtering, the response is also speed-invariant when the temporal filter is tuned to the temporal frequency of the spatiotemporal sinusoid, and the spatial frequency is below $\hat{\rho}$ (i.e., speeds above $\hat{v}$)). Thus, for a broadband signal, we can get a flicker- or speed-invariant estimate of contrast in the high-speed region, $v \gg \hat{v}$, by measuring the energy at the temporal frequency, $\hat{\omega}$. Again, since the response phase does not change at this temporal frequency, we can get away with just measuring the peak amplitude.

We should be aware that the circuit generates the energy we are measuring by amplifying energy at the spatial frequency $\rho = \hat{\omega}/v$. Therefore, we extract energy from lower spatial frequencies as speed increases, so the response is speed invariant only when the input energy is distributed uniformly across the spectrum.

In summary, we can obtain a speed-invariant estimate of contrast from the cone's

output by proceeding as follows:

- For low speeds, $v \ll \hat{v}$, measure the energy at the spatial frequency, $\hat{\rho}$.
- For high speeds, $v \gg \hat{v}$, measure the energy at the temporal frequency, $\hat{\omega}$.

This strategy assumes that the input energy is distributed uniformly across spatial frequency. Since natural images, and edges, have a $1/\rho^2$ power spectrum, it will be smarter to tailor the algorithm to such a colored spectrum. Additional filtering in the inner retina could achieve this optimization.

To achieve speed-invariance in the transition region between the low-speed regime and the high-speed regime, we must match the peak flicker sensitivity, $\tilde{H}_c(0, \hat{\omega}) = \sqrt{(\tau_h/\tau_c)}Q_t$, to the peak grating sensitivity, $\tilde{H}_c(\hat{\rho}, 0) = (\ell_h/\ell_c)Q_x$. Equating these two expressions, and neglecting the term $\epsilon_h(\ell_c/\ell_h)^2$, we find that

$$\frac{\ell_c}{\ell_h} = \frac{\epsilon_h}{2}\frac{\tau_c}{\tau_h}. \tag{4.5}$$

As we want ϵ_h to be small, to attenuate low frequencies, we must make the space constants of the cone and horizontal cell syncytia disparate, and the time constants of the cone and horizontal cell similar, to satisfy this constraint.

4.5.2 Contrast-Invariant Speed Estimation

I now turn to the question of how to estimate the speed of the motion from the outer-retina's response. This computation is relatively straightforward when the input energy is distributed uniformly across frequency. In this case, the distribution of energy in the output is determined entirely by the intersection of the support of the input spectrum with the model's spatiotemporal-sensitivity surface. Hence, it is easy to see how the motion of such a broadband stimulus is encoded by the model.

The strongest spatial- and temporal-frequency components in the output are selected by the bandpass filtering performed by the model. For low speeds, $v \ll \hat{v}$, a particular spatial frequency, $\hat{\rho}$, is selected, and the energy shifts to higher temporal frequencies, $\omega = \hat{\rho}v$, with increasing speed. For high speeds, $v \gg \hat{v}$, a particular

temporal frequency, $\hat{\omega}$, is selected, and the energy shifts to lower spatial frequencies, $\rho = \hat{\omega}/v$, with increasing speed.

In summary, the model predicts that we can obtain a contrast-invariant estimate of speed from the outer retina's response by proceeding as follows:

- For low speeds, $v \ll \hat{v}$, determine which temporal frequency, $\omega_{\max}$, has the most energy, and compute $v = \omega_{\max}/\hat{\rho}$.
- For high speeds, $v \gg \hat{v}$, determine which spatial frequency, $\rho_{\max}$, has the most energy, and compute $v = \hat{\omega}/\rho_{\max}$.

The most economical way to implement this algorithm would be to use just two broadly-tuned bandpass filters, one tuned to the low end of the range and the other tuned to the high end, and interpolate between these two filters to determine the frequency of the input signal. This two channels may correspond to the magno and parvo pathways [116].

This algorithm for computing speed will not work for an image with a $1/\rho^2$ power spectrum because the bandpass filter is intentionally designed to whiten such a spectrum, and it equalizes the energy for all frequencies in its passband. To make the spatial or temporal frequency tuned in by the outer retina's spatiotemporal bandpass stand out, we may use a highpass temporal filter or a highpass spatial filter to flatten such natural spectra. This strategy may be used by retina, since both of these highpass filtering operations occur in the inner retina.

4.5.3 Space–Time Effects

I bring my discussion about the outer-retina's motion responses to a close by leaving the frequency domain to take a look at the response to a moving edge in space–time. The pertinent question is: How does the simple intuitive picture that we have painted in frequency coordinates translate into space–time coordinates?

The response of the model to a moving edge is shown in Figure 4.9, for five different speeds—quarter, half, once, twice, and four times the pivotal speed. I obtained

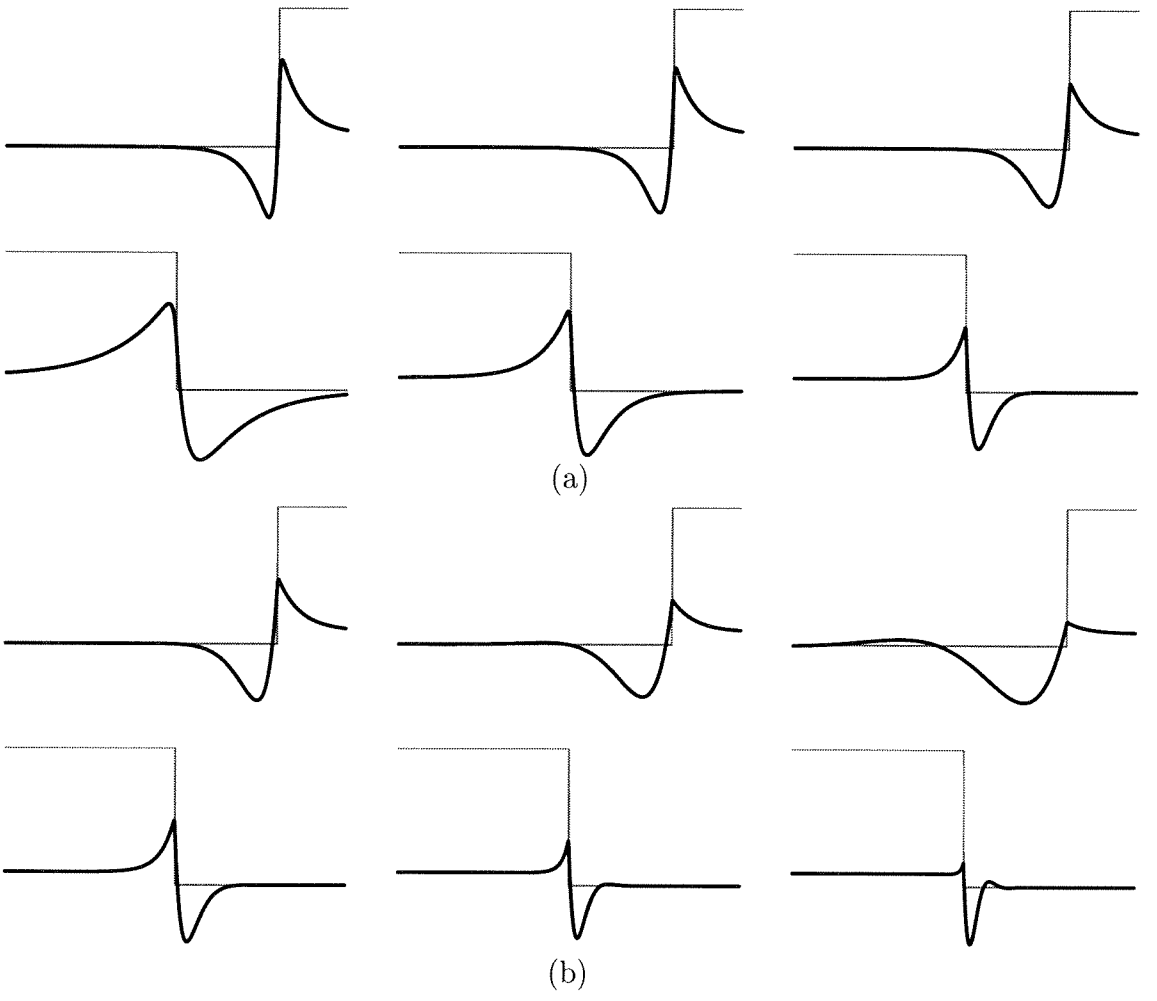


Figure 4.9: Response of Cone to Moving Edges

Each panel shows the responses of cones (bold line) to step edges (thin line) moving at three different speeds, obtained from the outer-retina circuit model: (a) Speeds equal to $0.25\hat{v}$, $0.5\hat{v}$, and $\hat{v}$. (b) Speeds equal to $\hat{v}$, $2\hat{v}$, and $4\hat{v}$. The input goes from 1 (white) to 0 (black) as the black region invades the white region. In each row, the edge's speed doubles from one graph to the next, going from left to right. In each panel, the top row shows responses plotted versus space, at a particular point in time, and the bottom row shows responses plotted versus time, at a particular location in space. In space coordinates, the edge moves to the right, and in time coordinates, the edge moves to the left.

analytical expressions for these responses by taking the inverse-Fourier transform of the model's frequency-transfer functions. I evaluated and plotted these expressions for a particular choice of parameter values. No antialiasing was performed, so these responses contain energy at all frequencies, out to infinity.

As the speed increases, the response transforms from the spatial response to a static edge, which consists of an overshoot and an undershoot on either side of the edge, to the temporal response to a step input, which consists of an exponentially damped sinusoid that starts after the step occurs. By the time the speed changes by a factor of eight, a complete transformation has occured, and the response changes from a perfectly symmetric spatial response to a completely assymetric temporal response.

Due to causality, the temporal component of the response always trails the edge, occuring to the right of the edge in temporal coordinates, or to the left of the edge in spatial coordinates. Only the spatial component of the response can preceed the edge; this effect occurs via long range transmission through the tightly-coupled horizontal-cell network. As the horizontal-cell network produces inhibition, the spatial signal gives rise to the overshoot to the right of the edge in space coordinates, or to the left of the edge in time coordinates. It takes time for signals to propagate through the network, and these inhibitory signals may be overtaken by excitatory signals from the short-range cone network if the edge moves fast enough.

Excitation overtakes inhibition when the speed exceeds the pivotal speed. At the pivotal speed, the time it takes for the edge to transverse the receptive filed equals the time it takes for the cone–horizontal-cell feedback loop to settle. Therefore, for speeds below the pivotal speed, the system settles, and the response looks like that to a static edge—there is no evidence of temporal behavior. Whereas, for speeds above the pivotal speed the system starts to respond after the edge has passed by, and the response looks like that to a full-field flash—there is no evidence of spatial behavior.

Consequently, two distinct behaviors are observed above and below the pivotal speed. Below the pivotal speed, the response is invariant with speed, when plotted versus position, whereas it is increasingly compressed when plotted versus time (See Figure 4.9a). Hence, the frequency responses look like the stationary grating re-

sponse, and the energy shifts to higher temporal frequencies with increasing speed, as shown in Figure 4.5a. Above the pivotal speed, the response is invariant with speed, when plotted versus time, whereas it is increasingly drawn out, when plotted versus space (See Figure 4.9b). Hence, the frequency responses look like the full-field flicker response, and the energy shifts to lower spatial frequencies with increasing speed, as shown in Figure 4.5b.

4.6 Discussion

A simple physical model, consisting of two reciprocally-connected diffusive (signal-spreading) layers, captures the qualitative aspects of spatiotemporal filtering in the retina. The model reproduces the dependence of spatial filtering on temporal frequency and the dependence of temporal filtering on spatial frequency. In particular, spatial filtering is bandpass at low temporal frequencies, but becomes lowpass at high temporal frequencies. Conversely, temporal filtering is bandpass at low spatial frequencies, but becomes lowpass at high spatial frequencies.

Models of the retina similar to the one that I study here have been proposed and analyzed. However, none of the previous studies analyzed the effect of the model's spatiotemporal inseparability on motion. By studying a minimal model, and treating space as a continuum—using the continuous approximation—just like time, I was able to obtain closed-form analytic solutions, and to develop a clear intuitive picture of the spatiotemporal behavior of the retina.

I showed that the model's spatiotemporal inseparability has serious consequences for how information about contrast and speed is encoded by the retina. It also results in suboptimal filtering, as the model's spatiotemporal behavior deviates from the optimal filter for the ensemble of natural images.

In following subsections, I show how spatiotemporal inseparability goes hand in hand with local connectivity. As a consequence, nature must choose between a costly spatiotemporally separable optimal filter or a cheap spatiotemporally inseparable sub-optimal filter, weighing coding efficiency against implementation efficiency. I also

provide a summary of the procedures that I proposed to extract of information about contrast and speed from the outer retina's outputs.

4.6.1 Spatiotemporal Inseparability and Local Connectivity

The interdependence of spatial filtering and temporal filtering is a direct consequence of the locally connected character of the signal-spreading networks. Signals diffuse in space as they are relayed from node to node by the internode conductances. Signals also diffuse in time as they accumulate on the node capacitances. Consequently, the temporal scale on which signals are processed is intimately connected with the spatial scale at which they occur, and vice versa.

Simultaneous spatial and temporal diffusion places a constraint on the sum of the spatial frequency and the temporal frequency. The current spreading through the internode conductances is proportional to the second spatial derivative of the voltage, and the current charging the node capacitance is proportional to the first temporal derivative. Consequently, the sum of the rates at which the signal changes in space and in time is constrained by the input current. This constraint translates into a constraint on the sum of the spatial frequency and the temporal frequency. Therefore, all the terms that appear in the transfer function of a locally-connected network involve sums of spatial frequency and temporal frequency, instead of products.

Spatiotemporal separability requires a multiplicative interaction between spatial frequency and temporal frequency, not a subtractive one. To obtain a multiplicative interaction, we must put a constraint on the product of spatial frequency and temporal frequency. Such a constraint produces frequency-sensitivity plots with contours running diagonally (for log–log coordinates), as shown in the frequency-response plot for the optimal filter (Figure 3.3b). In contrast, a sum constraint produces L-shaped contours (for log–log coordinates), as shown in the frequency-response plot of the model (Figures 4.7 and 4.8), and in the frequency response plot for humans (Figure 4.6b).

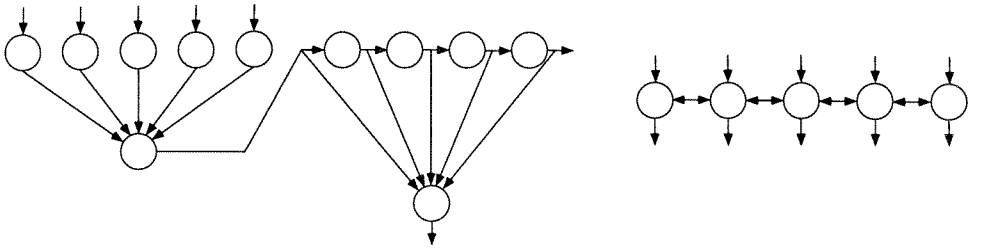


Figure 4.10: HARDWARE FOR SPATIOTEMPORAL FILTERS
Each node has the ability to form a weighted sum of its inputs and to provide a delayed version of the sum to its output; wires communicate signals instantaneously and do not attenuate or amplify them. Left: Separable spatiotemporal filter built from a spatial array and a tapped delay line. This configuration can realize any desired spatiotemporally separable filter. The number of nodes, and the number of wires, required per output is $O(n_r + n_t)$, where n_r is the order of the spatial filter and n_t is the order of the temporal filter. Right: Inseparable spatiotemporal filter built from a nearest-neighbor–connected spatial array. This configuration can realize only those inseparable filters whose spatiotemporal impulse response falls off smoothly. The number of nodes, and the number of wires, required per input is $O(1)$.

4.6.2 Efficient Coding Versus Efficient Implementation

A spatiotemporally inseparable filter cannot match the spectrum of the ensemble of natural images, which is more or less separable, over the entire range of spatial and temporal frequencies. It is possible to match the model's inseparable response to the separable response of the optimal filter, but in only certain restricted regions of the spatiotemporal frequency spectrum. We can acheive optimal spatial filtering at low temporal frequency and optimal temporal filtering at low spatial frequency. However, when matched to the optimal filter in these regions, the model does not filter out signals with poor SNR that occur at high spatial frequencies and high temporal frequencies. This mismatch between the model and the optimal filter predicts that the outer retina devotes more of its channel capacity to noise than is optimal.

Suboptimal outer-retina performance is a small price to pay for efficient implementation. The amount of hardware required to implement these two classes of filters—one separable and the other inseparable—is shown in Figure 4.10. By using nearest-neighbor connections, the inseparable network can share wires and nodes,

and thus can use the hardware efficiently. A factor of $n = n_r + n_t$, where n_r (n_t) is the order of the spatial (temporal) filter, reduction in hardware translates to a similar reduction in pixel size and to a similar reduction in power consumption. These improvements in efficiency allow smaller and faster pixels to be used, increasing the spatiotemporal bandwidth of the retina.

Additional spatiotemporal filtering in the inner retina may compensate for suboptimal behavior of the outer retina, such that excessive noise in the latter's output is not passed on to the optic nerve and transmitted all the way to the brain. Possibly, this result is achieved by the presence of two channels, with one tuned to low temporal frequencies and high temporal frequencies (parvo pathway), and the other is tuned to high temporal frequencies and low spatial frequencies (magno pathway); neither one is tuned to the noisy signals that occur at high temporal freqencies and high spatial frequencies.

4.6.3 Encoding of Contrast and Speed of Moving Images

Understanding how the outer retina responds to motion led me to develop a natural set of procedures for obtaining a speed-invariant estimate of contrast and a contrast-invariant estimate of speed from the outer-retina's output signals.

In particular, there is a pivotal speed that demarcates the border between two distinct regimes. An edge moving at the pivotal speed sweeps across the receptive field in exactly the time it takes for the system to settle. Below the pivotal speed, the response is dominated by energy at the spatial frequency to which the spatial bandpass is tuned, and this energy moves to higher temporal frequencies as the speed increases. Above the pivotal speed, the response is dominated by energy at the temporal frequency to which the temporal bandpass is tuned, and this energy moves to lower spatial frequencies as the speed increases.

We can estimate contrast by measuring the amplitude of the response at the frequencies to which the bandpass spatial filter and the bandpass temporal filter are tuned, and taking the larger value. To see why this strategy works, we recall that

the response is invariant when the sum of the temporal frequency and the spatial frequency is constant. So, by guaranteeing that the spatial frequency is high and the temporal frequency is low, we ensure that changing the temporal frequency has a negligible effect, making the response insensitive to speed. Making the temporal frequency high and the spatial frequency low also works, for the same reason.

We can estimate speed by finding the dominant spectral component, and taking the ratio between that component's temporal and spatial frequencies. In the low-speed regime, the dominant component occurs at the frequency where the spatial bandpass peaks. Hence, we already know the spatial frequency, and we need to determine only the temporal frequency. This strategy is analogous to using the spatial extent of the receptive field as a reference, and measuring the time it takes for the stimulus to cross the receptive field. In the high-speed regime, the strongest spectral component occurs at the temporal frequency where the temporal bandpass peaks. Hence, we already know the temporal frequency, and we need to determine only the spatial frequency. This strategy is analogous to using the temporal extent of the receptive field as a reference, and measuring how far the stimulus travels during that time.

The advantage of this biomorphic motion algorithm is that it uses signals that occur at either the same location or at the same time—unlike other motion algorithms, which try to match up signals that occur at different locations at different times [117, 118, 119]. In general, this **correspondence problem** is difficult to solve, since there are many candidate matches and the correct one can be found only if the features within the field of view are sufficiently distinct to disambiguate. Note that the algorithm proposed here computes only speed—unlike these more general motion algorithms, which compute direction as well. We can use information about speed, however, to eliminate candidate matches, making the correspondence problem more tractable.

Chapter 5 Electrodiffusion: From Nerve Membranes to Transistors

In this chapter, I compare the nerve membrane with the MOS transistor. The nerve membrane is a liquid-state device, with ionic species diffusing in water, whereas the transistor is a solid-state device, with electrons and holes diffusing in a crystal. On a microscopic scale, however, the movement of these charge carriers are the identical. Their motion is driven by the same forces, which are either of thermal or electrical origin. At the macroscopic scale, these forces give rise to diffusion and to drift, respectively. Hence, the transport mechanisms found in cells and in transistors are identical.

There are three important differences between these two devices, though.

First, the effectiveness of the transport mechanisms are drastically different. Diffusion coefficients and mobilities are six orders of magnitude smaller for ions in water, compared to electrons and holes in crystalline silicon. But the ions travel much shorter distances: A lipid bilayer is only 6nm thick, whereas the channel length of a typical transistor is around 1μm. Because decreasing the distance increases the driving force, this 2-decade reduction in length reduces the transit time by 4 decades.

Second, the nerve membrane is strictly a two-dimensional structure, with the same population of charges responsible for its electrostatics and for its electrodiffusion. In contrast, a transistor is fundamentally a three-dimensional structure, with two distinct populations of charges responsible for its electrostatics and for its electrodiffusion. The transistor's electrostatics involves primarily immobile charges on the gate, which is placed on top of the bulk crystal to control the potential at the surface of the crystal. And the transistor's electrodiffusion involves mobile charges at the surface of the bulk crystal that are totally isolated from the charge on the gate.

Third, several ionic species serve to transport charge across the nerve memebrane,

and selective transmembrane ion channels control the permeability of the membrane to each ion species independently. In contrast, a transistor uses a single charged species, and its gate potential controls the flux of this species through the channel. To capture the function of a variety of ion channels, you have to use a separate transistor for each channel type, and control the flux through each transistor independently using that transistor's gate voltage. You also have to copy the currents passed by each transistor onto a separate capacitor if you want to keep track of the concentration of each ionic species.

I will review electrodiffusion in nerve cells and in transistors by deriving expressions for the ionic fluxes and electric currents in these devices from first principles. The derivations are similar because the basic electrostatic and transport mechanisms present in these two devices are identical. The nonlinear partial differential equations that govern electrostatics and electrodiffusion cannot be solved analytically for these devices. To obtain explicit, closed-form solutions, we must make some simplifying assumptions. As the assumptions that hold in one case do not hold in the other case, and vise versa, we end up with different forms of solutions for the membrane and for the transistor. This comparative study—which, to my surprise, has not yet been done—will help us figure out how best to exploit the native physics of the transistor to model the biophysics of the nerve membrane.

5.1 Electrodiffusion in Membranes

The general outline of my review of electrodiffusion in nerve membranes is as follows. I begin by studying the Nernst–Planck equation, which relates the flux to the ion concentrations and the potential at each point within the membrane. The potential is related to the net charge concentration by Poisson's equation. This equation couples together the fluxes of all the ion species present, making it difficult to solve for the flux of each species. Assuming that the electric field in the membrane is constant allows us to obtain the potential profile across the membrane without solving Poisson's equation.

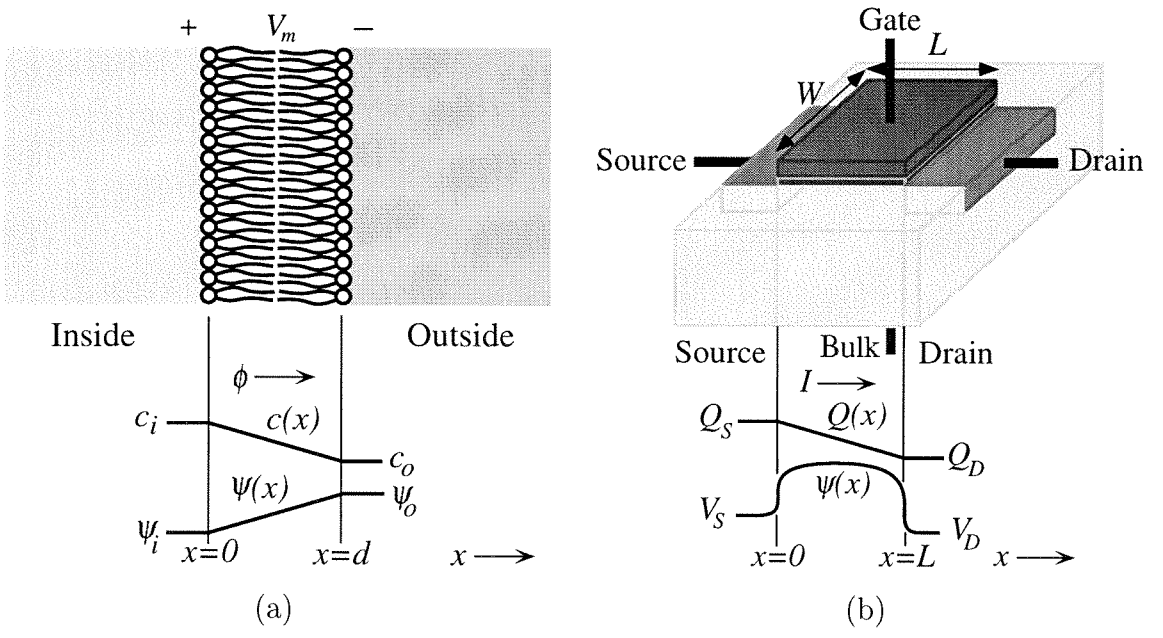


Figure 5.1: ELECTRODIFFUSION IN MEMBRANES AND TRANSISTORS
(a) Cross-section of a bilipid membrane with profiles of the ion concentration of the nth ion species, $c_n(x)$, and of the potential, $\psi(x)$. Ions diffuse down the concentration gradient and drift down the potential gradient; ϕ is the net flux through the membrane. By definition, positive flux flows out of the cell; the x axis also points in this direction. And, by definition, the membrane potential is the potential on the inside minus the potential on the outside. Hence, positive current flows from inside to outside; current and flux are in the same direction for positively charged ions. When the ion species diffusing across the membrane are distributed such that they maintain charge neutrality, the electric field is constant, and the potential changes linearly. The membrane's behavior under this contant–electric-field assumption is derived in the text. (b) Structure of an n-type MOS transistor, with profiles of the mobile-charge concentration, $Q(x)$, and the potential gradient, $\psi(x)$, along the surface of its channel. Just like in the cell membrane, the gradients in the charge concentration and in the potential make free electrons diffuse and drift across the channel; I is the net current through the channel. By definition, charge carriers flow from source to drain; the x axis also points in this direction. As an n-type MOS transitor uses negative charge carriers, its current is negative. The potential along the channel surface is determined by charges on the gate, which attract oppositely charged mobile charges to the surface of the channel, and repel similarly charged mobile charges away from the surface. When the charges repelled are much farther away than those attracted to the surface, we can ignore the effect of the former. In that case, we are left with the mobile charge at the surface and the charge on the gate, which form a parallel-plate capacitor. The transistor's behavior under this parallel-plate capacitor assumption is derived in the text.

The **constant field model** was proposed by Goldman over fifty years ago [1], in a study of biological membranes. Hodgkin and Katz latter applied this model to the giant axon of the squid [2]. This simple model treats the ions as though they diffuse freely: In fact, they travel through a pore that confines them to one degree of freedom, making independent movement unlikely. And it assumes that the pore excludes all the other ionic species: In fact, real pores do not have perfect selectivity. Nevertheless, in additon to being instructive, the Goldman model turns out to be quite robust and useful in practice.

5.1.1 The Membrane Flux

The derivation of the ion flux across the membrane starts with the **Nernst–Planck equation**, which governs the transport of ions in a solvent.[1] The Nernst–Planck equation is an application of Fick's Law, which governs diffusion, and of Ohm's law, which governs drift. The Nernst–Planck equation relates the molar flux, ϕ_n, of the nth ion species to concentration gradient of that species, $\partial c_n/\partial x$, and to the electrical potential gradient, $\partial\psi/\partial x$:

$$\phi_n(x) = -D_n\frac{\partial c_n(x)}{\partial x} - u_n z_n F c_n(x)\frac{\partial\psi(x)}{\partial x}, \tag{5.1}$$

where D_n is the diffusion coefficient of the nth ion species, u_n is the molar mechanical mobility, z_n is the valence, and F is the Faraday charge.

I am abiding by conventions used in biology and chemistry, where the charge quantum is 1 mole—6.022×10^{23} particles! To compute the diffusion component, you multiply the concentration gradient (in units of $\mathrm{mol/m^3/m}$) by the diffusion coefficient ($\mathrm{m^2/s}$), which gives you the flux ($\mathrm{mol/m^2/s}$). To compute the drift component, you multiply the molar concentration (in units of $\mathrm{mol/m^3}$) by the Faraday charge (C/mol), and by the valence of the nth ion (dimensionless), to obtain the charge concentration ($\mathrm{C/m^3}$). Multiplying this charge concentration by the electric field (N/C)—expressed

[1] The derivation of the membrane equations follows closely the treatment in Weiss' thorough, two-volume monograph, *Cellular Biophysics* [120].

as the potential gradient—gives you the force on a unit volume of ions ($\mathrm{N/m^3}$). Multiplying this force by the molar mechanical mobility ($\mathrm{mol \cdot m/s/N}$) gives you the flux ($\mathrm{mol/m^2/s}$).

We can rewrite the Nernst–Planck equation in the form

$$\phi_n(x) = -u_n z_n F V_{T_n} \left(\frac{\partial c_n(x)}{\partial x} + \frac{c_n(x)}{V_{T_n}} \frac{\partial \psi(x)}{\partial x} \right) \tag{5.2}$$

by using the chemist's version of the Einstein relation, $D_n = u_n RT$, to express the diffusion coefficient in terms of the molar mechanical mobility, where R is the molar gas constant (units of J/mol/K) and T is the temperature (K). I have defined a more appropriate unit of voltage $V_{T_n} \equiv (RT)/(z_n F)$, which corresponds to **thermal potential** of the nth ion (units of J/C).

The Nernst–Planck equation, and its device-physics counterpart, the **drift–diffusion equation**, make clear the balance of forces between drift and diffusion. At equilibrium, drift and diffusion cancel out each other, and the flux is 0.

We can solve the Nernst–Planck equation for the concentration profile of the nth ion at equlibrium, $c_{n_0}(\psi)$, by using the chain rule to express the concentration gradient $\partial c_n / \partial x$ as $\partial c_n / \partial \psi \cdot \partial \psi / \partial x$. Making this substitution gives us

$$\begin{aligned} \left(\frac{\partial c_{n_0}(\psi)}{\partial \psi} + \frac{c_{n_0}(\psi)}{V_{T_n}} \right) \frac{\partial \psi(x)}{\partial x} &\equiv 0 \\ \Rightarrow c_{n_0}(\psi) &= c_{n_0}(0) e^{-\psi / V_{T_n}}. \end{aligned} \tag{5.3}$$

At equilibrium, the concentration of the nth ion e-folds every time that the potential decreases by the thermal potential V_{T_n}—assuming it is positively charged. The exponential form arises because, once the dependence on the potential gradient is factored out, drift is proportional to the concentration, whereas diffusion is proportional to the rate at which the concentration changes with potential. Therefore, we must equalize the concentration and the derivative of the concentration to counterbalance drift with diffusion. The exponential is the only function whose derivative is proportional to itself.

Factoring out the potential gradient also makes evident the relative strengths of diffusion and drift away from equilibrium. The ratio between the fluxes caused by diffusion and by drift is equal to the ratio between the rate at which the concentration changes with potential and the concentration itself.

We can compute the ratio between the diffusion and drift components by taking the derivative of the logarithm of the concentration, because the derivative of the logarithm of a function gives the ratio between the derivative of the function and the function itself. Making use of this observation allows us to rewrite the Nernst–Planck equation in a simpler form:

$$\begin{aligned}\phi_n(x) &= -u_n z_n F V_{T_n} c_n(x) \left(\frac{c_n'(\psi)}{c_n(\psi)} - \frac{c_{n_0}'(\psi)}{c_{n_0}(\psi)} \right) \frac{\partial \psi(x)}{\partial x} \\ &= -u_n z_n F V_{T_n} c_n(x) \frac{\partial}{\partial x} \log \left(\frac{c_n(x)}{c_{n_0}(x)} \right), \end{aligned} \tag{5.4}$$

where $f'(u)$ is the derivative of f with respect to u. Notice that, by taking advantage of the equilibrium condition $c_{n_0}'(\psi)/c_{n_0}(\psi) + 1/V_{T_n} = 0$, we can express the thermal voltage in terms of the exponential equilibrium distribution and its derivative. This substitution makes the deviation from equlibrium explicit.

Equation 5.4 has the same form as the drift term in the electrodiffusion equation (Equation 5.1): The flux is given by the product of the concentration and the spatial derivative of a potential function. Therefore, I call Equation 5.4 the **driftlike formulation of electrodiffusion**. When the effects of both drift and diffusion are included in the potential function, it has the form of the logarithm of the ratio of the concentration profile and the concentration profile at equilibrium, with the potential expressed in units of the thermal potential.

The driftlike formulation provides an appealing intuitive interpretation of the behavior away from equilibrium. The average velocity of the particles—which is equal to the ratio between the flux and the concentration—is proportional to the rate at which the ratio between the concentration profile and the equilibrium concentration profile changes with position. Therefore, when the concentration decreases less than

the equilibrium concentration (in percentages) as we move up the potential gradient, the velocity points down the potential gradient (i.e., drift dominates). On the other hand, when the concentration decreases faster than the equilibrium concentration as we move up the potential gradient (in percentages), the velocity points up the potential gradient (i.e., diffusion dominates). Thus, both drift and diffusion tend to redistribute the particles so as to approach the equilibrium distribution, and hence the equilibrium distribution is stable.

We can express the electrodiffusion equation in yet another simple form by dividing Equation 5.4 by $c_{n_0}(x)$. Doing the division gives us the derivative of a quotient, and therefore the result simplifies to

$$\frac{\phi_n(x)}{c_{n_0}(x)} = -u_n z_n F V_{T_n} \frac{\partial}{\partial x} \frac{c_n(x)}{c_{n_0}(x)}. \tag{5.5}$$

Equation 5.5 has the same form as the diffusion term in the electrodiffusion equation (Equation 5.1): The flux is proportional to the spatial derivative of a function of the concentration—and does not depend on the concentration itself. Therefore, I call Equation 5.4 the **diffusionlike formulation of electrodiffusion**. When the effects of both drift and diffusion are included in the concentration function, it has the form of the ratio of the concentration profile and the equilibrium concentration profile, with the concentration expressed in units of the equilibrium concentration.

5.1.2 The Membrane Potential

Proceeding with the derivation of the membrane current, if there are N species of ions, we have to solve N transport equations,

$$\phi_n(x) = -D_n \frac{\partial c_n(x)}{\partial x} - u_n z_n F c_n(x) \frac{\partial \psi(x)}{\partial x}$$

(either Equation 5.4 or Equation 5.5 will do too) to obtain their fluxes; N continuity equations,

$$\frac{\partial J_n(x,t)}{\partial x} = -z_n F \frac{\partial c_n(x,t)}{\partial t},$$

to obtain their concentration profiles; and one electrostatic equation,

$$\frac{\partial^2 \psi(x,t)}{\partial x^2} = -\frac{\rho(x,t)}{\epsilon}$$

to obtain the potential profile.

The physical significance of the last two equations is as follows: The continuity equation relates the net flux into a region to the rate at which the particle population there increases, ensuring that each ionic species is conserved. These currents, which add or remove particles from a given region, are called **displacement currents**. In contrast, the flux moves particles through the boundaries of that region.

The electrostatic equation, which is called **Poisson's Equation**, relates the potential to the net charge distribution, $\rho(x,t)$, produced by all N ion species; ϵ is the permittivity of the membrane. The second spatial derivative of the potential is proportional to the charge density, because you integrate the charge to obtain the electric field (applying Gauss's law), and then integrate the electric field to obtain the potential.

Needless to say, it is extremely difficult to solve these coupled partial differential equations in closed form; therefore, we must find reasonable assumptions that decouple them.

The first assumption that we make is that *the membrane is in the steady state.* That is, the displacement currents are zero and the ion concentrations do not change with timea. Hence, the fluxes do no change with position. The steady-state assumption allows us to solve either Equation 5.4 or Equation 5.5 simply by integrating both sides with respect to x, if we know the concentration profile or the potential profile, respectively. The second assumption that we make is that *the electric field is constant.* That is, the potential, $\psi(x) - \psi(0) = -(x/d)V_m$, where $V_m \equiv \psi(0) - \psi(d)$ is the voltage across the membrane, and d is its thickness.

Knowing the potential profile, we can use Equation 5.3 to obtain the equilibrium concentration profile:

$$c_{n_0}(x) = c_{n_0}(0) \exp\left(\frac{V_m x}{V_{T_n} d}\right).$$

Knowing the equilibrium concentration profile, we can integrate Equation 5.4 to obtain the flux, and convert the flux to current density J_n (in units of $C/m^2/s$) by multiplying by the molar charge for the nth ion species, $z_n F$:

$$J_n = -u_n z_n^2 F^2 V_{T_n} \frac{c_n(d)/c_{n_0}(d) - c_n(0)/c_{n_0}(0)}{\int_0^d 1/c_{n_0}(x)dx} \tag{5.6}$$

$$= u_n z_n^2 F^2 \frac{V_m}{d} \left(\frac{c_n(d)/c_{n_0}(d) - c_n(0)/c_{n_0}(0)}{1/c_{n_0}(d) - 1/c_{n_0}(0)} \right). \tag{5.7}$$

Substituting the expression for $c_{n_0}(x)$ into Equation 5.7, and converting the molar mechanical mobility to electrical mobility, $\mu \equiv |z_n| F u_n$—a quantity commonly used by device physicists—gives us

$$J_n(V_m) = |z_n| F \frac{\mu_n V_m}{d} \frac{c_n(d) - c_n(0) e^{V_m/V_{T_n}}}{1 - e^{V_m/V_{T_n}}}.$$

It is easier to interprete our final expression for the membrane current intuitively when we use electrical mobility rather than mechanical mobility. Notice that, swapping $c_n(0)$ and $c_n(d)$ is exactly equivalent to swapping the sign of V_{T_n}; therefore, an anion channel that sees a higher concentration outside the cell behaves just like an cation channel that sees a higher concentration inside the cell.

Due to the constant-field assumption, the current is proportional to the product of the drift velocity of the ions ($\mu_n V_m/d$) and the charge carried by a mole of these ions ($|z_n|F$). The proportionality constant is determined by the concentration of charge carriers; it changes with the direction of the current because of the difference in concentrations on either side of the membrane. When the membrane voltage is large and positive, the current is equal to the drift velocity times the concentration at the membrane's inner boundary. In a similar vein, when the membrane voltage is large and negative, the current is equal to the drift velocity times the concentration at the membrane's outer boundary. For small membrane voltages, $V_m \ll V_{T_n}$, the exponential is close to 1, and $1 - \exp(V_m/V_{T_n}) \approx V_m/V_{T_n}$. Therefore, we have

$$J_n(V_m) = \frac{|z_n|}{z_n} \mu_n RT \frac{c_n(d) - c_n(0)}{d}.$$

Thus, the membrane passes current by diffusion when the membrane voltage is small, and there is current flow even when the voltage is 0—except in the degenerate case where the concentrations on either side are equal. The current goes to 0 when the the drift and diffusion components cancel each other, which happens when $V_m = \log(c_n(d)/c_n(0))$, as we would expect from the equilibrium concentration profile (Equation 5.3).

We can express the current in terms of the ion concentrations inside and outside the cell, c_n^{i} and c_n^{o}, if we know the partition coefficient $k_n \equiv c_n(0)/c_n^{\mathrm{i}} = c_n(d)/c_n^{\mathrm{o}}$. We can also make the equilibrium condition obvious by defining $E_n \equiv -V_{T_n} \log(c_n^{\mathrm{i}}/c_n^{\mathrm{o}})$; E_n is called the **reversal potential** for the nth ion, because the current carried by this ion changes sign when $V_m = E_n$. Making these substitutions, and referencing the potentials inside (V_{i}) and outside (V_{o}) the cell to a third potential, instead of to each other, gives us

$$J_n(V_m) = |z_n| F P_n c_n^o \frac{V_{\mathrm{i}} - V_{\mathrm{o}}}{V_{T_n}} \left(\frac{e^{(V_{\mathrm{i}} - E_n)/V_{T_n}} - e^{V_{\mathrm{o}}/V_{T_n}}}{e^{V_{\mathrm{i}}/V_{T_n}} - e^{V_{\mathrm{o}}/V_{T_n}}} \right), \tag{5.8}$$

where $P_n \equiv k_n D_n/d$ is the permeability of the membrane to ion n.

When several ionic species are present, the current due to each species is given by Equation 5.8, with the appropriate values of electrical mobility μ_n, valence z_n, and thermal voltage V_{T_n}. In this case, equilibrium occurs when the sum of all the currents is 0.

5.2 Electrodiffusion in Transistors

The general outline of my review of electrodiffusion in the MOS transistor is as follows. I begin by solving the classical drift–diffusion equation, which relates the current to the charges. Approximating the gate and the channel to a parallel-plate capacitor gives us a simple, physically intuitive, closed-form description of the current in terms of the charge concentrations at the channel boundaries. Our next task, then, is to solve for the charge concentration in terms of the potential at the surface of the

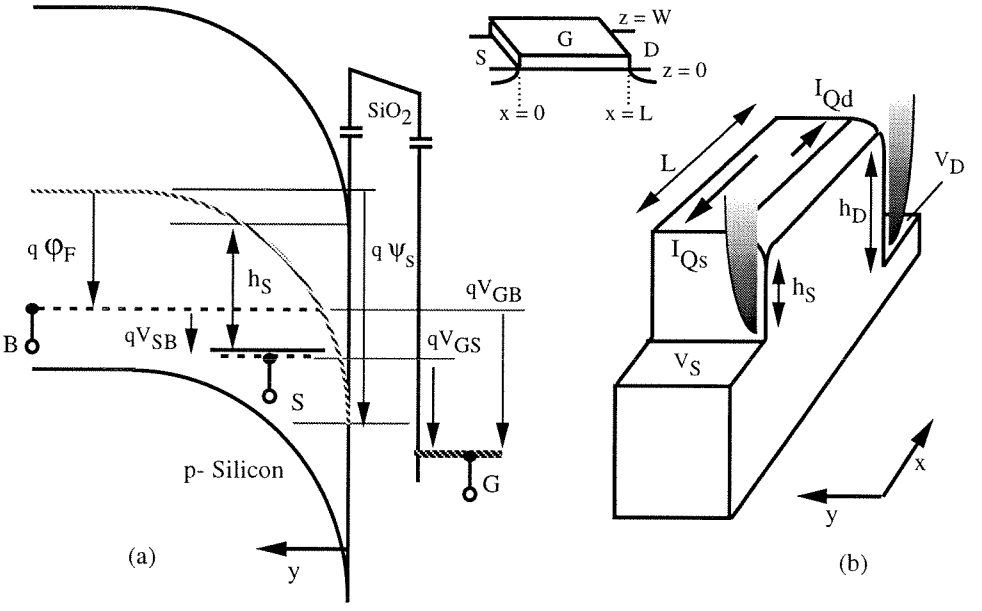


Figure 5.2: Potential Energy of Electrons in a nMOS Transistor
Potential profile along the length ($0 < x < L$) and the depth ($0 < y$) of the channel; the potential is assumed to be uniform along the channel width ($0 < z < W$). Electrons enter the channel at the source end (potential V_S), where they must overcome an energy barrier ($\psi(0, y) - V_S$). The barrier height is lower near the surface of the channel surface ($y = 0$), as positive charges on the gate (potential V_G) attract the electrons. Thus, all the mobile charge in the channel is in close proximity to the channel surface, dying off exponentially as the potential increases with depth. The electrons leave the channel at the drain end, where they drop down a higher potential barrier that excludes entry from that end.

channel, and of the potentials at the source and at the drain. Finally, we must relate the surface potential to the gate potential, to obtain an expression for the current that includes only the voltages applied to the transistor's four terminals. The structure of a transistor, and the profiles of charge density and potential along its channel, is shown in Figure 5.1b.

Maher and Mead proposed the **parallel-plate capacitor approximation** to solve the transport problem [121, 122]. And they used the **charge sheet approximation**, which assumes that all the mobile charge is right at the channel surface, to solve the electrostatic problem. My derivation of the current–charge relationship from the transport equations follows their treatment. But my derivation of the charge–voltage relationship from the electrostatic equations extends their derivation to take into account the distribution of mobile charge down the depth of the channel. This refinement is especially important at the onset of weak inversion, where the charge-sheet approximation breaks down.

5.2.1 The Channel Current

The derivation of the channel current of a MOS transistor begins with the drift-diffusion equation, which governs charge transport in semiconductors.[2] The **drift-diffusion equation** relates the current density, J, to the electrical potential gradient, $\partial\psi/\partial x$, and to the charge concentration gradient, $\partial Q/\partial x$:

$$J(x,y) = -D_{\mathrm{n}}\frac{\partial Q(x,y)}{\partial x} - \mu_{\mathrm{n}}Q(x,y)\frac{\partial \psi(x,y)}{\partial x},$$

where D_{n} is the electron's diffusion coefficient and μ_{n} is its electrical mobility. The current density is assumed to be uniform along the width of the channel (z dimension), but it may vary along both the length (x dimension) and the depth (y dimension). By design, the direction of the current is strictly along the length of the channel; the components along the width and along the depth are 0.

[2]The derivation of the current equations closely follows the treatment in Maher's appendix to Mead's book, *Analog VLSI and Neural Systems* [122].

I have switched to the conventions used in physics and engineering, where the charge quantum is one electron—as it is in reality! To compute the diffusion component, you multiply the charge concentration gradient (in units of $fC/\mu m^3/\mu m$) by the diffusion coefficient ($\mu m^2/ns$), and that gives you the current density ($\mu A/\mu m^2$). To compute the drift component, you multiply the charge concentration ($fC/\mu m^3$) by the electric field ($V/\mu m$)—expressed as the potential gradient—and that gives you the force on a unit volume of electrons ($fC \cdot V/\mu m/\mu m^3$). Multiplying this force by the electrical mobility ($\mu m^2/V/ns$) gives you the current density ($\mu A/\mu m^2$).

Unlike the potential profile across the membrane, which is determined entirely by the ion flux through the membrane, the potential profile across the transistor's channel is controlled by charges on the gate, and by mirror charges on ionized dopant atoms in the bulk, as well as by the electron flux. Charge separation between the channel and the gate violates charge neutrality. Consequently, the constant-field assumption is not applicable to the transistor.

The potential energy of electrons traveling along the channel is shown in Figure 5.2. A good first-order approximation is to assume that the mobile charge changes with the channel potential at the same rate everywhere along the channel—the **parallel-plate–capacitor assumption** [122]. When surface potential increases, more mobile electrons enter the channel, and this charge draws more positive charges onto the gate and screens them from the negatively-charged dopant atoms down in the bulk. Thus, the structure is analagous to two capacitors connected in parallel—one between the surface and the gate and the other between the surface and the edge of the depletion layer. This analogy is perfect if, for each of these capacitors, the charges remain at the same distance from the surface as we move along the length of the channel. In that case, the derivative of the mobile charge with respect to potential is the same everywhere along the channel.

At first glance, you would not expect the parallel-plate–capacitor assumption to hold when the mobile charge concentration may change along the channel length, because more dopant atoms are ionized where the concentration is lower. Consequently, the incremental charge comes from dopant atoms further away, as the depeletion

layer extends down into the bulk. This process is self-limiting, however, because the mobile charge at the surface and the fixed charge in the bulk compete to cover the gate charge. As the depletion layer grows, the charge contributed by the depletion layer reduces, and the charge contributed by the mobile charge increases. In other words, the parallel-plate–capacitor assumption holds as long as the depletion-layer capacitance is smaller than the gate-oxide capacitance. For a typical 2μm process, with a 40nm-thick gate oxide and a dopant concentration of $2.1 \times 10^4/\mu\text{m}^3$, the gate-oxide capacitance exceeds the depletion-layer capacitance when the surface potential exceeds 0.23V.

Proceeding with the derivation of the current, I define the channel capacitance at a perticular depth Y as

$$C(Y) \equiv \partial Q/\partial \psi|_{y=Y},$$

and make the substitution

$$\frac{\partial \psi}{\partial x} = \frac{\partial \psi}{\partial Q}\frac{\partial Q}{\partial x} = \frac{1}{C(y)}\frac{\partial Q}{\partial x}$$

into the drift-diffusion equation. The result is

$$J(x,y) = -\mu V_T \left(1 + \frac{Q(x,y)}{Q_T(y)}\right)\frac{\partial Q(x,y)}{\partial x}. \tag{5.9}$$

Thus, the parallel-plate capacitor assumption makes the drift term look like a diffusion term, with diffusion coefficient proportional to the charge concentration. Compare this equation with Equation 5.5, the diffusionlike formulation for the cell membrane. I used Einstein's relation between diffusivity and mobility, $D = \mu kT/q_e$, where k is Boltzman's constant, T is temperature, and q_e is magnitude of the charge on an electron. I also introduced appropriate units for voltage and for charge: the thermal voltage, $V_T \equiv kT/q_e$, and the thermal charge, $Q_T(y) \equiv C(y)V_T$.

Before integrating this drift-diffusion equation along the channel (x dimension) to get the current, let us convert the charge to units of the thermal charge, $Q_T(y)$. The

result is

$$j(x,y) = -\mu V_T(1 + q(x,y))\frac{\partial q(x,y)}{\partial x},$$

where the symbols in lowercase letters represent quantities in units of the thermal charge. Performing the integration gives us the current density at a particular depth, y, in terms of the charge concentrations at the channel boundaries:

$$\begin{aligned} j &= -\frac{\mu V_T}{L}\left(q_D + \frac{q_D^2}{2} - q_S - \frac{q_S^2}{2}\right) \\ &= -\frac{\mu V_T}{L}\left(1 + \frac{q_D + q_S}{2}\right)(q_D - q_S), \end{aligned} \tag{5.10}$$

where $q_S(y) \equiv q(0,y)$ and $q_D(y) \equiv q(L,y)$ are the charge concentrations at the source and drain ends of the channel. For clarity, I do not explicitly show the dependence of the current and the charges on the channel depth, y. To obtain the total current in the channel, you integrate Equation 5.10 along the depth of the channel (y dimension), as shown in Section 5.2.2, where integrals for the charge terms are computed.

The effects of diffusion and drift are evident in Equation 5.10. Both diffusion and drift are proportional to the charge difference, $q_D - q_S$, because the concentration gradient is proportional to the charge difference, and the potential gradient also is proportional to the charge difference (assuming capacitance is constant). Hence, we can factor out the charge difference, leaving the average charge density, $(q_D + q_S)/2$, and a constant—because the drift component is proportional to the number of carriers, whereas the diffusion component is not. Consequently, drift dominates when the mobile charge is large—compared with the thermal charge—and diffusion dominates when the mobile charge is small.

The drift component and the diffusion component are equal when the mobile charge is equal to the thermal charge, or, more precisely, when $(Q_D + Q_S)/2 = Q_T$. This point is defined as the threshold by Maher and Mead [122]. In the **subthreshold regime**, charge is transported primarily by diffusion; in the **above-threshold regime**, charge is transported primarily by drift. Hence, we can obtain the current

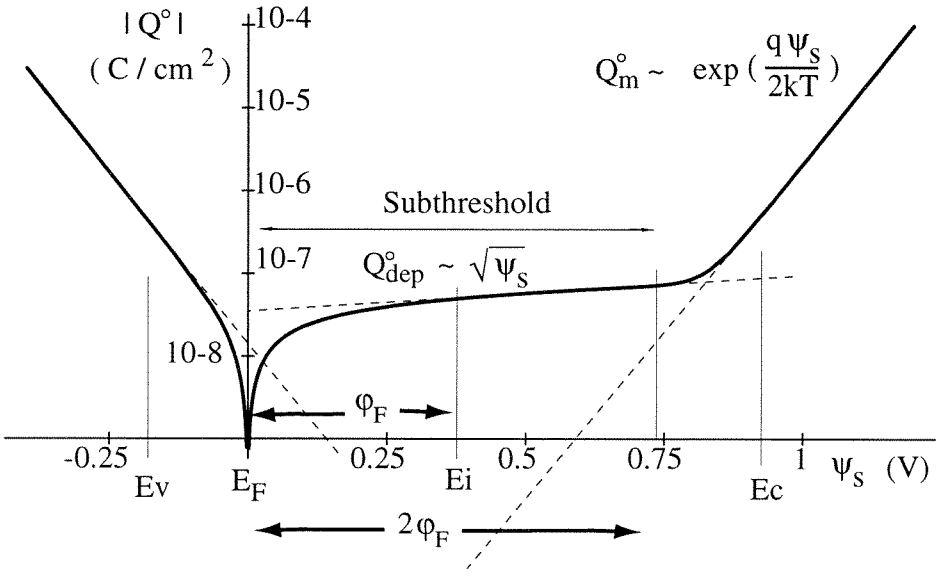


Figure 5.3: MOS Capacitor's Charge Versus Surface Potential
The positions of the conduction band E_c, valence band E_v, Fermi level E_F, and the intrinsic Fermi level E_i also are shown. We compute this plot from a closed-form expression for the total charge given in [123], which was obtained by solving Poisson's equation in the y dimension. The temperature is 300K, and the acceptor-dopant concentration is $2.1 \times 10^4/\mu m^3$—a typical value for a 40nm–gate-oxide process. Notice that the charge concentration grows at one-half of the rate you would expect, taking two thermal voltages to e-fold; this discrepancy is due to space-charge limitation effects in the MOS capacitor structure. The situation would be different in the channel of a transistor, because the heavily doped n-type drain and source regions would supply electrons to the channel.

either above threshold or below threshold using the following approximations:

$$J(Q_S, Q_D) \approx \frac{\mu V_T}{L} \frac{Q_S - Q_D}{Q_T}; \quad Q_S, Q_D \ll Q_T, \tag{5.11}$$

$$J(Q_S, Q_D) \approx \frac{\mu V_T}{L} \frac{Q_S^2 - Q_D^2}{2Q_T}; \quad Q_S, Q_D \gg Q_T. \tag{5.12}$$

5.2.2 The Channel Charge

The current into the bulk (y direction) is practically 0, so we can express the electron and hole concentrations in terms of the potential using the exponential relation be-

tween charge concentration and potential at equilibrium. The problem of finding the charge profile thus reduces to one of finding how the potential changes with depth, so we solve Poisson's equation in the y dimension:

$$\frac{\partial^2 \psi(x,y)}{\partial y^2} = -\frac{\rho(x,y)}{\epsilon_{\mathrm{Si}}},$$

where the net charge density is given by

$$\begin{aligned} \rho(x,y) &= q(p(x,y) - n(x,y) - N_{\mathrm{A}}); \\ n(x,y) &= n_0 e^{\psi(x,y)/V_T}, \\ p(x,y) &= p_0 e^{-\psi(x,y)/V_T}. \end{aligned}$$

N_{A} is the acceptor-dopant atom density, and n_0 and p_0 are the concentrations of free electrons and holes deep down in the p-type bulk material, where the potential is defined to be 0, and the material is charge neutral.

It is easier to compare the relative sizes of the electron and hole populations if we rewrite $n(x,y)$ and $p(x,y)$ in the following form:

$$n(x,y) = n_i e^{(\psi(x,y)-\phi_F)/V_T}, \tag{5.13}$$

$$p(x,y) = p_i e^{-(\psi(x,y)-\phi_F)/V_T)}; \tag{5.14}$$

$$\phi_F \equiv \frac{1}{V_T} \log\left(\frac{N_{\mathrm{A}}}{n_i}\right), \tag{5.15}$$

where $n_i = p_i$ are the electron and hole concentrations in intrinsic (undoped) silicon. The electron and hole concentrations in the p-doped bulk material are equal to the intrinsic concentration—and are equal to each other—when $\psi = \phi_F$. Therefore, ϕ_F is also the potential difference between the p-doped bulk and the undoped silicon when these two semiconductor crystals are in equilibrium. Because doping levels are typically millions of times larger than the intrinsic concentration ($n_i = 0.0145/\mu\mathrm{m}^3$ at 300K), ϕ_F is about 0.4V—close to one-half of the bandgap of silicon.

We have to be careful not to apply Equations 5.13 and 5.14 in close proximity

to the source and drain regions. These heavily doped regions are a rich source of electrons, and perturb the electron and hole concentration at the ends of the channel. Assuming quasi-static behavior, we can approximate the carrier distribution along the channel in the vicinity of the source and drain regions with the exponential equilibrium profile. In that case, replacing ψ by $\psi - V_S$ gives the distribution at the source end of the channel; a similar calculation gives us the distribution at the drain end. To see why, we observe that the built-in potential makes the electron and hole concentrations on the channel side equal to those in the bulk when the potential difference between the bulk and the source is 0.

As we increase the surface potential, we observe three distinct regions, with one of the three charged species in the majority in each region, as shown in Figure 5.3. When the surface potential is negative, holes are drawn to the surface and accumulate there, increasing exponentially as the surface potential decreases. When the surface potential is positive, holes are repelled from the surface, and a depletion layer with negatively charged dopant atoms develops. Electrons are drawn to the surface when the surface potential is positive, and increase exponentially, whereas the depletion charge increases as the square root of the surface potential. Hence, the electrons take over when the surface potential becomes large.

The terms **weak inversion** and **strong inversion** are used to describe the size of the invading electron population relative to the native hole population. The channel is said to be inverted when the electrons, which are normally the minority carriers in the p-type substrate, become the majority carriers. Inversion starts when the electron and hole concentrations become equal; that is, when ψ equals ϕ_F. The channel is said to be **strongly inverted** when the electron concentration exceeds the original hole concentration—that is, when ψ exceeds $2\phi_F$.

The transistor has three distinct regimes of operation, depending on whether the holes, the dopant atoms, or the electrons are the dominant charges. The first region is known as the **accumulation region** ($\psi_s - V_S < 0$), and the other two correspond to the **subthreshold regime** ($0 < \psi_s - V_S < 2\phi_F$) and the **above-threshold regime** ($2\phi_F < \psi_s - V_S$), which were distinguished in Section 5.2.1 by the mode of charge

transport.

Unfortunately, general solutions for the concentration-versus-depth profiles of electrons, holes, and dopant ions cannot be obtained in closed-form. Therefore, I use a first-order Taylor expansion at $\psi = \psi_s$ to obtain an approximation for the effective densities of states:

$$\int_{-\infty}^{\psi_s} F(\psi) e^{\psi/V_t} d\psi \approx (F(\psi_s) - V_T F'(\psi_s)) V_T e^{\psi_s/V_T}. \tag{5.16}$$

This approximations To use this result, we perform a change of variables from y to ψ:

$$\begin{aligned}
Q^n(x) &\equiv \int_{-\infty}^{y_s} \left(\frac{q_e n_i e^{(\psi(y)-\phi_F-\phi_c)/V_T}}{Q_T(y)} \right)^n Q_T(y) \frac{dy}{d\psi} d\psi \\
&= \int_{-\infty}^{\psi_s} \left(\frac{q_e n_i e^{(\psi-\phi_F-\phi_c)/V_T}}{C(\psi)V_T} \right)^n \frac{C(\psi)V_T}{-\mathcal{E}(\psi)} d\psi \\
&= -V_T \left(\frac{q_e n_i}{V_T} \right)^n \int_{-\infty}^{\psi_s} \frac{1}{\mathcal{E}(\psi)} \frac{1}{C(\psi)^{n-1}} e^{n(\psi-\phi_F-\phi_c)/V_T} d\psi \\
&\approx -\frac{V_T^2}{n} \left(\frac{q_e n_i}{V_T} \right)^n \left(\frac{V_T}{n} \frac{\mathcal{E}'(\psi_s)}{\mathcal{E}(\psi_s)} + \frac{(n-1)V_T}{n} \frac{C'(\psi_s)}{C(\psi_s)} + 1 \right) \frac{e^{n(\psi_s-\phi_F-\phi_c)/V_T}}{\mathcal{E}(\psi_s) C(\psi_s)^{n-1}},
\end{aligned} \tag{5.17}$$

where $\psi_s \equiv \psi(x, 0)$ is the potential at the surface. The reference voltage for the surface potential, ϕ_c, is equal to V_S at the source end of the channel, is equal to and V_D at the drain end, and is 0 in parts of the channel that are isolated from the source–drain regions. $\mathcal{E}(\psi_s)$ is the electric field in the direction normal to the surface, pointing away from the bulk.

We can compute the electric field at the surface, $\mathcal{E}(\psi_s)$, by applying Gauss' law, if we know the amount of charge between the surface and the reference point deep down in the bulk where the field is 0. We can ignore the mobile charge both above and below threshold when we compute the charge, because in the former case the mobile charge is confined to a thin sheet right at the surface, and in the latter case the mobile charge is negligible compared to the fixed charge in the depletion layer.

Ignoring the mobile charge as well as the holes, which are both negligible compared

to the depletion-layer charge in the subthreshold region, we have $\mathcal{E}(\psi_s) \approx -Q_{dep}/\epsilon_{Si}$, where

$$Q_{dep} = -\sqrt{2q_e\epsilon_{Si}N_A\psi_s}; \tag{5.18}$$

ϵ_{Si} is the permitivity of silicon and q_e is the electronic charge. We obtain this expression by assuming that the dopant atom density, N_A, is uniform from the surface down into the depth, and that the boundary of the depletion layer is defined sharply, with the charge density dropping abruptly from N_A to zero.

Using Equation 5.18 and Gauss' law, we obtain the function $\mathcal{E}(\psi_s)$, and we substitute this result into the expression for $Q^n(x)$ in Equation 5.17 above, with $n = 1$, to obtain the mobile-charge per unit area underneath the gate:

$$\begin{aligned} Q(x) &= \left(\frac{V_T}{2\psi_s} + 1\right) \frac{q_e\epsilon_{Si}n_iV_T}{\sqrt{2q_e\epsilon_{Si}N_A\psi_s}} e^{(\psi_s-\phi_F-\phi_c)/V_T}, \\ &= \left(\frac{V_T}{2\psi_s} + 1\right) \sqrt{\frac{2q_e\epsilon_{Si}N_A}{\psi_s}} \frac{n_iV_T}{2N_A} e^{(\psi_s-\phi_F-\phi_c)/V_T}, \\ &\approx C_{dep}(\psi_s)V_T e^{(\psi_s-2\phi_F-\phi_c)/V_T}; \end{aligned} \tag{5.19}$$

for $\psi_s \gg V_T$, where

$$C_{dep}(\psi_s) \equiv \frac{\partial}{\partial\psi_s}(-Q_{dep}) = \frac{1}{2}\sqrt{\frac{2q\epsilon_{Si}N_A}{\psi_s}}, \tag{5.20}$$

is the depletion-layer capacitance. As the surface potential increases, the depletion-layer capacitance decreases like the square root, because the depth of the depletion layer increases like the square-root of the surface potential. We can use Equation 5.19 to calculate the charge terms in the subthreshold current–charge relation (Equation 5.11).

The dependence of the subthreshold charge expression (Equation 5.19) on the depletion capacitance results in an inverse–square-root dependence in the pre-exponential factor. This dependence comes from integrating all the mobile charge from the surface down into the bulk. On one hand, when the potential at the surface is close to that in the bulk, the mobile charge spreads deep into the bulk. Hence, the integral is large, and it decreases rapidly as the surface potential deviates from the bulk

potential. Therefore, we must take into account the inverse–square-root dependence on ψ_s to predict low-level currents accurately. On the other hand, when the surface potential becomes large, the mobile charge concentration dies off rapidly away from the surface. Hence, the integral is small, and does not change dramatically with ψ_s. Therefore, the square-root dependence on ψ_s can be ignored at high-level currents.

5.2.3 The Surface Potential

Summing all the voltage drops that we encounter as we go from the bulk to the gate, we find that

$$V_{GB} = \psi_s + V_{ox} + \phi_{MS},$$

where V_{ox} is the voltage across the gate-oxide capacitor, C_{ox}, and ϕ_{MS} is the contact potential between the gate material and the bulk material. The voltage drop across the gate-oxide $V_{ox} = -(Q_{tot} + Q_0)/C_{ox}$; where Q_{tot} is the total charge in the bulk, and Q_0 is the charge due to electrons trapped at the oxide interface and ions implanted at the channel surface. We introduce the **flat-band voltage** $V_{FB} \equiv \phi_{MS} - Q_0/C_{ox}$ to account for the constant voltage offset due to these fixed charges and to the contact potential. Hence,

$$V_{GB} = V_{FB} + \psi_s - Q_{tot}(\psi_s, \phi_c)/C_{ox}. \tag{5.21}$$

When $V_{GB} = V_{FB}$, the surface potential is 0, and the semiconductor is charge neutral. By substituting an expression for the total charge, Q_{tot}, which is given by the sum of the mobile charge (Equation 5.19) and the depletion charge (Equation 5.18), into Equation 5.21, we can obtain a relation between the gate voltage and the surface potential.

Equation 5.21 tells us that the amount by which the surface potential changes when we change the gate voltage depends on the amount by which the total channel charge changes. The dependence of the total charge on the surface potential changes radically when we cross threshold. The total channel charge increases as the square root of the surface potential when the channel is weakly inverted, and

increases exponentially with the surface potential when the channel is strongly inverted. Consequently, when we are below threshold, the surface potential follows closely changes in the gate potential because the dependence of the total charge on the surface potential is weak. In contrast, when we are above threshold, the surface potential stays more or less constant because the dependence of the total charge on the surface potential is strong. The dependence transitions between these two extremes in the region $2\phi_F < \psi_s - \phi_c < 2\phi_F + 5kT/q_e$, where neither the mobile charge nor the depletion charge dominates.

In subthreshold, the total charge is approximately equal to the depletion charge. Substituting the expression for the depletion charge from Equation 5.18 into Equation 5.21 yields

$$V_{GB} = V_{FB} + \psi_s + \gamma\sqrt{\psi_s}, \tag{5.22}$$

where the constant γ is defined as follows. Observe that the voltage drop across the oxide capacitor, $\gamma\sqrt{\psi_s}$, equals the surface potential, ψ_s, when the surface potential is equal to γ^2. Hence, using the expression for Q_{dep} in Equation 5.18,

$$\gamma = \frac{\sqrt{2q\epsilon_{Si}N_A}}{C_{ox}}. \tag{5.23}$$

Solving Equation 5.22 for ψ_s gives us

$$\psi_s = \left(-\frac{\gamma}{2} + \left(\frac{\gamma^2}{4} + V_{GB} - V_{FB}\right)^{1/2}\right)^2. \tag{5.24}$$

Equation 5.24 is approximated well by a linear relation throughout the subthreshold region. Thus, we have

$$\psi_s(V_{GB}) = (1.5\phi_F + \phi_c) + \kappa(V_{GB} - V^*_{GB}), \tag{5.25}$$

where κ is the slope at V^*_{GB}. We choose V^*_{GB} such that the surface potential is in the

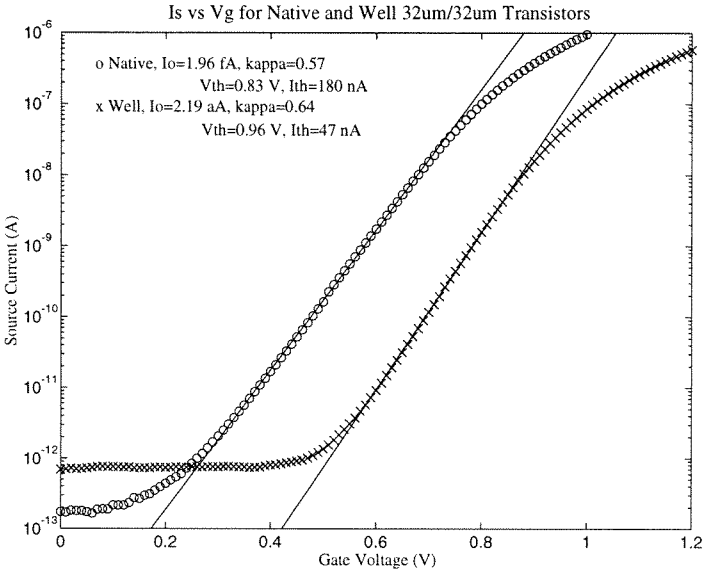


Figure 5.4: TRANSISTOR CURRENT VERSUS GATE VOLTAGE
Experimental and theoretical curves confirming the exponential relationship between the current and the gate voltage in the subthreshold region.

midpoint of the range; that is, $\psi_s(V^*_{\rm GB}) \equiv (1.5\phi_F + \phi_c)$. From Equation 5.22,

$$\kappa \equiv \frac{\partial \psi_s}{\partial V_{GB}}\Big|_{V_{\rm GB}=V^*_{\rm GB}} = 1 + \frac{1}{1+\gamma/(2\sqrt{\psi_{\rm s}(-V^*_{\rm GB})})}. \tag{5.26}$$

The physical significance of κ is apparent if we express this parameter in terms of the oxide capacitance and the depletion capacitance. Rewriting Equation 5.26, using Equation 5.20 and the definition of γ (Equation 5.23), we obtain

$$\kappa = \frac{C_{\rm ox}}{C_{\rm ox} + C_{\rm dep}}.$$

It becomes clear that the oxide and depletion capacitances form a capacitive divider between the gate and bulk terminals that determines the surface potential. Lighter doping reduces γ, reduces $C_{\rm dep}$, and pushes the divider ratio closer to unity. A larger surface potential also reduces $C_{\rm dep}$.

So far, we have derived four equations that describe the transistor in all its regimes

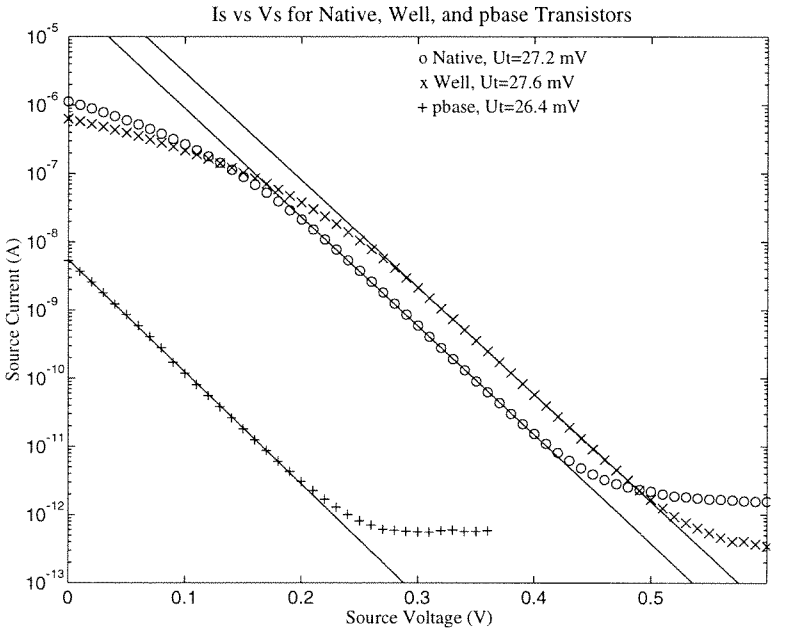


Figure 5.5: TRANSISTOR CURRENT VERSUS SOURCE VOLTAGE
Experimental and theoretical curves confirming the exponential relationship between the current and the source voltage in the subthreshold region.

of operation:

1. Equation 5.10 gives the current density, $j(y)$, at any depth in the channel in terms of the charge concentrations there.

2. Equation 5.17 gives the total charge and powers of the total charge, $Q^n(\psi_s, \phi_c)$, at any position along the channel length, in terms of the surface potential.

3. Equation 5.18 gives the charge in the depletion layer, $Q_{dep}(\psi_s, \phi_c)$, at any position along the channel in terms of the surface potential; differentiating this function gives the depletion-layer capacitance.

4. Equation 5.24 gives the surface potential, $\psi_s(V_{GB})$, in terms of the the gate voltage.

These equations give us a complete description of MOS transitor behavior—to the extent that the numerous assumptions we made are valid.

We have also obtained simpler versions of these equations that are valid over a limited range. Equations 5.12 and 5.11 give approximations for the current density above threshold and below threshold, respectively. Equation 5.19 gives approximations for the mobile charge below threshold; the depletion-layer capacitance is given by Equation 5.20. And Equation 5.25 gives an approximate value for the surface potential below threshold.

For subthreshold operation, these approximations give

$$\begin{aligned} Q_S(\psi_s, V_{SB}) &\approx -V_T C_{dep} \exp\left(\frac{\psi_s - V_{SB} - 2\phi_F}{V_T}\right), \\ Q_D(\psi_s, V_{DB}) &\approx -V_T C_{dep} \exp\left(\frac{\psi_s - V_{DB} - 2\phi_F}{V_T}\right), \end{aligned}$$

$$\psi_s(\phi_c, V_{GB}) \approx 1.5\phi_F + \phi_c + \kappa_n(V_{GB} - V^*_{GB}(\phi_c)),$$

$$I(Q_S, Q_D) \approx \mu V_T \frac{W}{L}(Q_S - Q_D). \tag{5.27}$$

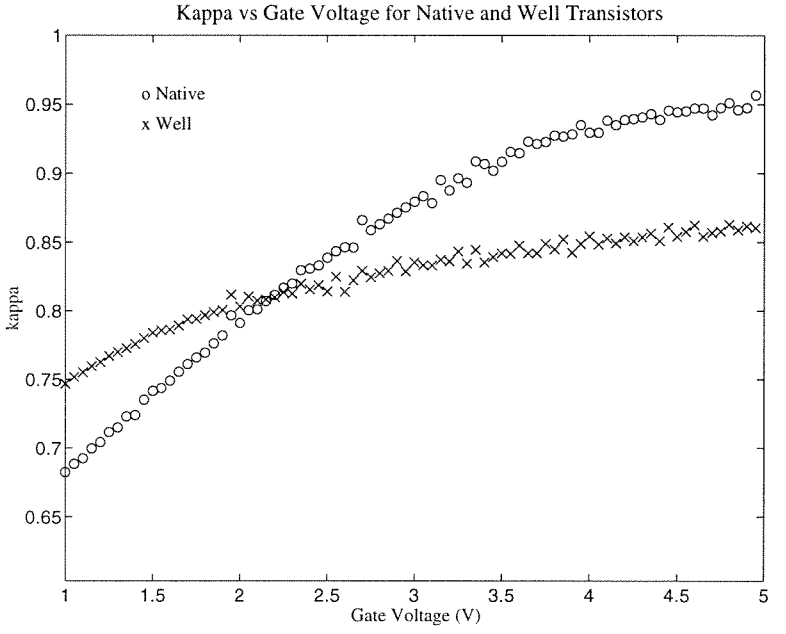


Figure 5.6: SURFACE-POTENTIAL SENSITIVITY VERSUS GATE VOLTAGE
When the current is constant, the difference between the surface potential and the source voltage remains fixed; and hence, the source voltage tracks the surface potential. Consequently, we can estimate the sensitivity of the surface potential to the gate voltage—called κ—by differentiating the V_{SB}-versus-V_{GB} curve. This curve was measured by sweeping the gate of a transistor with a constant current supplied to the source of the transistor.

Thus, we obtain the familiar result for an nMOS transistor:

$$I_{\rm DS} = SI_{no}e^{\kappa_n V_{\rm GB}/V_T}(e^{-V_{\rm SB}/V_T} - e^{-V_{\rm DB}/V_T}), \tag{5.28}$$

$$I_{n0} = \mu V_T^2 C_{\rm dep} \exp\left(-\frac{\phi_F/2 + \kappa_n V_{GB}^*}{V_T}\right); \tag{5.29}$$

where $S \equiv W/L$ is the ratio of the channel width to the channel length. The pMOS equation is similar except for a sign flip on the terminal voltages due to a change in sign of the charge carriers. The experimental data plotted in Figures 5.4 and 5.5 confirm the exponential dependence of the channel current on the gate and source voltage over at least four decades. The deviation at the low end is due to leakage currents in the measurement setup.

The process-dependent parameters, I_{no} and κ_n, can be measured experimentally. Note that I_{no} is not simply the point where a line running along the $\log(I)$ versus $V_{\rm GS}$ curve intercepts the $V = 0$ axis. Fitting Equation 5.28 and Equation 5.29, which take into account the $1/\sqrt{\psi_s}$ pre-exponential factor, is the only way to determine I_{no} accurately. We must also exercise care in obtaining the subthreshold slope coefficient κ, because it depends on the surface potential. The relationship between κ and the surface potential can be measured by fixing the current passed by the device and sweeping the gate voltage, as shown in Figure 5.6.

For submicron devices, we must adjust the drawn widths and lengths to obtain the dimensions of the channel. These adjustments must take into account channel-length modulation by the depletion layer at the drain end, fringing of electric field lines at the channel boundaries, and velocity saturation at extremely high fields.

The terminal voltages can be referenced to the source, instead of to the bulk:

$$I_{\rm DS} = SI_{no} \exp\left(\frac{(1-\kappa_n)V_{\rm BS} + \kappa_n V_{GS}}{V_T}\right)\left(1 - e^{-V_{\rm DS}/V_T}\right). \tag{5.30}$$

This form makes explicit the role of the bulk as a **back gate**. However, it obscures the symmetry between the drain and the source.

Ion Species	Cations $z_n > 0$	Anions $z_n < 0$
Higher concentration inside $c_n^{\mathrm{i}} \gg c_n^{\mathrm{o}}$	pMOS Transistor with $V_{\mathrm{SB}} = V_{\mathrm{m}};\ V_{\mathrm{DB}} = E_n < 0$	nMOS Transistor with $V_{\mathrm{SB}} = V_{\mathrm{m}};\ V_{\mathrm{DB}} = E_n > 0$
Lower concentration inside $c_n^{\mathrm{i}} \ll c_n^{\mathrm{o}}$	nMOS Transistor with $V_{\mathrm{SB}} = V_{\mathrm{m}};\ V_{\mathrm{DB}} = E_n > 0$	pMOS Transistor with $V_{\mathrm{SB}} = V_{\mathrm{m}};\ V_{\mathrm{DB}} = E_n < 0$

Table 5.1: MODELING PASSIVE PROPERTIES OF ION CHANNELS WITH TRANSISTORS

For devices that are biased with $V_{DS} \geq 4\,(kT/q_{\mathrm{e}})$, the drain current becomes independent of the drain voltage, and is simply

$$I_{\mathrm{DS}} = SI_{no} \exp\left(\frac{\kappa_n V_{\mathrm{GS}}}{V_T}\right), \tag{5.31}$$

assuming that the source and substrate terminals at the same potential. Devices operating in this region are said to be in **saturation**. However, channel-length modulation—which we have ignored completely—causes the current to increase slowly with the drain voltage. This effect must be included in the model if we want to predict the output conductance accurately [122].

5.3 Discussion

We have seen that there are similarities and differences between the nerve membrane and the MOS transistor. The similarities between these two structures are most evident at the microscopic level, since the physics that governs their behavior is the same. Balancing drift and diffusion results in equilibrium concentration profiles that decrease exponentially with potential in both devices. The differences between the two devices arise from the way in which the flux is controlled.

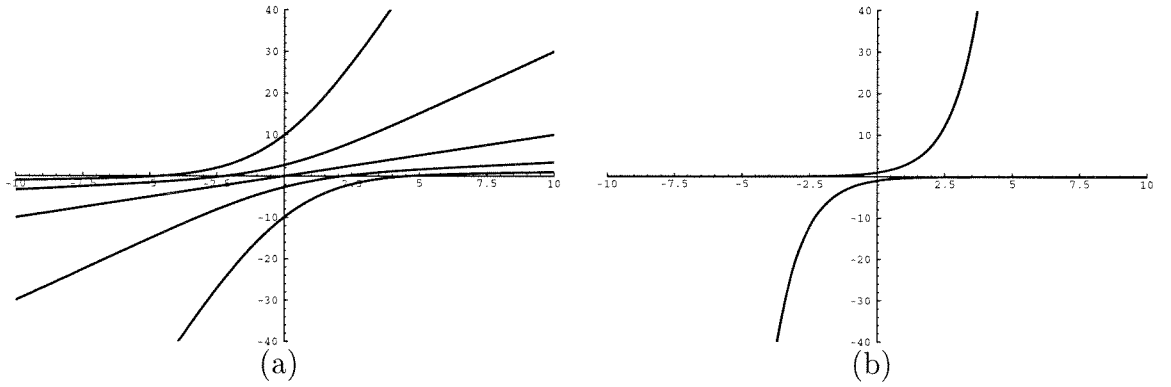


Figure 5.7: CURRENT VERSUS VOLTAGE FOR MEMBRANES AND TRANSISTORS
(a) Membrane curves from Equation 5.8. The five curves correspond to different ratios between the concentrations inside and outside the cell. The ratios (inside:outside) are $1:100$, $1:10$, $1:1$, $10:1$, and $100:1$, going from the lowest curve to the highest one. Voltages are in units of V_{T_n} and currents are in units of $|z_n|FP_n\sqrt{c_n^i c_n^o}$. The ions are cations, and they flow out of the cell (direction for positive current) when the membrane potential is positive. Therefore, steep slopes occur for positive currents when the concentration inside the cell is higher than that outside; the situation is exactly reversed for anions. (b) Transistor curves from Equation 5.28. The top curve is for a pMOS transistor and the bottom one is for an nMOS transistor. The voltage applied to one of the source–drain terminals of the transistor is plotted; this terminal corresponds to the inside of the cell. A voltage equal to the desired reversal potential ($\pm 4.6V_T$ in this case) is applied to the other end of the channel. Voltages are in units of V_T and currents are in units of $SI_{n0}\exp(\kappa V_{GB}/V_T)$. For the nMOS device, which uses electrons, the inside of the cell becomes the source when the current is negative, and the current increases rapidly below the reversal potential. For the pMOS device, which uses holes, the inside of the cell becomes the source when the current is positive, and the current increases rapidly above the reversal potential.

5.3.1 Ion Channels Versus Transistors

We can separate the properties of nerve membranes and transistors into active ones and passive ones. **Passive properties** describe the behavior of the devices given a fixed conformational state of the ion channel, or a fixed voltage applied to the gate. And **active properties** describe the behavior of the devices when the conformational state of the ion channel, or the voltage applied to the gate, are varied to modulate the flux through the channel.

With respect to their passive properties, these devices are qualitatively similar, as shown in Figure 5.7. A pMOS device reproduces the qualitative behavior of a cation channel that sees a higher concentration inside the cell, or of an anion channel that sees a higher concentration outside the cell. And an nMOS transistor reprodces the qualitative behavior of a cation channel that sees a higher concentration outside the cell, or of an anion channel that sees a higher concentration inside the cell. These analogies between membranes and transistors are summarized in Table 5.1.

There is an important quantitative difference the transistor and the membrane, however, which is obvious from the expressions for the currents in these two devices. Exponentials of the voltages inside and outside the cell appear in the denominator as well as the numerator or the membrane equation (Equation 5.8)—resulting in linear asymtotic behavior. In contrast, the transistor current is not normalized in this fashion—resulting in exponential asymptotic behavior.

This difference arises because the concentrations of holes in the the drain–source regions of a pMOS transistor are millions of times larger than the concentration of holes in the n-type bulk. A similar situation holds for electrons in the nMOS transistor. In contrast, the ions that are primarily responsible for the electrical properties of the cell—namely, K^+ and Na^+—have concentration ratios of 1 or 2 decades. We could get the transistor to match the ion-channel's current-voltage curve quantitatively by reducing the doping of the source–drain regions by four or more decades.[3]

[3]The derivation of the MOS transistor current reviewed here does not apply in this low-doping case because we assumed quasistatic behavior. This assumption does not hold when the current levels are large compared to the doping levels, as there are large deviations from the equilibrium charge concentration profile (so-called strong-injection effects). These effects limit the current to the

As far as active properties go, the transistor and the cell membrane are are not even qualitatively similar. The active gating properties of ion channels arise from conformational changes in channel proteins that physically close or open a pore. The pore is a microscopic device whose dimensions are matched to those of the ions, making it selective for a particular size. The voltage dependence of the gating mechanism comes from charged subunits on the channel protien that move from one side of the membrane to the other to open or close the channel; going up or down in energy, depending on the sign of the membrane voltage. In contrast, the transistor modulates the current through its channel not by changing the voltage across the channel, but rather by changing the energy barriers that the carriers must overcome to gain entry into the channel. We need to add more circuitry to monitor the voltage difference across the channel and to modulate the gate voltage appropriately, if we wish to model the active properties of ion channels.

5.3.2 Single-Cell Model

When all the ion channels see the same voltage difference—as they do when they are part of the same cell—the relative differences between the currents in different ion-channel populations may be reproduced fairly well using transistors. This proportionality arises because the exponentials in the denominator of the ion-channel current expression are the same for all the channels, and may therefore be factored out. As the exponentials in the numerator dominate the linear term in the membrane equation when the concentration gradient is large, the ion-channels and transistors behave alike. Except that as the voltage levels increase, the normalizing action of the denominator in the membrane equation limits the current, whereas the transistor current increases exponentially. To mimick this effect, we would have to make the gate voltage of the transistor track the source–drain voltages.

In any case, we can build a fairly decent single-cell model in a standard CMOS process by using a single transistor to model each population of ion channels. In

level predicted by Ohm's law, makeing the asymtotic behavior identical to the membrane's.

particular, the model will reproduce the behavior of the cell at equilibrium (i.e., the dependence of the resting membrane potential on channel permeability, which is described by the Goldman–Hodgkin–Katz equation [1, 2]), because scaling all the pearmeabilities by the same factor does not change the equilibrium point—it just scales all the currents by the same amount. But the model will not reproduce the behavior away from equilibrium because it does not reproduce the actual current levels in the cell. The model could be improved by going to a CMOS process with low doping levels and adding a gate-bias circuit that senses the voltages applied to the source–drain terminals and adjusts the voltage applied to the gate appropriately.

In Chapter 6, we go beyond the single cell to study multiple-cell networks. In particular, we propose transistor-based models for gap-junction–coupled cell syncytia. Such syncytia are common in the retina, and they occur in other parts of the brain as well.

Chapter 6 Linear Networks: By Diffusion in MOS Transistors

In this chapter, I extend the device-level charge-based formulation of the MOS transistor to the circuit level by introducing the concepts of terminal and node charges, and the equivalence principle. With this formalism, we can exploit the linear current–charge relationship of the MOS transistor at the circuit level, enabling us to simulate the diffusion of ions in cell syncytia, or the spread of current in resistive netorks, extremely efficiently.

Ions spread from cell to cell in a syncytium through ion channels that are part of the gap-junction synaptic complex formed between these cells. At gap junctions, the membranes of two cells are in close juxtaposition, and pores in the two membranes are lined up. Hence, a channel is formed, and ions cross from the intracellular fluid of one cell directly into the intracellular fluid of another.

When the ion channel sees a small concentration gradient—as it does when a gap junction is formed between cells of the same type—transport of ions is primarily by drift. As we saw in Chapter 5, in the constant-field approximation, drift produces a linear current–voltage relationship. It is difficult to reproduce this linear current–voltage characteristic with the transistor. The transistor's behavior is close to linear over a range of only a few thermal voltages. For higher voltages, its current either increases exponentially or saturates, depending on the polarity of the voltage.

I show here how we can obtain linear behavior by exploiting the inherently linear current–charge relationship of the transistor in the subthreshold regime.

6.1 Symmetric MOS Transistor Model

Using the charge-based formulation for the channel current in a MOS transistor, we can develop an intuitive, physically accurate, circuit-level abstraction of this device. The parallel-plate–capacitor approximation yields a simple and symmetric relationship between the channel current and the charge at the ends of the channel. As it turns out, this intuitively appealing model does indeed provide a good fit to the experimental measurements [121, 122]. Thus, we can confidently use it to develop circuit models of the transistor.

In addition to preserving our intuition about the symmetrical construction of the MOS transistor, a symmetric model is easier to use than an asymmetric one. When you use asymmetric models, you have to figure out which terminal is the source, and then reference the voltages of all the other terminals to that source terminal before you can apply the model. With a symmetric model, all voltages are referenced to the bulk, and it is not necessary to know a priori which terminal is the drain and which is the source. Because the roles of the channel terminals are determined by the direction in which the charge carriers flow, source and drain can be determined by the circuit design, by the bias conditions—and even by the input signals. Using a symmetric model makes it easier to understand the behavior of circuits with such flexibility, and enhances our ability to design circuits that use unidirectional currents as well as bidirectional currents.

I adopt the conventions for voltages and currents in pMOS and nMOS transistors shown in Figure 6.1a, to preserve symmetry between the channel terminals. For example, in an n-well process, the local reference for the nMOS transistors is the p-substrate V_{NBB} (usually V_{SS}); for the pMOS transistors, it is the n-well V_{PBB} (usually V_{DD}). This notation is consistent with that used by Mead [122], and also with that commonly used in the subthreshold MOS literature [124]. Because the drain and source terminals are treated symmetrically by these conventions, and because the circuit model for the device is itself symmetric, we can assign labels of source and drain to the channel terminals arbitrarily, without regard to the actual direction of

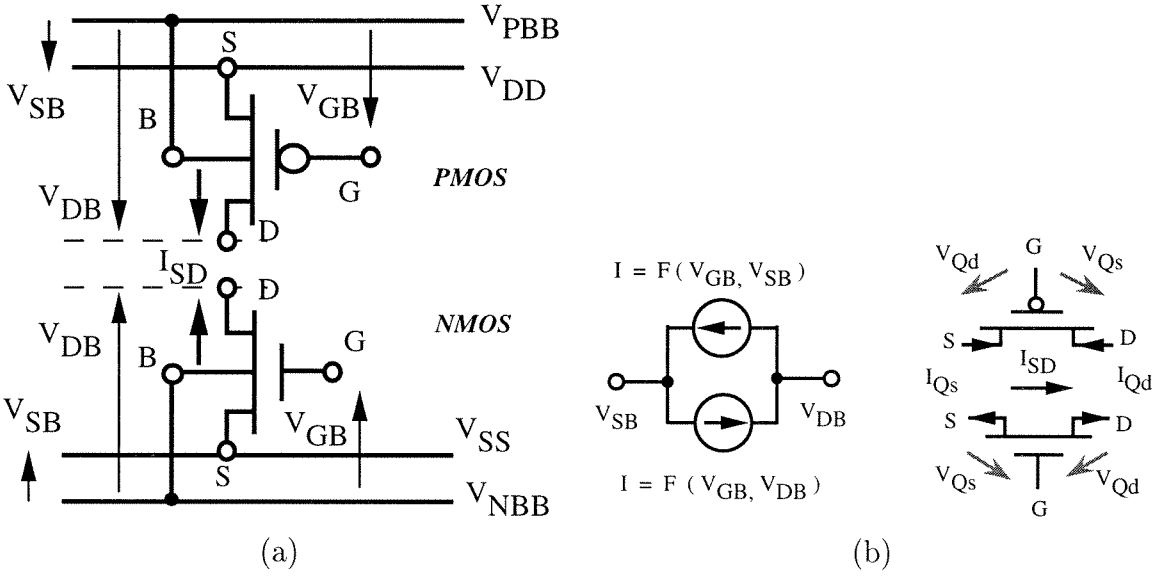


Figure 6.1: Circuit Conventions for the MOS Transistor
(a) Adopted conventions for voltages polarities and currents directions. (b) Symmetric decomposition of channel current into source and drain components.

the channel current.

6.2 Source- and Drain-Current Components

As shown in Section 5.2.1, for the parallel-plate capacitor approximation, both diffusion and drift are proportional to the charge-concentration gradient. This linear relation yields a quadratic expression for the current that consists of two symmetric, independent, opposing components (see Equation 5.10), as has been shown by other device physicists [125, 126, 123]. Therefore, the channel current can be decomposed into a **source component** and a **drain component**, as illustrated in Figure 6.1b.

One component is a function of the mobile charge at the source end of the channel; the other component is a function of the mobile charge at the drain end. Because the charge concentrations are related to voltages applied to the source terminal and to the drain terminal, respectively, as well as to the gate voltage, we can express the

channel current, per unit width, of an nMOS transistor in the form

$$I_{\mathrm{SD}} = \mathcal{P}(L)(\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}}, V_{\mathrm{SB}}) - \mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}}, V_{\mathrm{DB}})). \tag{6.1}$$

We can write a similar equation for the pMOS transistor in terms of $\mathcal{Q}_{\mathrm{p}}(V_{\mathrm{GB}}, V_{\mathrm{SB}})$.

As we saw in Section 5.2.2, the source component is related to the source voltage and the gate voltage by *exactly* the same function that relates the drain component to the drain voltage and the gate voltage. Therefore, which component we call the drain and which we call the source has nothing to do with the device—it is purely a question of which direction we prefer for the current in the device.

As we saw in Section 5.2.2, the charge–voltage relationship, $\mathcal{Q}_{\mathrm{n}}(V_1, V_2)$, has a complicated form, due to the exponential dependence of charge concentration on potential, and the compressive relationship between surface potential and gate voltage. These highly nonlinear dependencies arise because three different charged species are involved in the electrostatics. The complexity of the electrostatics obscures the simplicity of the drift–diffusion transport mechanisms that determine the charge–current relationship (Equation 6.1).

As shown in Section 5.2.1, in the subthreshold regime, the mobile-charge concentration grows exponentially as the gate voltage increases. This behavior is due to the linear relationship between gate voltage and surface potential when the mobile-charge concentration is much smaller than the depletion-layer charge. When the relationship is exponential, changing the gate voltage multiplies the mobile-charge concentrations at the source and the drain ends of the channel by the same amount. This multiplicative effect, together with the strictly linear relationship between current and charge in the diffusion-dominated subthreshold regime, allows us to factor out the dependence on the gate voltage. Hence, we can express the channel current per unit width in the form

$$I_{\mathrm{SD}} = \frac{\mathcal{K}_{\mathrm{n}}(V_{\mathrm{GB}})\mathcal{D}}{L}(\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{SB}}) - \mathcal{Q}_{\mathrm{n}}(V_{\mathrm{DB}})), \tag{6.2}$$

for an nMOS transistor; and similarly for a pMOS transistor in terms of $\mathcal{K}_{\mathrm{p}}(V_{\mathrm{GB}})$ and

$\mathcal{Q}_{\mathrm{p}}(V_{\mathrm{CB}})$, where

$$\mathcal{D} = \mu V_T, \tag{6.3}$$

$$\mathcal{K}_{\mathrm{n}}(V_{\mathrm{GB}}) = n_0 e^{\kappa_n V_{\mathrm{GB}}/V_T}, \tag{6.4}$$

$$\mathcal{K}_{\mathrm{p}}(V_{\mathrm{GB}}) = p_0 e^{-\kappa_p V_{\mathrm{GB}}/V_T}, \tag{6.5}$$

$$\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{CB}}) = -C_{\mathrm{dep}} V_T e^{-V_{\mathrm{CB}}/V_T}, \tag{6.6}$$

$$\mathcal{Q}_{\mathrm{p}}(V_{\mathrm{CB}}) = C_{\mathrm{dep}} V_T e^{V_{\mathrm{CB}}/V_T}. \tag{6.7}$$

6.3 Conditions for Symmetric Current Decomposition

In general, current decomposition is difficult to achieve for devices whose charge transport is governed by the drift–diffusion process. The difficulty arises because decomposition precludes any dependence of $\mathcal{P}(L)$ on the terminal voltages. Integrating the diffusionlike formulation of the Nernst–Planck equation (Equation 5.5), with the flux held constant, tells us that

$$\frac{1}{\mathcal{P}(L)} \propto \int_0^L e^{\psi(x)/V_{T_n}}\, dx.$$

Hence, $\mathcal{P}(L)$ is constant when the integral is constant. The preceeding integral is independent of the voltages applied at the ends of the channel only if the dependence of the potential profile on these voltages is constrained appropriately. This constraint is not satisfied in general; in particular, the cell membrane does not satisfy it. The constant-field assumption makes the potential change linearly along the channel, with its values at the ends equal to the voltages applied there. Hence, the equivalent permeability of the membrane is not independent of the voltages on either side of the membrane, as is evident in the solution for the membrane current (Equation 5.8).

Any voltage dependence of the nominally constant physical and geometrical properties of the device also violates decomposition. For the transistor, the susceptible

physical constants that are factored into $\mathcal{P}(L)$ and $\mathcal{D}$ are the mobility and the channel length L. The mobility degrades at high electric fields due to velocity saturation, and the effective channel length is reduced by the depletion layers at the source–bulk and drain–bulk junctions. The widths of these depletion layers depend on the voltages across those junctions. The channel width is also prone to modulation due to field fringing. Both velocity saturation and channel-length and channel-width modulation are negligible for devices with dimensions well over 1μm, but they become critically important for submicron devices.

The susceptible constant that is factored into $\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}}, V_{\mathrm{SB}})$ and $\mathcal{Q}_{\mathrm{p}}(V_{\mathrm{GB}}, V_{\mathrm{DB}})$ is the depletion-layer capacitance. This voltage-dependent capacitance is fairly constant below threshold, because the surface potential is virtually independent of the source and drain voltages below threshold—it is determined primarily by the gate voltage. Above threshold, the mobile-charge concentration is limited by the gate-oxide capacitance, which is constant. Therefore, the voltage dependence of the depletion capacitance does not limit decomposition.

Symmetry is violated by any changes in doping profile along the channel, or by differences in doping between the source–drain regions, such as in the lightly doped drain (LDD) structure employed in submicron devices. Symmetry is also violated by differences in area between the source and drain regions. When such differences are present, we can no longer use the same function, $\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_n}, V_i)$, to compute the terminal charge at both ends of the channel. Instead, we must use one function $\mathcal{Q}_{\mathrm{L_n}}(V_{\mathrm{GB}_n}, V_i)$, for the left end of the channel, and a different function, $\mathcal{Q}_{\mathrm{R_n}}(V_{\mathrm{GB}_n}, V_i)$, for the right end.

6.4 The Terminal Charge

I introduce the concept of **terminal charge** to exploit the inherent linearity of the MOS transitor. Each drain–source terminal is assigned a fictitious charge that is given by $\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_n}, V_i)$, where i is the label of the node to which the terminal is connected, and n is the label of the transistor to which the terminal belongs. Terminal charge

is negative for an nMOS transistor because electrons serve as charge carriers, and is positive for a pMOS transistor because holes serve as charge carriers.

In terms of these terminal charges, the current that flows from node i to node j, via transistor n, is in general given by

$$I_{i,j} = W_n \mathcal{P}(L_n)(\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_n}, V_i) - \mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_n}, V_j)), \tag{6.8}$$

from Equation 6.1. There is a perfectly *linear* relationship between the difference in terminal charges and the current, as though terminal-charge transport occurs by diffusion across a device with **permeability** $\mathcal{P}(L_n)$ per unit width. The effective permeability $W_n \mathcal{P}(L_n)$ is fixed because it depends on only physical constants, such as mobility, thermal voltage, and the width and the length of the channel—the designer specifies the channel width and the channel length.

In the special case, where we restrict operation to the subthreshold regime, the current is given by

$$I_{i,j} = W_n \frac{\mathcal{K}(V_{\mathrm{GB}_n})\mathcal{D}}{L_n}(\mathcal{Q}_{\mathrm{n}}(V_i) - \mathcal{Q}_{\mathrm{n}}(V_j)), \tag{6.9}$$

from Equation 6.2. Therefore, in subthreshold, the permeability can be factored into a **diffusitivity** $\mathcal{D}$, which is constant, and a **partition coefficient** $\mathcal{K}(V_{\mathrm{GB}})$, which is a function of the gate voltage. Thus, we can use the gate voltage to control how the charge concentration partitions between the source–drain regions and the channel, and thereby we can control the permeability of the device electronically. In contrast, the only way to change the permeability in the general case is to change the channel length or the channel width, and we cannot do that after the device has been fabricated.

I have shown how, in theory, we can transform the nonlinear MOS transistor into a linear element by performing a mapping $\mathcal{Q}(V_1, V_2)$ on the voltages applied to its terminals. We may achieve this linearity in practice if we adhere to the design and operation constraints under which symmetric current decomposition is valid, as

dicussed in Section 6.2.

6.5 Diffusors, Pseudoconductances, and Ohm's Law

I draw analogies between the MOS transistor operating below threshold and diffusion across a permeable membrane because Equation 6.9 arises from diffusion-dominated charge transport in the transistor. This mode of transport gives rise to an inherently linear charge–current relationship—unlike for drift-dominated transport. Therefore, when I use transistors that exploit this linear relationship in a circuit, I call them **diffusors** [5]. The analogy with the physical process of diffusion serves our intuition well, and allows us to make comparisons with neurobiology; I got the idea of using a transistor in this inherently linear fashion by making an analogy between the transistor and a gap junction [127].

The analogy between charge flow in a transistor and diffusion of uncharged particles across a porous membrane is perfect in the subthreshold regime, where transport is due primarily to diffusion, and the charge on the mobile carriers is negligible compared to the charge on the gate and the immobile charge in the depletion layer. Above threshold, however, transport is primarily due to drift, and the mobile charge is the dominant charge species. To the extent that the parallel-plate–capacitor model is valid, the derivative of charge concentration with respect to potential is constant, and therefore the charge-concentration profile along the channel is simply a scaled version of the potential profile. Therefore, at a microscopic level, we can model the charge transport as a diffusion process with a diffusion coefficient proportional to the local charge concentration, as shown by Equation 5.9, the diffusionlike formulation for the transistor. Therefore, the analogy is not disingenuous, as long as we bare in mind the proportionality between permeability and charge concentration above threshold. This dependence explains the macroscopic quadratic relationship between current and charge above threshold.

Other researchers have proposed viewing the transistor as a pseudoconductance, with a linear relationship between current and pseudovoltage [128, 129]. However, this

view disconnects us from the physics of the transistor. Conductance, or resistance, is a property of devices that obey Ohm's law, which states that the current density is proportional to the potential gradient. Transistors do not obey Ohm's law because the charge-carrier–concentration gradient along the channel is not zero. Linearity between potential gradient and current density holds only when the concentration is constant, and hence transport is due entirely to drift, and drift is linear at the macroscopic level when the charge-carrier–concentration profile is flat. However, in the above-threshold regime, where carriers drift, the concentration profile along the channel is not flat, and, in the subthreshold regime, carriers do not drift.

Actually, it is physically impossible to achieve linearity at the macroscopic level by satisfying Ohm's law at the microscopic level. Constant flux and constant concentration along the length of a conductor imply constant electric field. But this uniformity in both the charge density and the electric field is inconsistent with Gauss' law, which states that the electric field is the integral of the charge. Introducing an oppositely charged species to neutralize the charge does not produce linear behavior either, as we saw for the membrane. Carbon-film resistors, as well as other varieties of resistors, use thousands of elements connected in series to achieve linearity, by ensuring that the voltage drop across each element is smaller than the thermal voltage. Such small-signal operation makes the nonlinear behavior of these element irrelevant, effectively linearizing the element.

We can satisfy Ohm's law at the microscopic level by introducing another pair of terminals, in addition to the pair that conduct the current, as Tsividis and his colleagues have shown [130]. We achieve the correct microscopic behavior by placing voltage gradients on the gate and on the bulk that match the voltage gradient at the channel surface, such that $C_{\mathrm{ox}}(V_{\mathrm{G}}(x) - \psi_{\mathrm{s}}(x)) + C_{\mathrm{dep}}(\psi_{\mathrm{s}}(x) - V_{\mathrm{B}}(x))$ does not change with x. This arrangement makes the mobile-charge concentration, as well as the depletion charge and the hole charge, constant everywhere, thereby satisfiying Ohm's law for the entire range of operation of the device. The concentration of mobile charge can be controlled linearly above threshold by the potential difference between the gate and surface, or exponentially below threshold by the potential difference

between the gate and the source–drains. Thus, the conductivity of the material can be controlled electronically. We need additional circuitry, however, to make the pair of voltages applied to the ends of the gate, and the pair of voltages applied to the ends of the bulk, track the voltages on the source and the drain, and we must use a highly-resistive gate layer to limit power dissipation.

By thinking in terms of terminal charges and viewing the transistor as a diffusor, I created a simple circuit-level abstraction for the device. This viewpoint also gives us a powerful analogy between transistors and the porous membranes found in nerve cells. In particular, when operation is restricted to the subthreshold regime, we can control the permeability of the device by changing the gate voltage, which modulates the partition coefficient. Electronic control gives us the ability to model active properties of ion channels in the cell membrane. However, we must extend this abstraction to the circuit level, if we wish to use our abstract diffusors to build a real linear network.

6.6 The Node Charge

I introduce the concept of **node charge** to extend the device-level terminal-charge concept to the circuit level. It is possible to extend the latter concept to the circuit level if the **equivalence property** holds:

> All terminals connected to the same node have the same terminal charge.

Equivalence between node voltage and terminal charge allows devices to communicate their terminal charges using the voltage on the common node to which they are connected. When equivalence holds, we can replace device-level terminal charges with circuit-level node charges. Thus, equivalence allows us to exploit the linearity between terminal charge and current at the circuit level.

In the general case (Equation 6.1), equivalence holds if

$$V_i = V_j \;\Rightarrow\; \mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_m}, V_i) = \mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_n}, V_j)) \;\Rightarrow\; V_{\mathrm{GB}_m} = V_{\mathrm{GB}_n}.$$

Therefore, equivalence limits the ways in which transistors can be connected: Transistors connected to the same node cannot have different gate voltages. When two source–drains are connected together, the gates of the corresponding devices must be connected together as well—as must their bulks. Thus, for transistors that are part of the same circuit, all the gates must be connected together, and all the bulks must be connected together. For the special case of subthreshold operation (Equation 6.2), there is no such restriction, because $\mathcal{Q}_{\mathrm{n}}$ does not depend on the gate voltage in this region. In both regions of operation, however, equivalence precludes us from connecting together the drain–source terminals of devices of different type, and from connecting channel terminals to gate terminals.

Equivalence implies a one-to-one relationship between node charge and node voltage. Such an invertible relationship requires that terminal charge, $\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_n}, V_i)$, be a strictly monotonic function of terminal voltage, V_i. For the nMOS transistor, the mobile electrons decrease with increasing source or drain voltages. Hence, $-\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_n}, V_i)$ is a montonically decreasing function. And, for a pMOS transistor, the mobile holes increase with increasing source voltage. Hence, $\mathcal{Q}_{\mathrm{p}}(V_{\mathrm{GB}_n}, V_i)$ is a monotonically increasing function. These functions change either exponentially or quadratically with the voltage on the channel terminals.

Due to the expansive nature of $\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_n}, V_i)$, we have

$$\begin{aligned} V_i \gg V_j \quad &\Rightarrow \quad |\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_m}, V_i)| \ll |\mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_n}, V_j)| \\ &\Rightarrow \quad \left|\frac{d\mathcal{Q}_{\mathrm{n}}}{dV_i}\right| \ll \left|\frac{d\mathcal{Q}_{\mathrm{n}}}{dV_j}\right|. \end{aligned}$$

The same relationships apply to the pMOS transistor when the signs of the latter's voltages are reversed. Hence, we have

$$\begin{aligned} V_i \gg V_j \quad &\Rightarrow \quad I_{i,j} \approx I_j \equiv -W_n \mathcal{P}(L_n) \mathcal{Q}_{\mathrm{n}}(V_{\mathrm{GB}_n}, V_j) \\ &\Rightarrow \quad \left|\frac{dI_{i,j}}{dV_i}\right| \ll \left|\frac{dI_{i,j}}{dV_j}\right|. \end{aligned}$$

Therefore, when the voltage on node i is much larger than that on node j, the node charge at i becomes negligible, and the current asymtotically approaches I_j, the value of the component driven by node j.

In the dichotomous **ohmic–saturation** voltage-mode viewpoint, the device is said to enter a different regime of operation when the current becomes independent of the node voltage, called the **saturation region**. When the current is decomposed into source and drain components, however, there is no such dichotomy. For an nMOS transistor, we have

$$\begin{aligned} V_i > V_j + V_{\text{sat}} &\Rightarrow |I_i| \ll |I_j| \Rightarrow I_{i,j} \approx I_j, \\ V_j > V_i + V_{\text{sat}} &\Rightarrow |I_j| \ll |I_i| \Rightarrow I_{i,j} \approx I_i, \\ |V_i - V_j| < V_{\text{sat}} &\Rightarrow I_j \simeq I_i \Rightarrow I_{i,j} = I_i - I_j, \end{aligned}$$

where $V_{\text{sat}} = 5V_{T_n}$ below threshold, and $V_{\text{sat}} = V_{\text{GB}_n} - V_{\text{thr}}$ above threshold. The functional dependence of the current components, I_i and I_j, on the terminal voltage, V_i and V_j, is fixed, and remains the same throughout the ohmic and saturation regions. There is no dichotomy between these two regions from the current-component perspective: We split the current into two components, rather than split the voltage range into two regions. I prefer to split the current because this choice preserves the symmetry of the device, whereas splitting the voltage does not.

The linear dependence of the current on only one of the terminal charges in the saturation region gives us the capability to measure our fictitious terminal charges and node charges. Gaining access to the node charge is extremely important if we want to apply external inputs to the circuit, process them using the inherent linearity of the transistor, and read out the results from the circuit. Therefore, it is most convenient to use currents as our inputs and outputs, if we want to exploit the inherent linearity of the MOS transistor.

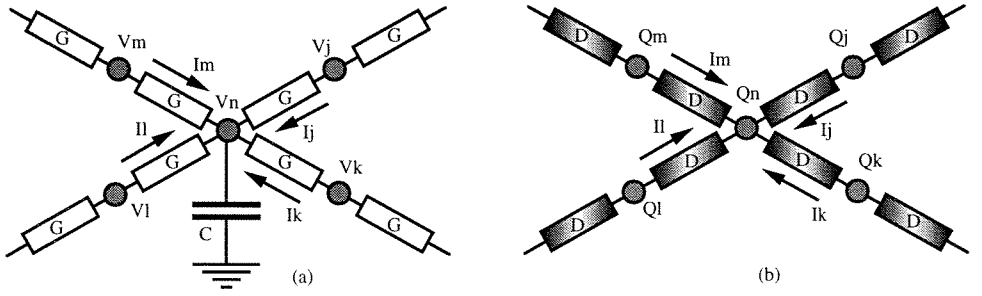


Figure 6.2: LOCAL AGGREGATION
(a) Aggregation using voltage–current linearity and conductances. (b) Aggregation using charge–current linearity and diffusors.

6.7 Diffusive Networks

Local aggregation—the linear summation of signals over a confined region of space—is a computation that occurs throughout the nervous system. A voltage-mode circuit that peforms this extremely useful computation is described by Mead (Chapter 6 of [122]); he uses this computation in several examples of neuromorphic systems presented in the book. In this section, I present a more efficient current-mode technique for performing local aggregation. My technique exploits diffusion in subthreshold MOS devices, much as cell syncytia in the nervous system use diffusion to distribute and sum signals over a local neighborhood.

The diffusion of particles through a continuous medium—or of heat in a solid—is described by the following partial differential equation:

$$\frac{dc}{dt} = D\nabla^2 c(x, y), \tag{6.10}$$

where $c(x, y)$ is the concentration profile over space—it is assumed to be uniform in the third (z) dimension. Here, $\nabla^2 \equiv \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the Lapalcian operator, and D is the diffusion coefficient. This equation is an application of Fick's law, which governs diffusion, and of the continuity equation, which guarantees conservation, as we discussed in Sections 5.1.1 and 5.1.2.

The discrete networks shown in Figure 6.2 both simulate diffusion in a continuous

medium. The first network (Figure 6.2a) uses voltages, currents, and conductances; its node equation is

$$\frac{dV_n}{dt} = \frac{4G}{C}\left(\frac{V_j + V_k + V_l + V_m}{4} - V_n\right), \tag{6.11}$$

which is homologous with Equation 6.10. The term in parentheses is the second-difference approximation to the Laplacian, when distance is measured in units of the internode spacing, as we discussed in Section 4.1. This solution, however, is not amenable to VLSI integration, because we must expend large amounts of area and power to make the nonlinear conductances of transistors appear linear over a voltage range larger than a few thermal voltages.

The second network (Figure 6.2b) uses charges, currents, and diffusors; its node equation is

$$\frac{dQ_n}{dt} = 4D\left(\frac{Q_j + Q_k + Q_l + Q_m}{4} - Q_n\right). \tag{6.12}$$

Note that dQ_n/dt is the same as the current supplied to node n by the network. This solution is amenable to VLSI implementation. We can realize diffusion with diffusors—transistors operating below threshold, as described in Section 6.6. The diffusion coefficient D is related to the diffusitivity, $\mathcal{D}$, and to the partition coefficient, $\mathcal{K}(V_{\mathrm{GB}})$, of the diffusors by

$$D = W\mathcal{K}(V_{\mathrm{GB}})\mathcal{D}.$$

In both of these networks, we can set up the boundary conditions by injecting current into the appropriate nodes. In the voltage-mode network, the solution is the node voltages, and we can read these voltages without disturbing the network. In the current-mode network, however, the solution is the node charges, and these fictitious charges are not directly accessible. We can infer the node charge from the node voltages V_i if we have an accurate description of $\mathcal{Q}(V_i)$. In practice, we can use a transistor to compute $\mathcal{Q}(V_i)$, by tying it to the node in question, and operating it in saturation so that it passes a current proportional to the node charge, as we discussed in Section 6.6. Unfortunately, this approach draws current from the node, and the

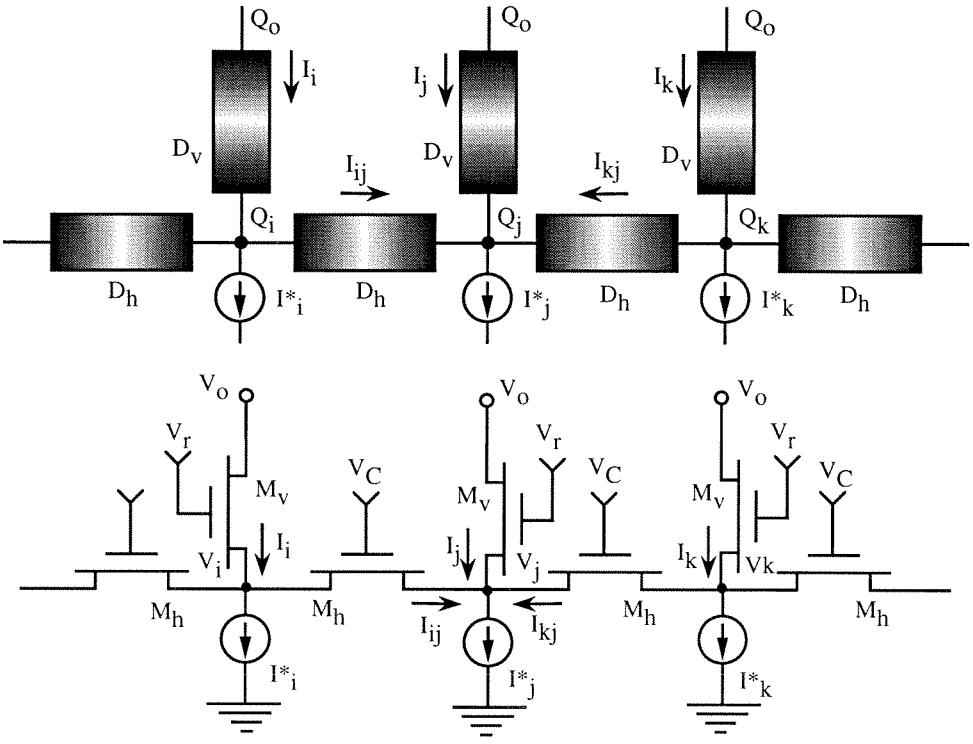


Figure 6.3: CELL-SYNCYTIA CIRCUIT MODEL
The lateral diffusors model gap junctions between cells; the vertical diffusors model the membrane leakage. (a) Schematic representation. (b) MOS transistor implementation.

permeability of the added device must be extremely small so that the disturbance is negligible.

Biological diffusive media, such as cell syncytia, are hardly ever isotropic (i.e., D varies from place to place). Nerve cells make gap junctions of varying area, and neuromodulators such as dopamine can vary the pore permeability. Thus, nerve cells can control actively the permeability of membranes between them and neighboring cells or the extracellular fluid. The dependence of the diffusor's partition coefficient on its gate voltage (see Equation 6.2) gives us the ability to control permeability locally in our circuit model of the diffusion network.

We can add a loss term to the diffusion equation to model the sequestering of

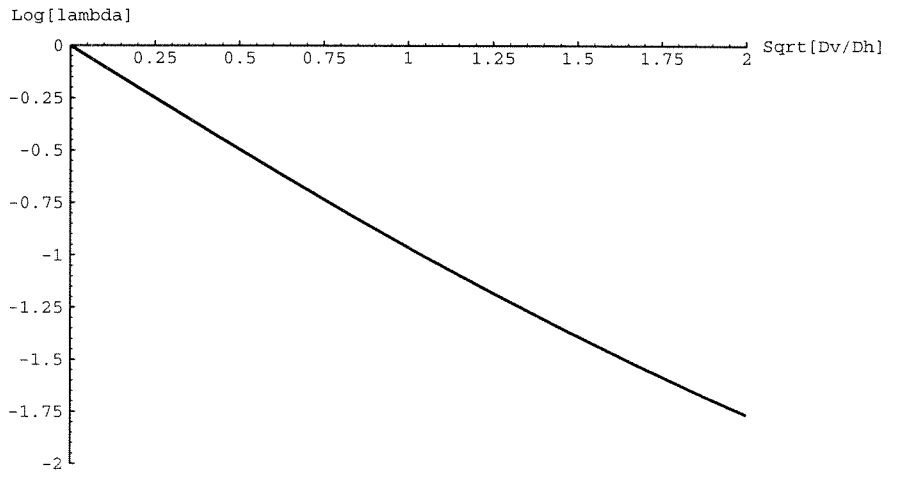


Figure 6.4: CONCENTRATION DECAY RATES FOR DIFFUSION
The exponential decay rate in the discrete network, $\log(\lambda)$, is plotted versus that in a continuous medium, $1/L = \sqrt{D_v/D_h}$, with the same values of vertical and horizontal diffusivities, D_v and D_h. The decay rates are nearly equal for low decay rates—the error is 12.5 percent when the charge density decays by a factor of e^2 each time that x changes by L.

particles by buffers or the leakage of particles out of the diffusive medium:

$$\frac{dc}{dt} = D_\mathrm{h}\nabla^2 c(x,y) - D_\mathrm{v}c(x,y). \tag{6.13}$$

This version of the diffusion equation is realized by the diffusor network shown in Figure 6.3, for the one-dimensional case. Its node equation is

$$I_j^* = I_{ij} + I_{kj} + I_j = \mathcal{K}(V_\mathrm{c})\mathcal{D}(Q_i + Q_k - 2Q_j) - \mathcal{K}(V_\mathrm{r})\mathcal{D}(Q_j - Q_0). \tag{6.14}$$

When $V_\mathrm{o} > \max_{\forall i}(V_i + V_\mathrm{sat})$, Q_0 is negligible and the node equation becomes the discrete analog of Equation 6.13. In this case, we can read out the node charges simply by monitoring the currents at the drains of the vertical elements as these currents flow into the voltage source applying V_o.

The discrete version of the lossy-diffusion equation admits solutions of the form

$Q_i = \lambda^i$, where

$$\lambda = 1 + \frac{D_v}{2D_h}\left(1 - \sqrt{1 + \frac{4D_h}{D_v}}\right),$$

and $D_v \equiv \mathcal{DK}(V_\mathrm{r})$, and $D_h \equiv \mathcal{DK}(V_\mathrm{c})$. The solution to the continuous lossy-diffusion differential equation in one dimension has the form $Q(x) = \exp(-x/L)$, where

$$L = \sqrt{\frac{D_h}{D_v}}.$$

Thus, the discrete simulation reproduces the exponential form of the decay; its decay rate matches that of the continuous medium when $\log(\lambda) \approx 1/L$. These quantities are plotted against each other in Figure 6.4; the slope is close to unity for low decay rates.

Using the exponential dependence of the partition coefficients on the gate voltages, we can relate the space constant to the biases voltages V_r and V_c applied to the diffusors:

$$L \approx \exp\left(\frac{\kappa_n(V_\mathrm{c} - V_\mathrm{r})}{2V_T}\right).$$

It becomes obvious why the ratio has an exponential dependence on the voltage difference between V_c and V_r if you observe that M_h and M_v constitute a differential pair operating in subthreshold. These devices act as a current-divider for current driven by the charge at their common node. The divider ratio is set by their effective widths, which depend on the geometrical width as well as on the surface potential. Here, we have used the κ approximation to relate the surface potential to the gate-bulk voltage. The surface potential is constant as long as the gate and bulk voltages are fixed—assuming that the mobile charge is negligible. Therefore, the divider ratio is constant, and linear division occurs. However, as we enter the transition and above-threshold regions, this assumption fails, and the surface potential starts to follow the source voltage. Consequently, the divider ratio is no longer independent of the current level. This variation of the divider ratio limits the dynamic range of the current divider, and hence the linear operating range of the diffusor network.

The diffusor network is a particularly attractive circuit for implementing local

aggregation because of the area efficiency that we realize by using a single device to model a linear element, the power efficiency that we obtain by operating with subthreshold currents, and the enhanced functionality available with electronically adjustable coupling strength.

The diffusive network in Figure 6.3 has been described in terms of pseudoconductances [129]. I prefer the charge-based formulation using diffusors, originally proposed in [5] and elaborated in [131, 132], because of the physically accurate intuition that it provides. The essence of this approach is the representation of variables and parameters by charge, current, and diffusivity—voltages and conductances are not used explicitly.

Bult and Geelen proposed an identical network for linear current division above threshold, and used it in a digitally controlled attenuator [128]; they also analyzed its subthreshold behaviour. However, they stipulated that all gate voltages must be identical, and controlled the division by manipulating the geometrical factor W/L of the devices. I showed here, and previously in [5], that this constraint can be relaxed in subthreshold without disrupting linear operation. This flexibility is a real bonus, because it allows us to modify the divider ratio or space constant of the network after the chip is fabricated by varying $V_c - V_r$. Tartagni and colleagues have demonstrated a current-mode centroid network [133] using subthreshold MOS devices whose operation is described by the diffusors discussed here.

6.8 Test Results

In this section, I present results from experiments designed to demonstrate the linearity of diffusor circuits, and to measure the dependence of diffusitivity on the gate voltage, the dynamic range of operation, and the spread of currents in diffusor networks.

The circuit designed to test the functionality of the current divider and the diffusor is shown in Figure 6.5. The measurements obtained from this test circuit are shown in Figures 6.6 and 6.7. These measurements demonstrate the linearity of the current

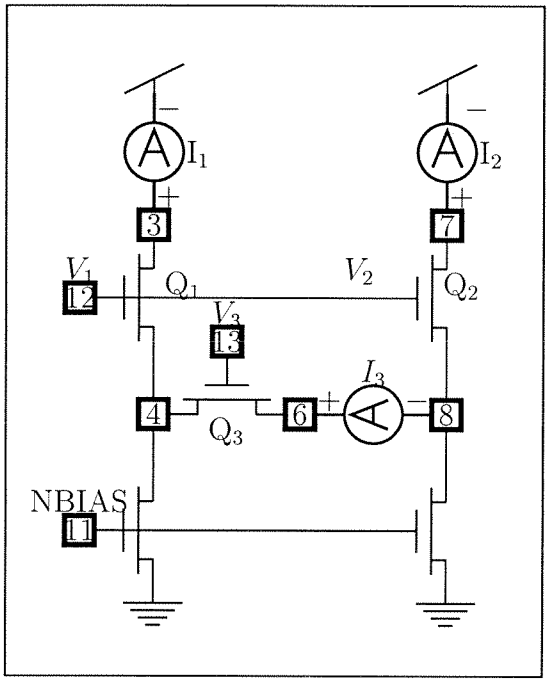


Figure 6.5: Current Divider and Diffusor Test Circuit

Using this test circuit, I measured current division in the differential pair formed by transistors Q_1 and Q_2 when Q_3 is shorted, and the linear dependence of the current, I_3, in transistor Q_3 on the current differential, $I_1 - I_2$, between transistors Q_1 and Q_2. I have plotted the current-divider measurements, taken for various voltage differentials $(V_1 - V_2)$, in Figure 6.6, and the diffusor measurements, taken for various voltage differentials ($VR - VG$, where $VR = V_3$, $VG = V_1 = V2$), in Figure 6.7.

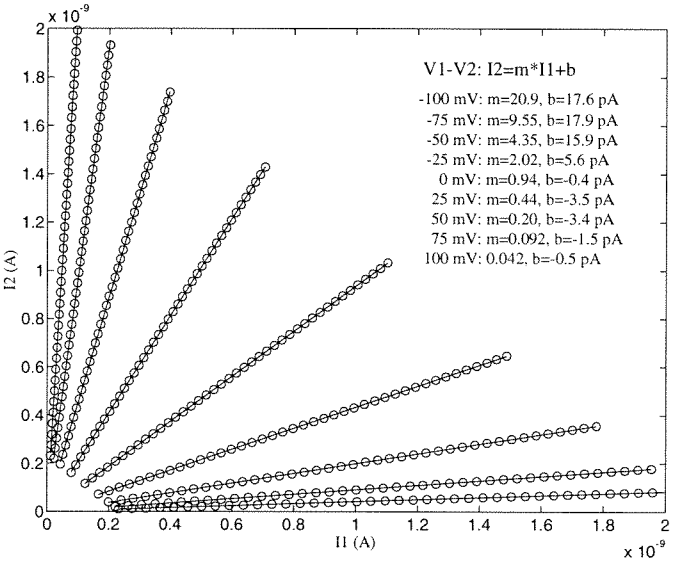


Figure 6.6: CURRENT-DIVIDER CURRENTS

These measurements, obtained from the test circuit shown in Figure 6.5, demonstrate that the differential pair splits its tail current between its two arms according to a fixed ratio, and this ratio depends on the voltage differential. The ratios $I1/I2$ obtained from the slopes of these curves are plotted against the voltage differentials $V1 - V2$ in Figure 6.8.

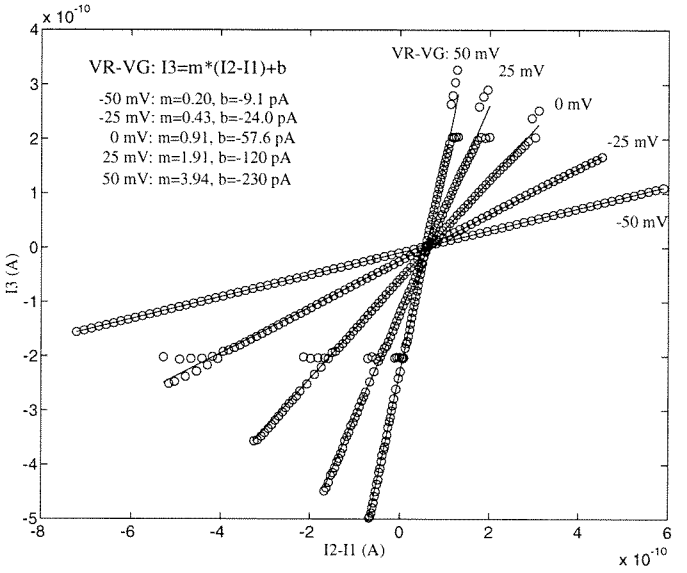


Figure 6.7: Diffusor Current versus Current-Difference

These measurements, obtained from the test circuit shown in Figure 6.5, demonstrate that the current in the horizontal diffusor is directly proportional to the current differential in the vertical diffusors, and the ratio $I3/(I1 - I2)$ depends on the voltage differential $VR - VG$ between the horizontal and vertical diffusors. The ratios obtained from the slopes of these curves are plotted against the voltage differential in Figure 6.9.

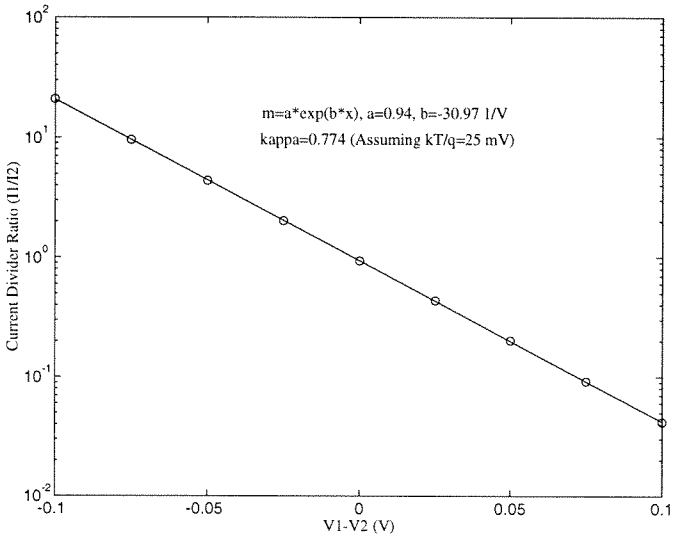


Figure 6.8: CURRENT-DIVIDER RATIO VERSUS VOLTAGE DIFFERENTIAL
These measurements demonstrate an exponential dependence of the current-divider ratio on the voltage differential, just as predicted by the theory.

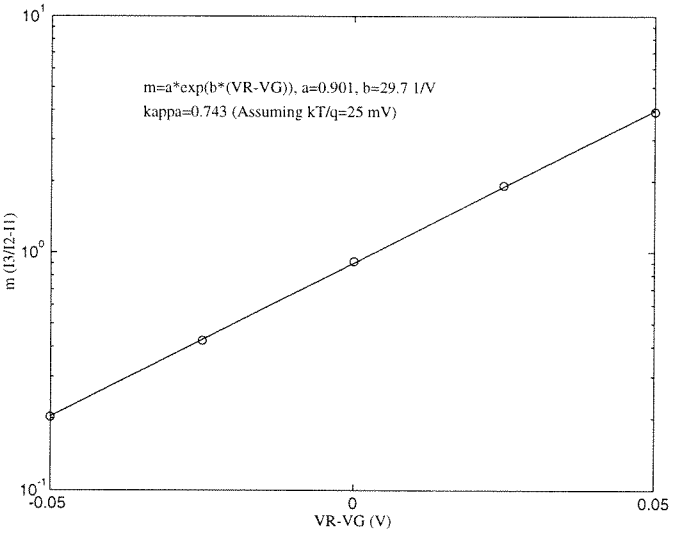


Figure 6.9: DIFFUSOR PERMEABILITY VERSUS GATE VOLTAGE
These measurements demonstrate an exponential dependence of permeability on gate voltage, just as predicted by the theory.

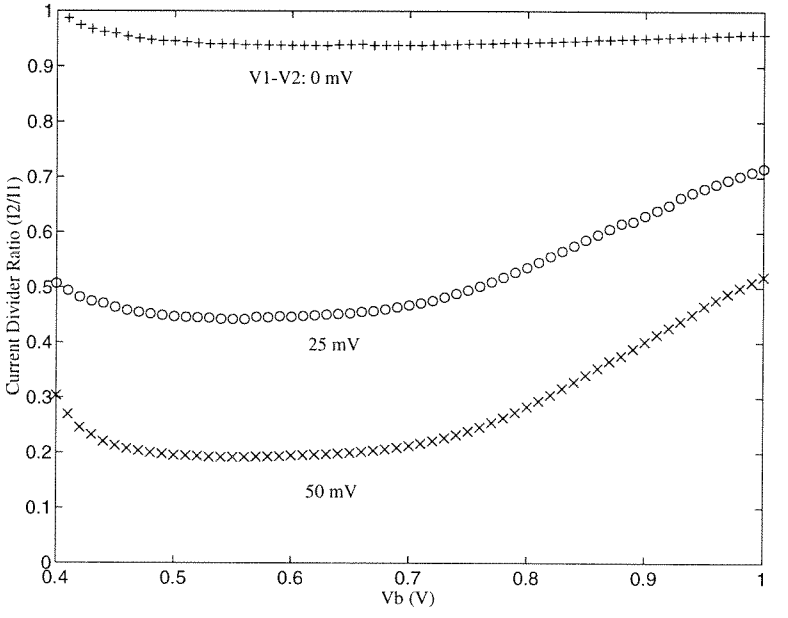


Figure 6.10: DYNAMIC RANGE OF CURRENT DIVIDER CIRCUIT

These measurements, which I took by sweeping the voltage applied to pin NBIAS of the test circuit shown in Figure 6.5, and taking the ratio between the currents in $Q1$ and $Q2$, with $Q3$ shorted, for various voltage differentials $V1 - V2$, show the limited range of operation of current divider, and diffusor, circuits.

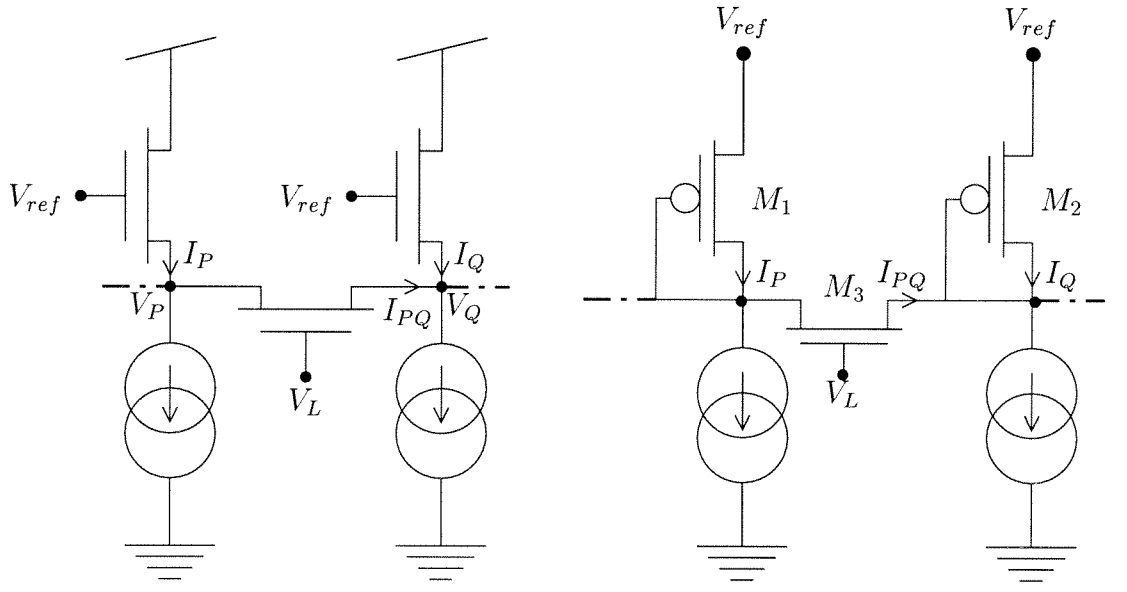


Figure 6.11: Linear and Nonlinear Diffusor Circuits
The vertical elements in the ideal diffusor network shown in (a) can be replaced by diode-connected complementary devices, as shown in (b). Measurements from these two circuits are shown in Figure 6.12.

divider and the diffusor, thereby confirming the basic functionality of these circuits. The dependence of the current-divider ratio and of the diffusor's permeability on the gate voltages is characterized by the plots in Figures 6.8 and 6.9, which confirm the exponential dependence predicted by the theory.

Finally, to check the range of current levels over which the operation of these circuits is linear, I swept the tail current of the differential pair and measured the ratio of the currents in its two arms, at three different voltage differentials. These data are plotted in Figure 6.10; the ratio was constant for bias voltages ranging from 0.45V to 0.70V, correspond to about two decades of current. Outside this range, the ratio approached unity, due to current levels approaching the leakage levels in the experimental set up, at one extreme, and approaching above-threshold levels, at the other extreme. When the voltage differential was zero, however, the ratio was constant over the entire range. In this degenerate case, current division is determined entirely by the device geometries.

Experimental data from diffusor networks are shown in Figure 6.12, for the two

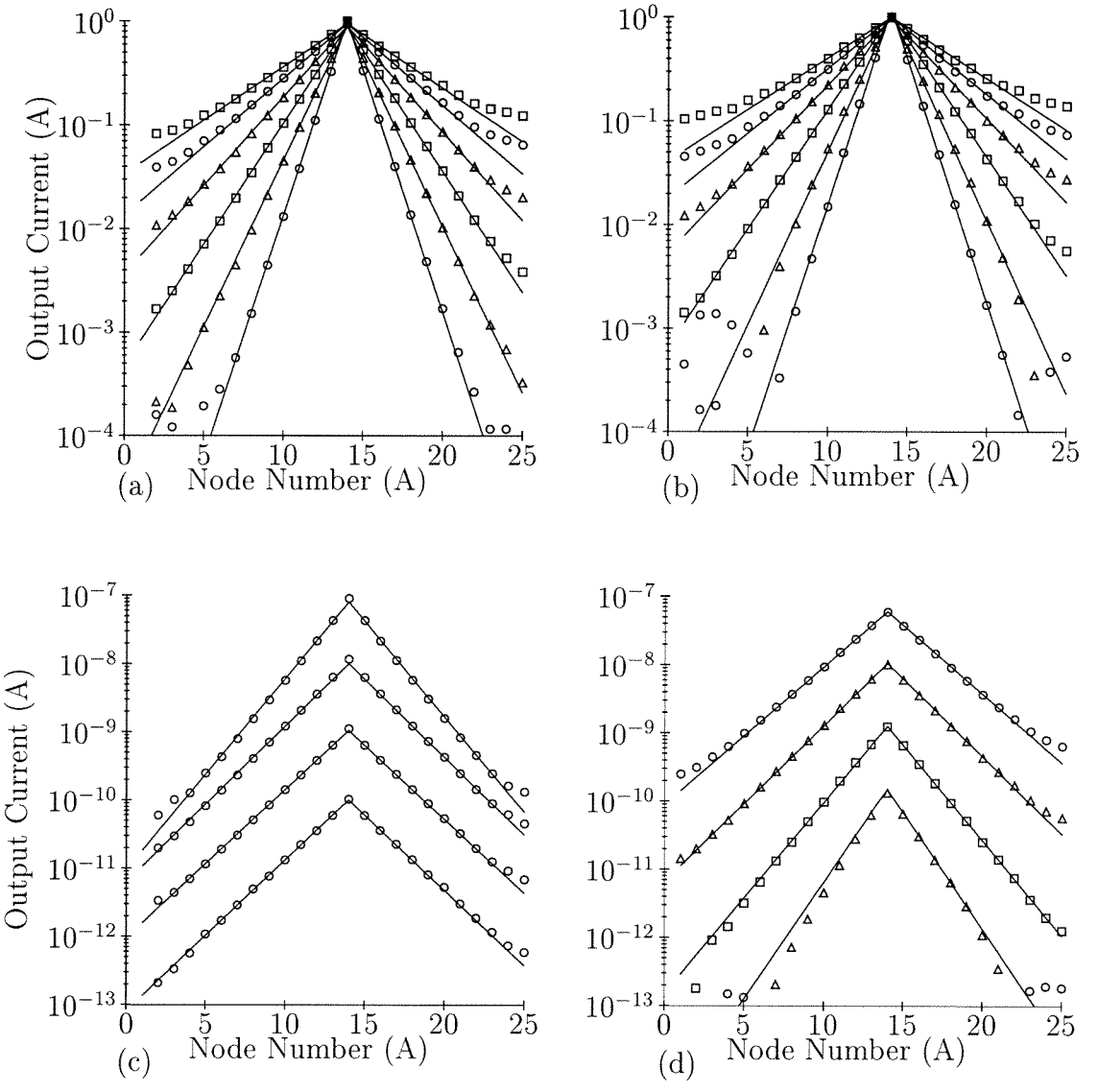


Figure 6.12: Current Spreading in Diffusor Networks

The curves on the left and on the right are for ideal and nonideal diffusor networks, respectively. The ideal network (shown in Figure 6.11a) is built entirely out of nMOS transistors; the nonideal network (shown in Figure 6.11b) uses nMOS transistors for the horizontal-spreading elements and diode-connected pMOS devices for the vertical-leakage elements. (a, b) The exponential decay of an input current applied to the center node of a one-dimensional network. The space constant increases as $V_c - V_r$ increases by 25mV from one curve to the next. The deviations at larger space constants are due to boundary effects. (c, d) The effect of varying the input level with $V_c - V_r$ constant. The source diffusor network's space constant is independent of input level—the hallmark of a linear circuit—except when currents approach above-threshold levels. In the second network, the space constant inceases with input level, because the currents in the horizontal nMOS transistors increase more rapidly with voltage than do the currents in the vertical pMOS transistors, due to the effect of κ on the gate voltages of the pMOS devices.

kinds of networks shown in Figure 6.11. The first network (Figure 6.11a) shows linear behavior for subthreshold current levels, as we expect from the equivalence of the terminal charges of the source–drain terminals of the vertical and horizontal elements. The second network (Figure 6.11b) shows weakly nonlinear behavior; its space constant increases slowly as the current levels increase. This nonlinear behavior arises because equivalence is violated. The terminal charges of the gate terminals of the vertical pMOS transistors are not equivalent to the terminal charges of the source–drain terminals of the horizontal nMOS transistors.

Using the subthreshold current–voltage relationships, and assuming that the pMOS device is in saturation, we can show that [5]

$$I_{PQ} = e^{\kappa V_L - V_{ref}}(I_Q^{1/\kappa} - I_P^{1/\kappa}). \tag{6.15}$$

Hence, there is an expansive relationship between the horizontal current I_{PQ} and the vertical currents I_P and I_Q. Expansion occurs because, as we change the voltages V_P and V_Q, the currents in the horizontal elements increase faster than do the currents in the vertical elements, as the source is more effective at changing the current than the gate is—by a factor of $1/\kappa$ on a log plot. When κ is close to unity, the gate and source terminals are nearly equivalent, and the behavior becomes almost linear. As we shall see in Chapter 7, however, we are forced to couple diffusor networks to other networks using gates if we wish to model the effects of chemical synapses. In such cases we can push κ close to one by increasing the source–bulk voltage. This technique can yield values close to 0.95 as shown in the experimental data plotted in Figure 5.6.

6.9 Summary

We have seen how to extend the simple and elegant description of device operation in terms of current and charge to the circuit level, using the equivalence principle and the concept of node charges. When node charges stand in for membrane voltages, we

may model the linear current–voltage relationship of the gap junction with the linear current–charge relationship of transistors in the subthreshold regime. This analogy enables us to simulate the spread of ions in cell syncytia extremely efficiently.

We can use diffusors to model the lateral spread of these ions, as well as the loss of ions through leakage into the extracellular fluid. These two mechanisms define a local neighborhood over which signals summate, and we can control the size of this region by the relative strengths of the lateral coupling between nodes in the network and the leakage path from these nodes to ground. When we use diffusors, we can control the size of this region electronically, and thereby we can actively regulate local aggregation. The extent of local aggregation determines the extent of collective computation. Cell syncytia can regulate the extent of local aggregation as well. The retina exploits this ability to trade off signal-to-noise ratio for bandwidth, as we discussed in Chapter 3.

Chapter 7 Neuromorphing: From Neural Circuits to CMOS Circuits

In this chapter, I show how we can use neural circuits as blueprints for VLSI CMOS circuits. By replacing each neuron with a single-cell circuit model and each synapse with a synapse circuit model, and connecting everything correctly, we can transform neural circuits into CMOS circuits. This process is called **neuromorphing**.[1] Once we have silicon-based circuit modules for the single cells, and for electrical synapses (gap junctions) and chemical synapses, neuromorphing is straightforward: We simply replace each nerve cell by our single-cell module and connect these subcircuits together with the appropriate synapse modules.

We already have single-transistor models of ion channels. In Chapter 5, we saw that each ion-channel population in the cell can be modeled by a single transistor, because the ion channel's nonlinear current–voltage relationship, for large concentration gradients, is similar to the transistor's current versus drain–source voltage relationship. In Chapter 6, we saw that a gap junction can be modeled with a single transistor as well, because the gap junction's linear current–voltage relationship is similar to the transistor's current–charge relationship in the subthreshold regime.

I show here how we can model excitatory and inhibitory chemical synapses with single transistors. Together with the single-transistor model of gap junctions, I use these neural analogs to morph the neurocircuitry in the outer retina into silicon. The result is a CMOS circuit that models bandpass spatiotemporal filtering in the outer retina—at the same level of abstraction as the linear electrical circuit model that we studied in Chapter 4. In contrast to the linear physical model in Chapter 4, the CMOS circuit includes a local gain-control mechanism. This nonlinear mechanism

[1] As far as I know, this verb-form of the word *neuromorphic* first appeared in a report written by Muriel Ross and her coworkers on the "From Neurons to Nanotechnology" workshop sponsored by NASA. The workshop was held at Moffet Field in October 1995.

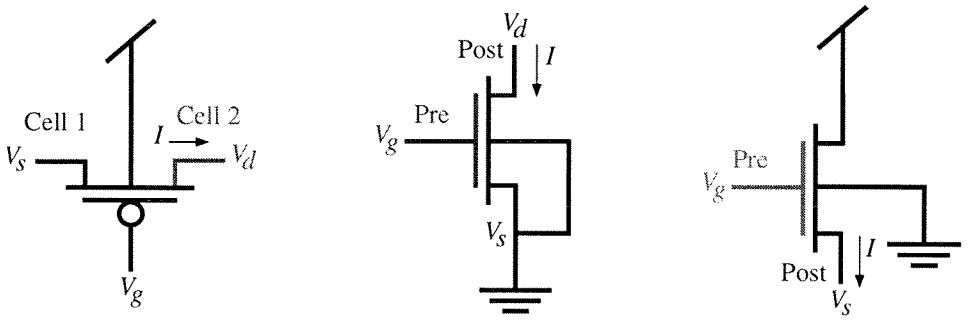


Figure 7.1: SINGLE-TRANSISTOR MODELS OF SYNAPSES
We model gap junctions, inhibitory synapses, and excitatory synapses by connecting a transistor between two nodes, as shown here. The voltages on these nodes represent the membrane potentials of the pair of cells that communicate via the synapse.

models the effect of shunting inhibition from the horizontal cells to the cones.

Unlike the abstract theoretical circuit model, the actual parameters of nominally identical circuit elements on the chip vary from location to location, due to the vagaries of the fabrication process. Consequently, building the model in silicon helps us to understand the effects of structural peturbations and quantum fluctuations on performance, as well as the effects of local gain control on bandpass filtering. It also forces us to address structural constraints, such as the energy and area costs of communication versus computation, which we discussed briefly in Section 4.6.2.

7.1 Modeling of Excitatory and Inhibitory Chemical Synapses

By making the same simplifying assumptions that we used to obtain a linear electrical circuit model of the outer retina in Chapter 4, we can model excitatory and inhibitory synapses, as well as gap junctions, using a single transistor.

In particular, we assume that these ligand-gated ion channels have extremely low conductances, compared to the combined conductance of all the other current-conducting elements in a cell membrane. Therefore, an excitatory synapse is equiva-

lent to a current source in the postsynaptic membrane, controlled by the membrane voltage of the presynaptic cell; the magnitude of the current increases as the voltage increases. Similarly, an inhibitory synapse is equivalent to a current sink in the postsynaptic membrane, controlled by the membrane voltage of the presynaptic cell; the magnitude of the current increases as the voltage increases.

We can realize these extremely simple abstractions of excitatory and inhibitory chemical synapses by connecting a transistor between circuit nodes that represent the pre- and postsynaptic cells, as shown in Figure 7.1. This figure also shows how we model a gap junction between two cells using a diffusor.

The postsynaptic currents can be expressed in terms of the node charges as follows:

$$\begin{aligned} I_{\text{junc}} &= \mathcal{K}_{\text{p}}(V_{\text{g}})\mathcal{D}(\mathcal{Q}_{\text{p}}(V_1) - \mathcal{Q}_{\text{p}}(V_2)), \\ I_{\text{inhib}} &= -\mathcal{K}_{\text{n}}(V_{\text{pre}})\mathcal{D}\mathcal{Q}_{\text{n}}(V_{\text{Gnd}}), \\ I_{\text{excit}} &= -\mathcal{K}_{\text{n}}(V_{\text{pre}})\mathcal{D}\mathcal{Q}_{\text{n}}(V_{\text{post}}); \end{aligned}$$

assuming that the synapse transistors are in saturation. We can express the partition coefficients in terms of the node charges using the expressions given in Section 6.2 (Equations 6.4 and 6.7); these expressions are repeated here, with all terminal voltages referred to a common reference, which we call ground (Gnd), and with bulks tied to sources (i.e., $V_{\text{PBB}} = V_{\text{DD}}$ and $V_{\text{NBB}} = V_{\text{SS}}$) :

$$\mathcal{K}_{\text{n}}(V_{\text{G}}) = n_0 e^{\kappa_n (V_{\text{G}} - V_{\text{SS}})/V_T}, \tag{7.1}$$

$$\mathcal{K}_{\text{p}}(V_{\text{G}}) = p_0 e^{-\kappa_p (V_{\text{G}} - V_{\text{DD}})/V_T}, \tag{7.2}$$

$$\mathcal{Q}_{\text{n}}(V_{\text{C}}) = -Q_T e^{-(V_{\text{C}} - V_{\text{SS}})/V_T}, \tag{7.3}$$

$$\mathcal{Q}_{\text{p}}(V_{\text{C}}) = Q_T e^{(V_{\text{C}} - V_{\text{DD}})/V_T}; \tag{7.4}$$

where $Q_T \equiv C_{\text{dep}}(V^*_{\text{GB}})V_T$.

At the circuit level, we define the node charges to be equivalent to the source–drain terminal charges of the pMOS transistors, and we obtain expressions for the synaptic currents entirely in terms of these arbitrarily defined node charges. First, we express

$\mathcal{K}_{\rm n}(V_{\rm pre})$ and $\mathcal{Q}_{\rm n}(V_{\rm post})$ in terms of $\mathcal{Q}_{\rm p}(V_{\rm pre})$ and $\mathcal{Q}_{\rm p}(V_{\rm post})$; then, we substitute the resulting expressions into the current equations. The results are

$$
\begin{aligned}
i_{\rm junc} &= \alpha_{\rm p}(V_{\rm g})\,(q_1 - q_2), && (7.5)\\
i_{\rm inhib} &= \beta_{\rm p}(V_{\rm DD} - V_{\rm SS})\,q_{\rm pre}^{\kappa_{\rm n}}, && (7.6)\\
i_{\rm excit} &= \beta_{\rm p}(V_{\rm DD} - V_{\rm SS})\,q_{\rm pre}^{\kappa_{\rm n}}/q_{\rm post}, && (7.7)\\
\alpha_{\rm p}(V_{\rm g}) &\equiv p_0\, e^{-\kappa_p(V_{\rm g}-V_{\rm DD})/V_T}, && (7.8)\\
\beta_{\rm p}(V_{\rm dd}) &\equiv n_0\, e^{V_{\rm dd}/V_T}, && (7.9)
\end{aligned}
$$

when charge is given in units of $Q_T \exp(-(V_{\rm DD} - V_{\rm SS})/V_T)$.

As expected, the gap-junction current is proportional to the node-charge difference. In contrast, the inhibitory and excitatory currents are a slightly compressive function of presynaptic node charge, because $\kappa_{\rm n}$ is less than 1. The excitatory current is also shunted (reduced) by postsynaptic activity, because it is inversely proportional to the postsynaptic node charge.

The roles of the nMOS and pMOS transistors can be interchanged; that is, we may use nMOS transistors for the gap junctions, pMOS transistors for chemical synapses, and the source–drain terminal charges of the nMOS transistors for the node charges. We can use Equations 7.5 and 7.7 for these complementary assignments by reversing the directions of the currents, replacing κ_n with κ_p, and $-\kappa_p$ with κ_n, and interchanging $V_{\rm SS}$ and $V_{\rm DD}$, and n_0 and p_0.

7.2 Outer-Plexiform–Layer Circuit

The neurocircuitry in the outer retina that subserves the cone pathway is shown in Figure 7.2. This simplified schematic includes only a single horizontal-cell type and a single cone type. It is based on the red-cone system of the turtle; similar circuitry is found in all vertebrate retinae. This simple neural circuit implements elegantly both local gain control and bandpass spatiotemporal filtering.

We employ **shunting inhibition** to compute a normalized output that is pro-

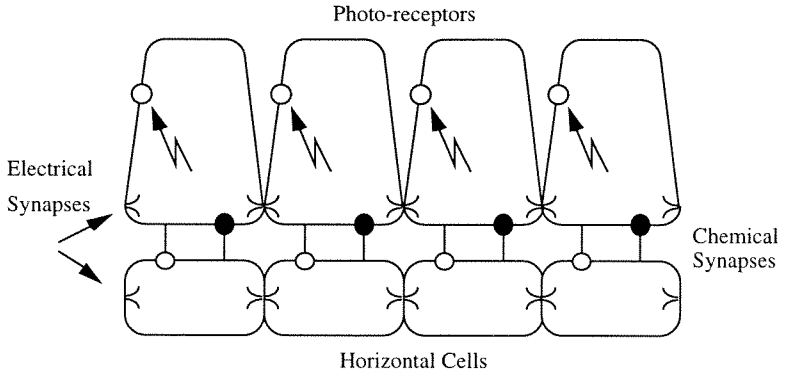


Figure 7.2: NEUROCIRCUITRY OF OUTER PLEXIFORM LAYER

The white and black circles are excitatory and inhibitory chemical synapses, respectively. Electrical gap junctions occur at the points of contact between cells. The photoreceptors are activated by light; they produce activity in the horizontal cells through excitatory chemical synapses. The horizontal cells reciprocate by suppressing the activity of the receptors through inhibitory chemical synapses. Receptors and horizontal cells are electrically coupled to their neighbors by gap junctions; these junctions allow ionic currents to flow from one cell to another. The junctions between horizontal cells are larger in area than those between cones.

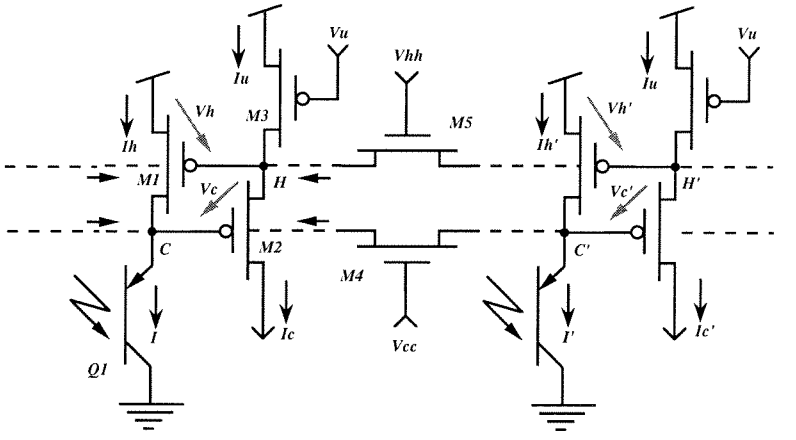


Figure 7.3: CMOS CIRCUIT MODEL OF OUTER PLEXIFORM LAYER

There are two diffusive networks coupled together by synapse transistors. The diffusive networks model the syncytia formed by the cone and horizontal-cell layers, and the synapse transistors model the excitatory and inhbitory synapses between these two layers. Nodes in the upper layer correspond to horizontal cells; nodes in the lower layer correspond to cones. Devices M_1 and M_2 model chemical synapses; M_4 and M_5 model gap junctions. Q_1 models the outer segment of the cone, and M_3 models a leak in the horizontal-cell membrane.

portional to contrast. The horizontal cells pool their signals to obtain a measure of the local light-intensity level, and they modulate conductances in the membranes of the cones proportionally. Because the current supplied by the cone outer segment is divided by this conductance to produce the membrane voltage, the cone's response is proportional to the ratio between its photo input and the local average. This description is a simplified abstraction of the complex ion-channel dynamics involved.

Computing contrast at the very first cell in the retina is extremely advantageous to the animal, because using this representation extends the dynamic range of the sensor and removes redundant information. At any particular background-intensity level, however, the outer plexiform layer behaves much like a linear system, and bandpass filters the image in space and time—more or less like the linear coupled-layer network that we studied in Chapter 4.

Neuromorphing the neurocircuitry of the outer-plexiform layer shown in Figure 7.2 into a CMOS circuit gives us the circuit shown in Figure 9.4. We have replaced each cell with a single node, and each synaptic element with a single transistor. Devices M_1 and M_2 model the reciprocal chemical synapses, and M_4 and M_5 model the gap junctions; their permeabilities are set globally by the bias voltages V_G and V_F. The phototransistor M_6 models the light-sensitive input from the cone outer segment. Transistor M_3, with a fixed gate bias V_U, is analogous to a leak in the horizontal-cell membrane that counterbalances synaptic input from the cone.

In this simple model, we reduce a single cell to a node in the electrical circuit, ignoring the nonsynaptic membrane conductances and the effect of the cell's morphology on its electrotonic properties. This simplification is valid for the special case where the membrane conductances are small compared to the synaptic conductances and the cell is electrotonically compact.

The CMOS analog of the outer retina operates as follows. When the photocurrent I increases, the excess current discharges node C; this effect represents excitation. The decrease in voltage at C increases the current in M_2. The excess current in M_2 discharges node H, so this node also is excited. The decrease in voltage at H increases the current in M_1. The excess current in M_1 counterbalances the excess photocurrent,

and tends to restore node C to its original voltage. Thus, M_1 inhibits the effect of the photocurrent. In short, M_2 mimics excitation from the cone layer (node C) to the horizontal-cell layer (node H), and M_1 mimics inhibition from the horizontal-cell layer (node H) to the cone layer (node C).

When the inhibitory network is biased such that its time constant and space constant are longer than those of the excitatory network, signals that change rapidly over time or space will escape the inhibition, resulting in a highpass frequency response. However, the response starts to roll off when the period approaches the time and space constants of the excitatory network, resulting in an overall bandpass response in spatial and temporal frequency. This aspect of the circuit's behavior is described by the linear model that we analyzed in Chapter 4.

The lowpass-filtered version of the image from the inhibitory layer is a measure of the local intensity level. Therefore, the OPL circuit uses this signal for gain control. The gain is controlled by the shunting effect of the node charge at node H on the excitatory synapse from node C to node H. This shunt makes the synaptic current I_C proportional to the ratio of the node charge at C (cone's activity) and the node charge at H (horizontal cell's activity). Therefore, I_C is proportional to the local contrast, and it provides a normalized signal that serves as the output of the OPL circuit.

We can write the node equations for this circuit using the continuous version of the diffusive network given in Section 6.7 (Equation 6.13). The result is

$$\begin{aligned} I_h(x,y) + D_{\mathrm{c}}\nabla^2 Q_{\mathrm{c}}(x,y) &= I(x,y), \\ I_u + D_{\mathrm{h}}\nabla^2 Q_{\mathrm{h}}(x,y) &= I_c(x,y), \end{aligned}$$

in steady state. Substituting the expressions for the synaptic currents and the diffusitivities given in Section 7.1 (Equation 7.5 through Equation 7.7) gives us

$$\beta_{\mathrm{n}} q_{\mathrm{h}}^{\kappa_{\mathrm{p}}} + \alpha_{\mathrm{cc}}\nabla^2 q_{\mathrm{c}} = i, \tag{7.10}$$

$$i_u + \alpha_{\mathrm{hh}}\nabla^2 q_{\mathrm{h}} = \beta_{\mathrm{n}} q_{\mathrm{c}}^{\kappa_{\mathrm{p}}}/q_{\mathrm{h}}, \tag{7.11}$$

where $\alpha_{cc} \equiv \alpha_n(V_{cc})$, $\alpha_{hh} \equiv \alpha_n(V_{hh})$; with charge expressed in the appropriate units.

Equations 7.10 and 7.11 are homologous with the equations for the linear model that we studied in Chapter 4 (Equation 4.1) when $\kappa_p = 1$, except that node charges, instead of node voltages, are used to represent the cell activities, and the signs of the synaptic currents and the input currents have been flipped. Nevertheless, we can use the solutions that we obtained for the linear network with

$$\begin{aligned} g_{ch} &= \beta_n, \\ g_{hc} &= \beta_n / q_h, \\ 1/r_{cc} &= \alpha_{cc}, \\ 1/r_{hh} &= \alpha_{hh}. \end{aligned}$$

The OPL CMOS circuit is linear if q_h is constant, and is approximately linear over regions where q_h changes slowly relative to q_c. The space constant of the horizontal-cell layer is much longer than that of the cone layer, so this assumption is reasonable.

Hence, the output current of the OPL circuit is given by the following expression:

$$i_c \approx \beta_n \frac{q_c}{\langle q_h \rangle} = \beta_n \frac{i}{\langle i \rangle} \frac{\ell_h^2 \rho^2}{\ell_c^2 \ell_h^2 \rho^4 + 1}; \tag{7.12}$$

$$\ell_c = \sqrt{\alpha_{cc}/\beta_n}, \tag{7.13}$$

$$\ell_h = \sqrt{\alpha_{hh} q_h / \beta_n},$$

in steady state, where $\langle i \rangle$ is the local average of the photocurrent. Notice that local gain control actually takes out the absolute intensity, and the output current is proportional to the ratio between the photocurrent and its local average.

The dependence of the horizontal-cell layer's space constant ℓ_h on q_h makes the spatial filtering dependent on the intensity level. In particular, if we use the results from Section 4.3, the position and height of the peak in the bandpass characteristic

is given by

$$\hat{\rho} = \frac{1}{\sqrt{\ell_c \ell_h}} = \left(\frac{\beta_n^2}{\alpha_{cc} \alpha_{hh} q_h} \right)^{1/4}, \tag{7.14}$$

$$H_c(\hat{\rho}) = \sqrt{\alpha_{hh} q_h / \alpha_{cc}}. \tag{7.15}$$

Hence, the peak moves to lower frequencies as the coupling strengths in either network increase. Increasing the light intensity causes q_h to increase, and also makes the peak move to lower frequencies. The height of the peak increases as the coupling strength in the horizontal-cell layer increases; increasing the intensity also makes the height increase. In contrast, the peak height decreases as the cone-layer coupling increases.

The local–gain-control mechanism makes the response to low and high frequencies independent of intensity, but it is not effective at frequencies close to the peak frequency. In particular, the sensitivity to these intermediate frequencies increases with intensity. This behavior is reflected in the impulse response of the system, where we observe a corresponding increase in the strength of the central part of the receptive field. This effect is evident when we compute the inverse Fourier transform to obtain the unit impulse response. For the one-dimensional case, we can obtain a closed-form analytical solution:

$$i_c(x) = \frac{\beta_n^2}{\alpha_{cc}} \frac{\ell}{2\sqrt{2}} e^{-|x|/\ell} \sin\left(\frac{|x|}{\ell} - \frac{\pi}{4} \right). \tag{7.16}$$

7.3 Test Results

The impulse responses of a one-dimensional outer-retina CMOS circuit are shown in Figure 7.4. These data were measured from a VLSI chip fabricated in a 2μm p-well CMOS process. As the cone coupling increases, the gain decreases and the excitatory and inhibitory subregions of the receptive field become larger, precisely as predicted by the theoretical expressions obtained in Section 3.3.5. Increasing the horizontal-cell coupling also enlarges the receptive field, but in this case the gain increases; again, both effects are consistent with the theory. The gain increases because stronger

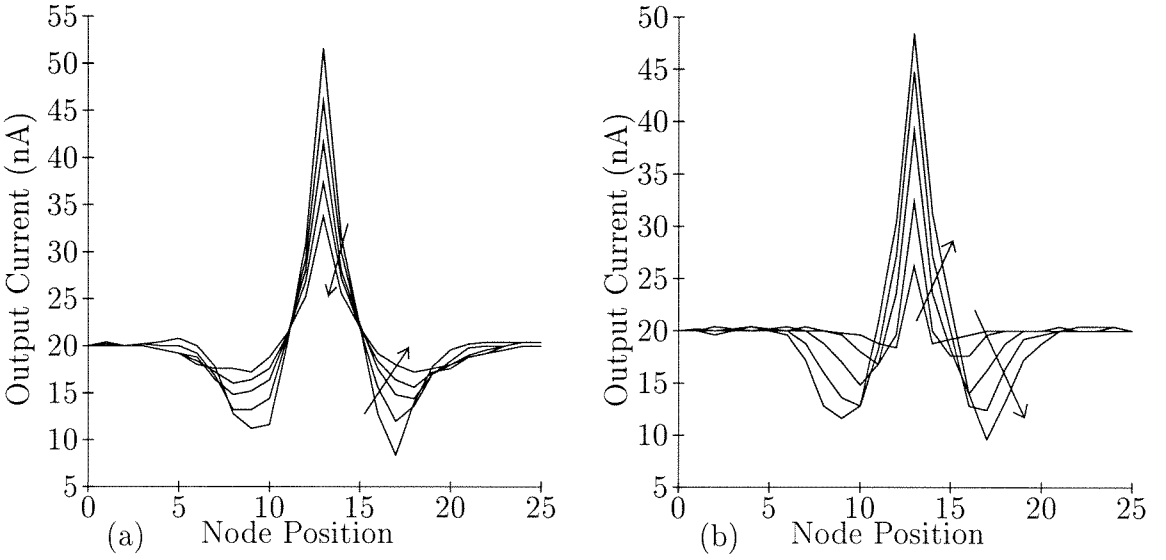


Figure 7.4: Impulse Responses of Outer-Retina CMOS Circuit
The receptive fields are measured for a 25 × 1 pixel chip; arrows indicate increasing diffusor gate voltages. The inputs were 50nA at the center and 10nA elsewhere, and the unit output current I_U was set to 20nA. (a) Increasing interreceptor diffusor voltages in 15mV steps. (b) Increasing inter–horizontal-cell diffusor voltages in 50mV steps.

diffusion results in weaker signals locally, so the inhibition decreases. Figure 7.5a shows the variation of receptive field size with intensity—roughly doubling in size for each decade. This rate indicates a one-third power dependence, which comes close to the theoretical prediction of one-fourth from the linear model. The discrepancy is due to the body effect on transistor M_2 (see Figure 9.4): that effect makes the diffusor strength increase with a power of $1/\kappa^2$.

Contrast-sensitivity measurements are shown in Figure 7.5b. The S–shaped curves are plots of the Michaelis–Menten equation used by physiologists to fit responses of cones [27]:

$$V = V_{\max} \frac{I^n}{I^n + \sigma^n}, \tag{7.17}$$

where σ is the background intensity, and the exponent n determines the slope of the S curve; I included a vertical offset to account for the dependence of transistor mismatch on the intensity level. The circuit deviates at high intensities due to increasing interreceptor coupling strength. For these fits, n is 1.2 in both cases, compared to

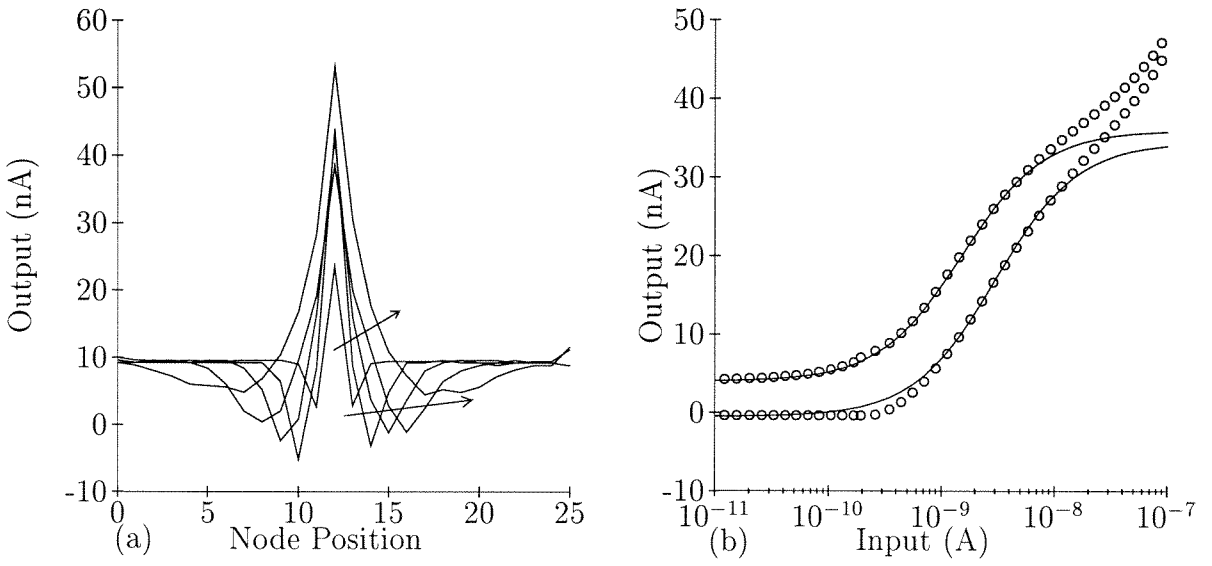


Figure 7.5: Intensity Dependence of Outer-Retina CMOS Circuit
(a) Dependence of receptive field on intensity; arrows indicate increasing intensity. Center inputs were 500pA, 5nA, 15nA, 50nA, and 500nA. The background input was always one-fifth of the center input. (b) Contrast-sensitivity measurements at two background-intensity levels. Lines are fits of the Michaelis–Menten equation given in the text.

the physiologically observed value of 1.0 for cones [27], and σ is 1.5nA and 3.0nA; the actual background intensities correspond to photocurrent levels of 0.56nA and 1.8nA. Thus, the responses are centered at a higher intensity and did not shift horizontally as much as expected with intensity. These discrepancies are due to the difference in gain for inputs above and below the background level. As the inputs decrease the cone coupling reduces, and so the gain increases. Hence, there is a smaller range of operation below the background level.

7.4 Tradeoffs in Outer-Retina Design

Certain filtering characteristics promote efficient encoding in the retina, as I discussed in Sections 3.1 and 3.2.3. Two of the most desirable characterisitcs are

1. A large attenuation of low spatial and temporal frequencies

2. A gain that is inversely proportional to intensity

The parameters that determine these characteristics also determine the temporal stability and the frequency tuning of the outer retina. Therefore, we must make tradeoffs to achieve the desired behavior.

7.4.1 Low-Frequency Attenuation Versus Temporal Stability

We must trade off good temporal stability for low-frequency attenuation. We need a high-gain cone–to–horizontal-cell synapse (i.e., small ϵ_h) to attenuate the cone's response to low spatial and temporal frequencies, since $\tilde{H}_c(0,0) = \epsilon_h / g_{ch}$. However, increasing the gain of the cone–to–horizontal-cell synapse means reducing ϵ_h, and that makes the circuit ring. It rings because the damping factor is small when the Q is large, and the Q is given by $Q_t = (\epsilon_c \sqrt{\frac{\tau_h}{\tau_c}} + \epsilon_h \sqrt{\frac{\tau_c}{\tau_h}})^{-1}$.

To restore temporal stability, we can compensate for the reduction in the second term of the expression for Q_t by increasing the first term, $\epsilon_c \sqrt{(\tau_h / \tau_c)}$. To increase it, we can make ϵ_c larger, decreasing the gain of the horizontal-cell–to–cone feedback synapse.

We can guarantee temporal stability, even when ϵ_h is set close to zero to attenuate low frequencies, by setting $\epsilon_c \sqrt{(\tau_h / \tau_c)} = 1$. This condition gives $\tau_c = \epsilon_c^2 \tau_h$, or

$$\frac{c_{h0}}{c_{c0}} = \frac{g_{ch} g_{hc}}{g_{c0}^2} = \frac{g_{h0}}{g_{c0}} A_{\text{loop}}, \tag{7.18}$$

where $A_{\text{loop}} \equiv (\epsilon_c \epsilon_h)^{-1}$ is the loop gain. Therefore, the horizontal-cell's membrane capacitance must grow like the loop gain. Since it is impossible to make the horizontal-cell's membrane capacitance arbitrarily large, maintaining temporal stability requires keeping the loop gain in a limited range.

Smith and Sterling realized the constraint imposed by temporal stability on the loop gain; they proposed using feedforward inhibition to second-order cells (bipolar cells) to attenuate the DC response [115]. However, it would be better if these low frequencies were removed at the cone, because that solution would achieve the most efficient use of the channel capacity of the cone–to–bipolar-cell synapse.

7.4.2 Gain Control Versus Frequency-Tuning Invariance

We must trade off frequency tuning invariance for local gain control. Equation 7.16 indicates that the most direct way to do gain control is to make the intercone gap-junction conductance ($1/r_{cc}$, or α_{cc}) proportional to intensity—provided that the space constant does not change. However, changing r_{cc} makes the space constant grow as the fourth root of the intensity, since $\ell = (r_{cc} g_{ch} r_{hh} g_{hc})^{-1/4}$. The change in space contant almost entirely cancels the effect of r_{cc} on the gain. Also, the frequency to which the spatial bandpass filter is tuned changes like the fourth root of intensity.

Consequently, we must change another parameter to compensate for the effect of the intercone conductance, $1/r_{cc}$, on the space constant, ℓ. One possibility is to increase g_{ch} proportionally, as we decrease r_{cc} with increasing intensity. Increasing g_{ch} increases the gain from the horizontal cell to the cone, which has the desirable side effect of increasing the attenuation of low-frequency energy ($\tilde{H}_c(0,0) = \epsilon_h / g_{ch}$) as the intensity goes up. Increasing g_{hc} achieves an identical result, since it compensates for the effect of r_{cc} on ℓ and reduces $\epsilon_h \equiv g_{ho}/g_{hc}$, and that reduction increases low-frequency attenuation as well.

Increasing either g_{ch} or g_{hc} will increase the loop gain, and the gain may become arbitrarily large as the intensity goes up. To avoid compromising temporal stability, we must make $g_{c0} \propto \sqrt{I_{\text{photo}}}$. Thus, performing gain control without disturbing frequency tuning (ℓ constant), or temporal stability (Equation 7.18), requires adjusting at least three parameters in coordination.

7.5 Discussion

By using transistors to model gap junctions and chemical synapses, we were able to morph the outer plexiform layer of the retina into a silicon circuit. This circuit performed spatiotemporal bandpass filtering as well as local gain control. We had to make performance tradeoffs to get these two functions to coexist within the same structure. In particular, when we tried to attenuate low-frequency temporal and

spatial signals, we found that the high loop gain required to achieve this goal in a negative-feedback circuit resulted in temporal instability. When we then tried to control the gain by modulating the intercone coupling conductance in proportion to the local intensity, we observed that the receptive field expanded alarmingly.

These severe shortcomings of the simple circuit model of the outer retina that I built forced me to review the retina literature in search of mechanisms that decouple spatiotemporal filtering and local gain control. I found that autofeedback in horizontal cells could provide an elegant solution to this dilemma.

7.5.1 Horizontal-Cell Autofeedback and Temporal Stability

Feedback of horizontal-cell signals back on to the horizontal cells was demonstrated by Kamermans a just few years ago in the tiger salamander [134]. Horizontal cells, which are known to use the inhibitory neurotransmitter GABA, also express GABA-gated Cl-channels. These channels have a reversal potential of -20mV and therefore depolarize the cell when they are opened, forming a positive-feedback loop. Kamermans and Werblin showed that this autofeedback loop could account for the extremely slow dynamics of horizontal cells, increasing the time constant from 65ms to 500ms. My analysis of the tradeoffs involved in outer-retina design has yielded further insights into the role of autofeedback.

As we have seen, there are tradeoffs between small low-frequency response and good temporal stability. Linear-system theory predicts that a high-gain cone–to–horizontal-cell synapse is required to attenuate the cone's response to low spatial and temporal frequencies using feedback inhibition. However, increasing the gain of cone–to–horizontal-cell synapse makes the circuit ring. To maintain stability, we must decrease the gain of the horizontal-cell–to–cone feedback synapse and reduce the horizontal cell's time constant. Unfortunately, both these changes reduce the sensitivity of the cone.

Smith recognized this tradeoff and proposed using feedforward inhibition to attenuate the low-frequency signals at the bipolar cells [115]. However, highly redundant

low-frequency signals make poor use of the limited dynamic range of the cone, and of the limited information-carrying capacity of the cone–to–bipolar-cell synapse. I propose that we eliminate this tradeoff by using slow autofeedback in the horizontal cells. With autofeedback, decreasing the gain of the cone–to–horizontal-cell synapse will not reduce the attenuation of low-frequency signals. Hence, we can achieve temporal stability, extend the dynamic range of the cone, increase the cone's sensitivity, *and* attenuate the low-frequency signals at the cones, making more efficient use of the available dynamic range.

7.5.2 Horizontal-Cell Autofeedback and Receptive-Field Invariance

We have also seen that there is a tradeoff between intensity normalization and invariant receptive field size. Linear-system theory predicts that the gain of the cone is equal to the space constant divided by the intercone coupling conductance. Hence, we can normalize the response by making the intercone conductance proportional to intensity. The intercone coupling may change automatically in the biological retina because gap-junction conductance is the product of permeability and ionic concentration. Unfortunately, the space constant also depends on the gap-junction conductances: $\ell = (r_{cc} g_{ch} r_{hh} g_{hc})^{-1/4}$, where r_{cc} and r_{hh} are the intercone and inter–horizontal-cell resistances and g_{ch} and g_{hc} are the horizontal-cell–to–cone and cone–to–horizontal-cell transconductances, respectively. Thus, as we decrease r_{cc} to reduce the gain, the receptive field expands because ℓ increases.

Must we forego intensity normalization for constant receptive field size? Or vice versa? Autofeedback in the horizontal cells provides an elegant solution to this dilemma: We make the horizontal-cell activity proportional to intensity, and multiply the cone signal by the local horizontal-cell signal to obtain the input to the horizontal cell! The effective cone–to–horizontal-cell transconductance g_{hc} is now proportional to intensity as well, and it cancels the effect of the intensity-dependent intercone resistance r_{cc} on the space constant ℓ. A second-messenger system, or a

metabotropic receptor, may monitor the GABA-mediated horizontal-cell autofeedback pathway and modulate the glutamate-mediated feedforward pathway from the cone to the horizontal cell appropriately.

Chapter 8 Adaptive Quantization: Circuit Models of Spiking Neurons

In this chapter, I discuss the properties of a compact, adaptive, spiking neuron circuit, and I analyze its behavior. We will use this circuit to model retinal ganglion cells. Retinal ganglion cells convert the graded signals that they receive into all-or-nothing spikes. They transmit trains of spikes down their axons which run from the retina to the rest of the brain.

With the exception of ganglion cells, and of a few amacrine cells with extremely long processes, all the cells in the retina encode information using graded, continuously varying signals. Thus, retinal signal processing occurs in the analog domain; quantization—in time and in amplitude—is performed only for the purpose of transmitting information over long distances.

Analog filters and amplifiers with good signal-to-noise ratios can be built with relatively little hardware. However, analog signals are poor at carrying information, because they trade bandwidth for signal-to-noise ratio. Since information capacity increases linearly with bandwidth, but only logarithmically with signal-to-noise ratio, low signal-to-noise, wide-band signals are much more effective at carrying information.

Consequently, a mixed-mode approach gives us the best of both worlds: We encode information using narrow-band, high signal-to-noise signals for processing. And, we encode information using wide-band, low signal-to-noise signals for transmission. Of all the wide-band, low signal-to-noise signals we could use, we choose extremely brief pulses. This choice allows us to conserve power, since we know how to design circuits that dissipate power only for the duration of the pulse, and do not dissipate any power in the quiescent state.

8.1 Information Encoding in Spiking Neurons

We must adopt a strategy for encoding information in pulse trains, just as engineer's do for the analog-to-digital converters they use. The simplest possible encoding is **rate coding**: The frequency of pulses is proportional to the input intensity. But neurons do not simply integrate their input current and fire at a rate that is linearly proportional to the input current level. They encode information differently in two important respects.

First, neurons show **spike-frequency adaptation:** They give priority to changes in the input signal, responding to increases with a high-frequency burst. Their firing rate falls to a much lower level after they adapt to the new level of the stimulus. Spike-frequency adaptation is due to accumulation of calcium in the cell, and to calcium-dependent potassium channels that subtract out the short-term temporal average of the input activity. I have reproduced this behavior by using a simple two-transistor current-mode integrator to model the accumulation of calcium in the cell and the calcium-dependent potassium channel.

Second, neurons show **membrane–time-constant adaptation:** They are exquisitely sensitive to small changes in their inputs and can respond by generating a spike with extremely short latency. I propose a biologically plausible mechanism for membrane time-constant adaptation that shortens the time constant of membrane when the voltage is far below threshold by activating fast, low-threshold, voltage-dependent potassium channels. Hence, the membrane voltage returns rapidly to the threshold and then sits just below the threshold, homing in slowly. I modeled this mechanism by placing a capacitor between the membrane-voltage node and the calcium-integration node that controls the potassium channel.

These two properties make the temporal resolution of neurons much better than that of a simple **integrate-and-fire neuron** in two important respects.

First, the interspike interval of the integrate-and-fire model is proportional to the average level of the input during that interval. This model cannot capture changes that occur on a shorter time scale than the interspike interval. In contrast, a neuron's

interspike interval is proportional to the temporal derivative of the signal as a result of spike-frequency adaptation. Therefore, the more rapidly the input changes, the shorter the interspike interval becomes. This is exactly what an adaptive system ought to do.

Second, the latency of the integrate-and-fire model depends on how long ago the neuron fired, because, in general, the voltage ramps up linearly from the reset level to the threshold. In contrast, a neuron spends most of its time extremely close to threshold as a result of membrane–time-constant adaptation. Therefore, its latency is largely independent of its state. Thus, neurons can signal the timing of small brief perturbations precisely, and all the neurons that respond to this event will fire in synchrony. Paradoxically, the conditions that make precise spike timing possible also make the neuron highly susceptible to noise: As a result, the neuron's interspike intervals are highly stochastic in the absence of any input.

Real neurons show adaptation at all stages of transmission: spike frequency, transmitter release, channel gating, dendritic time-constant, and soma time constant. Our simple neuron models only spike frequency adaptation and membrane–time-constant adaptation at the soma.

8.2 Concept and Circuit

A simplified block diagram of the **adaptive neuron circuit** is shown in Figure 8.1. The spiking mechanism generates a spike whenever the membrane potential exceeds the threshold. Each time a spike occurs, a small fixed amount of charge is added to the leaky integrator on the right that models the intracellular calcium concentration. We use the integrator's output to adjust the conductance of the calcium-dependent potassium channel, $g_{\mathrm{K}}(\mathrm{Ca})$, accordingly.

The other loop adapts the membrane time constant. Rapid changes in membrane potential are amplified by the differentiator and therefore cause a corresponding change in the potassium-channel conductance. This change in conductance acts to restore the membrane voltage, and the effective membrane conductance goes like

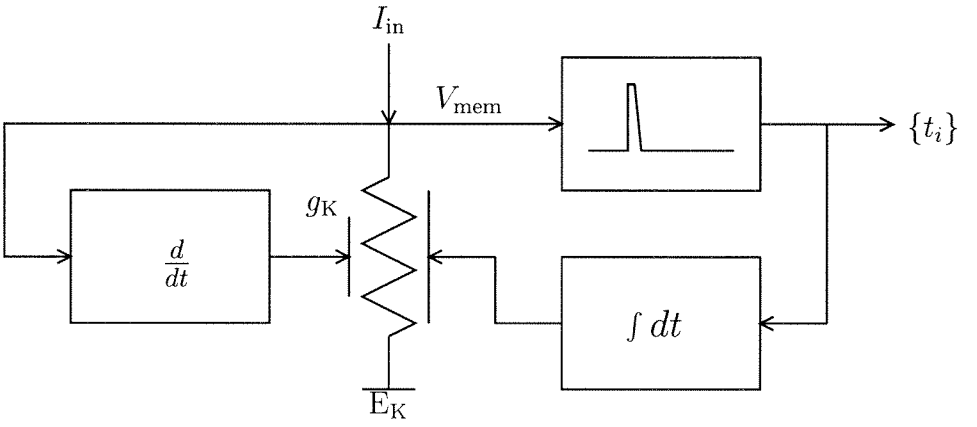


Figure 8.1: Block Diagram of Adaptive Neuron Circuit
The circuit consist of a pulse generator that models the regenerative and restorative parts of the spiking mechanism; a variable conductance, that models the potassium channel population; a leaky integrator, that models the accumulation and buffering of calcium in the cell; and a differentiator, that models the fast voltage-dependence of the potassium channels. The feedback loop on the right adapts the firing rate; the feedback loop on the left adapts the membrane time constant.

the gain times the actual conductance. Since the gain increases with temporal frequency, the time constant is shortened drastically when the membrane is hyperpolarized rapidly. Consequently, the membrane repolarizes rapidly, returning close to the threshold voltage.

In my circuit, I model the potassium channels as a single homogeneous population with both voltage and calcium dependence. The intracellular calcium concentration changes on a slow time scale and sets the baseline for the number of channels that are open. The voltage dependence acts on a much faster time scale and modulates the number of open channels around this baseline. Steady state is reached when the growth rate of the membrane voltage just balances the decay rate of the leaky integrator's charge. Since the leaky integrator's time constant is relatively long, the membrane voltage homes in on the threshold slowly. Thus, these two adaptation mechanisms keep the membrane voltage just below the threshold by regulating the potassium conductance to match the input current level.

The actual circuit is shown in Figure 8.2. The behavior of this circuit is described by two simultaneous differential equations, which I derived by applying Kirchoff's cur-

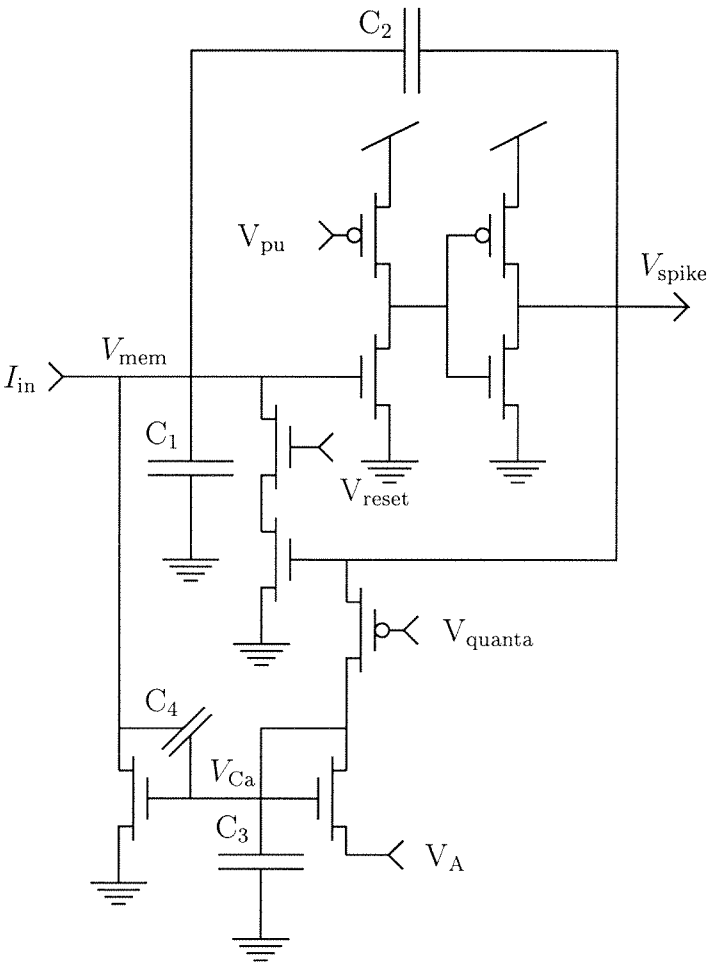


Figure 8.2: ADAPTIVE NEURON CIRCUIT

The high-gain noninverting amplifier, implemented by the two digital invertors, and the positive feedback pathway, implemented by the capacitive divider C_2/C_1, generate the spike. The switched current-sink tied to the input node, implemented by the two transistors connected in series, terminates the spike. The capacitor C_3 accumulates charge and the diode-connected transistor leaks charge away. We meter charge onto C_3 each time a spike occurs by turning on a switched current-source, implemented by the p-type device whose gate is tied to V_{quanta}. The remaining transistor, which turns on slowly as charge accumulates on C_3, and turns on rapidly when V_{mem} increases, due to capacitive coupling through C_4, shunts the input current.

rent law to the membrane-voltage node, labeled V_{mem}, and to the calcium-integration node, labeled V_{Ca}:

$$\begin{aligned} \mathrm{C_{mem}}\frac{dV_{\mathrm{mem}}}{dt} &= I_{\mathrm{in}} - \left(1 + \frac{\mathrm{r_{mc}}}{\mathrm{A}}\right) I_{\mathrm{K}} + (\mathrm{r_{mc}q_{\alpha}} - \mathrm{Q_{th}})\delta(V_{\mathrm{mem}} - \mathrm{V_{th}}), \\ \mathrm{C_{Ca}}\frac{dV_{\mathrm{Ca}}}{dt} &= \mathrm{r_{cm}}(I_{\mathrm{in}} - I_{\mathrm{K}}) - \frac{1}{\mathrm{A}} I_{\mathrm{K}} + (\mathrm{q_{\alpha}} - \mathrm{r_{cm}Q_{th}})\delta(V_{\mathrm{mem}} - \mathrm{V_{th}}), \end{aligned} \tag{8.1}$$

where $\mathrm{C_{mem}}$ and $\mathrm{C_{Ca}}$ are the equivalent node capacitances (i.e., $\mathrm{C_{mem}} \equiv \mathrm{C_1} + \mathrm{C_2} + \mathrm{C_3C_4}/(\mathrm{C_3} + \mathrm{C_4})$, $\mathrm{C_{Ca}} \equiv \mathrm{C_3} + \mathrm{C_4}(\mathrm{C_1} + \mathrm{C_2})/(\mathrm{C_4} + \mathrm{C_1} + \mathrm{C_2})$), and $\mathrm{r_{mc}}$ and $\mathrm{r_{cm}}$ give the fraction of the charge dumped on node V_{Ca} that goes to node V_{mem}, and vice versa (i.e., $\mathrm{r_{mc}} \equiv \mathrm{C_4}/(\mathrm{C_3} + \mathrm{C_4})$ and $\mathrm{r_{cm}} \equiv \mathrm{C_4}/(\mathrm{C_1} + \mathrm{C_2} + \mathrm{C_4})$); $\delta(\cdot)$ is the unit impulse. Here, $\mathrm{Q_{th}}$ is the repolarization charge subtracted from the input node and $\mathrm{q_{\alpha}}$ is the charge increment added to the leaky integrator each time a spike occurs. In this circuit, $\mathrm{Q_{th}}$ is equal to $\mathrm{C_2V_{dd}}$ and $\mathrm{q_{\alpha}}$ is determined by the gate bias voltage V_{quanta} and by the reset voltage $\mathrm{V_{res}}$, which sets the pulse width. The current-mirror gain A relates the current in the diode-connected transistor to the current in the mirror's output transistor, which represents the potassium-channel current I_{K}; the gain increases exponentially with the source bias voltage V_{A}.

The dependence of the potassium-channel current I_{K} on the calcium-node voltage V_{Ca} is given by $I_{\mathrm{K}} = \mathrm{I_0}\exp(\kappa V_{\mathrm{Ca}}/\mathrm{U_T})$, if the device operates in the subthreshold regime, where $\mathrm{U_T} \equiv \mathrm{kT/q}$ is the thermal voltage, and κ is the subthreshold slope coefficient. We can use this relationship to obtain a differential equation for I_{K} by eliminating V_{Ca} from Equation 8.1. Thus, we obtain the following system of equations:

$$\mathrm{C_{mem}}\frac{dV_{\mathrm{mem}}}{dt} = I_{\mathrm{in}} - \left(1 + \frac{\mathrm{r_{mc}}}{\mathrm{A}}\right) I_{\mathrm{K}} + (\mathrm{r_{mc}q_{\alpha}} - \mathrm{Q_{th}})\delta(V_{\mathrm{mem}} - \mathrm{V_{th}}), \tag{8.2}$$

$$\frac{\mathrm{C_{Ca}U_T}}{\kappa I_{\mathrm{K}}}\frac{dI_{\mathrm{K}}}{dt} = \mathrm{r_{cm}}(I_{\mathrm{in}} - I_{\mathrm{K}}) - \frac{1}{\mathrm{A}} I_{\mathrm{K}} + (\mathrm{q_{\alpha}} - \mathrm{r_{cm}Q_{th}})\delta(V_{\mathrm{mem}} - \mathrm{V_{th}}). \tag{8.3}$$

Now it is evident that the time scale for the potassium current is inversely proportional to the latter's amplitude: It changes faster when the current is larger. This scaling is a direct result of the exponential current–voltage relationship; I say more about this diode-capacitor–style dynamics in Section 8.3. In relation to biophysics, this variable

time scale is analogous to a rate constant for calcium buffering that is proportional to the intracellular calcium concentration.

The first equation captures the dependence of the membrane voltage on the input current and on the potassium-channel current. The fast transient currents responsible for generating and terminating the spike have been omitted intentionally. These currents are analogous to the sodium current and the delayed-rectifier potassium current in the Hodgkin–Huxley model. I ignore these currents because I am not interested in reproducing the detailed profile and shape of the spike. My goal is to reproduce the dependence of the interspike interval on the input current. Therefore, I am modeling the slow currents that shape the membrane-voltage trajectory during the interspike interval, such as the calcium-dependent potassium current. I have lumped the effect of the fast currents into a net repolarization charge, Q_{th}, that is subtracted from the membrane capacitance after each spike, returning the membrane voltage to the reset level.

The second equation captures the dependence of the potassium current on the spiking activity of the cell and on the trajectory of the membrane voltage. The q_α term models calcium entering the cell through the sodium channels that rapidly depolarize the membrane during a spike; the $r_{cm}Q_{th}$ term models calcium leaving the cell through the potassium channels that quickly repolarize the membrane to terminate the spike. The $r_{cm}(I_{in} - I_K)$ term models the dependence of the potassium current on the derivative of the membrane voltage. This term makes the rate of change of the potassium current proportional to the rate of change of the membrane voltage, since, from Equation 8.2, $I_{in} - I_K$ is the net current available to charge the membrane capacitance during the interspike interval, assuming that $r_{mc}/A \ll 1$. There is also a leakage term I_K/A that models buffering of calcium within the cell.

Note that, as implemented in this circuit, the potassium channel does not inactivate. That is, it turns on when the membrane voltage rises rapidly, and it *stays* on at a steady level of depolarization—until calcium buffers reduce the intracellular-calcium concentration. Hence, the slow calcium dependence is the analog of the inactivation variable h used for the fast sodium conductance in the Hodgkin–Huxley model. Of

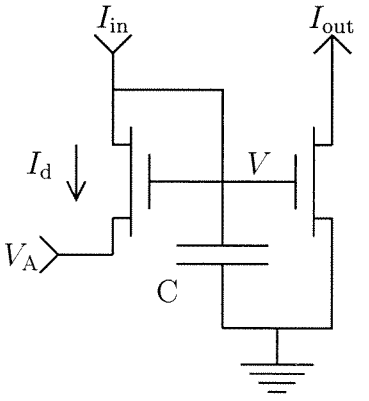


Figure 8.3: CIRCUIT DIAGRAM OF DIODE-CAPACITOR INTEGRATOR
The time constant is set by the current in the diode-connected input device I_d. We can set the output-current level independently by applying an appropriate bias voltage, V_A, to this device's source terminal. This voltage controls the gain of the current mirror.

course, the potassium current in my circuit also can be turned off by rapid hyperpolarizations through the reduction of activation. A noninactivating channel is crucial for membrane–time-constant adaptation because a steady current is required to match the input current and to hold the membrane voltage just below the threshold.

In Section 8.3, I analyze the diode-capacitor integrator circuit used to model the intracellular calcium concentration. In Section 8.4, I describe the axon-hillock circuit used to generate spikes. In Section 8.5, I discuss how to quantify the timing precision of neurons and I give definitions for latency and synchronicity. I analyze the complete adaptive neuron circuit in Section 8.6 and present measurements of its timing precision in Section 8.7. I close the chapter with a short discussion in Section 8.8.

8.3 Leaky Integration with a Capacitor and a Diode

The **diode-capacitor integrator** is shown in Figure 8.3; it is based on the well-known current-mirror circuit. The large capacitor at the input node accumulates charge, and the diode-connected transistor leaks charge away. For subthreshold current levels, the current has an exponential dependence on the gate voltage, and there-

fore the small-signal conductance of the diode-connected transistor is proportional to the current. Hence, the time constant will change as the current level changes. This dependence makes the circuit nonlinear, so we cannot obtain a solution for any arbitrary input waveform. However, the circuit's dynamic behavior is described by a simple differential equation

$$\mathrm{C}\frac{dV}{dt} = I_{\mathrm{in}}(t) - I_{\mathrm{d}} = I_{\mathrm{in}}(t) - \tfrac{1}{\mathrm{A}} I_{\mathrm{out}}(t),$$

where $\mathrm{A} = \exp(V_{\mathrm{A}}/\mathrm{U_T})$. Since the current passed by a MOS transistor is related to the gate and source voltages by $I_{\mathrm{ds}} = \exp((\kappa V_{\mathrm{g}} - V_{\mathrm{s}})/\mathrm{U_T})$ for subthreshold operation, we can eliminate the voltage V and rewrite this equation solely in terms of the input and output currents:

$$\mathrm{Q_T}\frac{dI_{\mathrm{out}}}{dt} = I_{\mathrm{out}}(t)(I_{\mathrm{in}}(t) - \tfrac{1}{\mathrm{A}} I_{\mathrm{out}}(t)), \tag{8.4}$$

where $\mathrm{Q_T} \equiv \mathrm{CU_T}/\kappa$ is the amount of charge required to e-fold the current. We can gain insight into the circuit's behavior by rewriting this differential equation in the following form:

$$\frac{\mathrm{Q_T}}{I_{\mathrm{in}}(t)}\frac{d(1/I_{\mathrm{out}})}{dt} = \frac{1}{\mathrm{A}I_{\mathrm{in}}(t)} - \frac{1}{I_{\mathrm{out}}(t)}.$$

This equation is a simple first-order ordinary differential equation in $1/I_{\mathrm{out}}$ with time constant $\mathrm{Q_T}/I_{\mathrm{in}}$—which is fixed only if $I_{\mathrm{in}}(t)$ is constant. Hence, for a steady nonzero input current, $1/I_{\mathrm{out}}$ changes exponentially with time constant $\tau = \mathrm{Q_T}/I_{\mathrm{in}}$. If the input current is 0, $1/I_{\mathrm{out}}$ decays linearly at the rate $1/(\mathrm{AQ_T})$.

8.3.1 A General Solution

In the most general case, we may write the solution in the form of the integral equation:

$$I_{\mathrm{out}}(t) = \frac{I_{\mathrm{out}}(t_0)}{\frac{I_{\mathrm{out}}(t_0)}{\mathrm{AQ_T}}\int_{t_0}^{t} \exp\left(-\frac{1}{\mathrm{Q_T}}\int_u^t I_{\mathrm{in}}(s)ds\right)du + \exp\left(-\frac{1}{\mathrm{Q_T}}\int_{t_0}^t I_{\mathrm{in}}(s)ds\right)}. \tag{8.5}$$

If we divide both the numerator and the denominator by the second term in the denominator, and express the resulting numerator as an integral of its derivative, we can rewrite this result in the following form:

$$I_{\text{out}}(t) = \frac{\frac{1}{\text{Q}_\text{T}}\int_{t_0}^{t} I_{\text{in}}(u)\exp\left(\frac{1}{\text{Q}_\text{T}}\int_{t_0}^{u} I_{\text{in}}(s)ds\right)du + 1}{\frac{I_{\text{out}}(t_0)}{\text{AQ}_\text{T}}\int_{t_0}^{t}\exp\left(\frac{1}{\text{Q}_\text{T}}\int_{t_0}^{u} I_{\text{in}}(s)ds\right)du + 1} I_{\text{out}}(t_0).$$

Now it is evident that the circuit computes a normalized weighted average by assigning weights to past inputs that decay exponentially with a time constant $\tau = \text{Q}_\text{T}/\langle I_{\text{in}}\rangle_{0}^{t}$, for $I_{\text{in}} \neq 0$, where $\langle\cdot\rangle_{t_1}^{t_2}$ is the mean value in the interval from t_1 to t_2. This time constant is equal to the time that it takes to change the output current by a factor of e when all the input current goes to charge the capacitor C. We need the constant offsets in the numerator and in the denominator to satisfy the boundary condition at $t = 0$; they become negligible for times

$$t \gg \tau \ln\left(\frac{\text{A}\langle I_{\text{in}}(t)\rangle_{0}^{t}}{I_{\text{out}}(0)} + 1\right).$$

Naturally, the time over which past inputs are forgotten depends on how different they are from the current state ($I_{\text{out}}(t) \simeq \text{A}I_{\text{in}}(t)$) and on the value of the time constant τ.

8.3.2 Response to Step Changes

For a step change in the input to I_1 at $t = 0$, we can obtain closed-form expressions for the integral equation (Equation 8.5):

$$I_{\text{out}}(t) = \frac{\text{A}}{(1/\text{I}_1) + (1/\text{I}_0 - 1/\text{I}_1)\exp(-t/\tau)},$$

where $\tau = \text{Q}_\text{T}/\text{I}_1$, $\text{I}_0 = I_{\text{out}}(0)/\text{A}$, and $\text{I}_0 \neq \text{I}_1 \neq 0$. As expected, the reciprocal of the output current, $1/I_{\text{out}}(t)$, follows a single-exponential–decay from $1/\text{I}_0$ to $1/\text{I}_1$, which is a characteristic of first-order linear systems. Consequently, the time course of the output current is described by a Fermi function—or by a tanh if I_1 is greater than I_0,

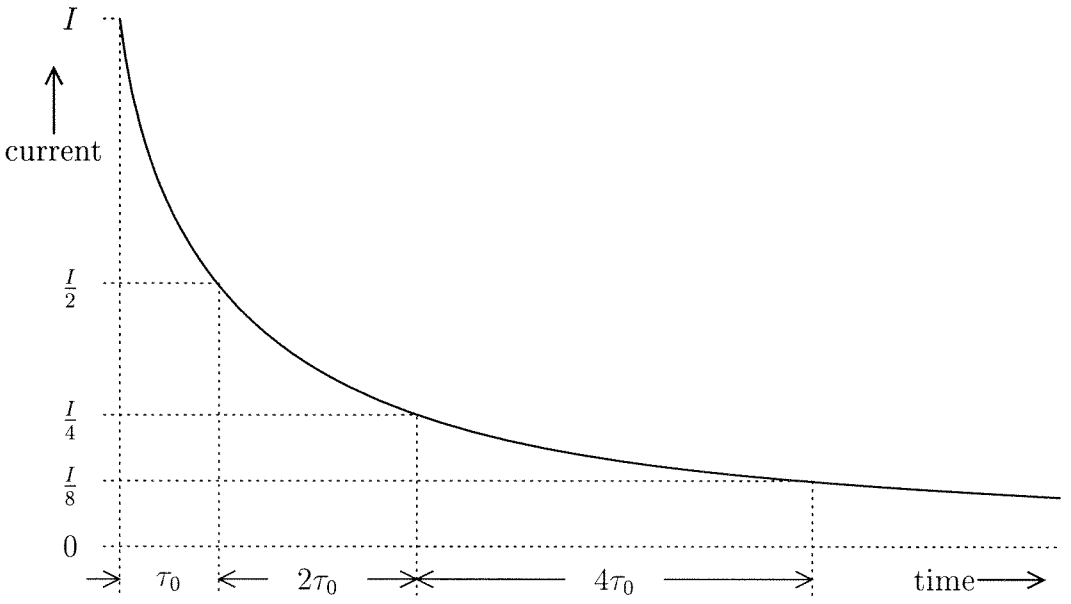


Figure 8.4: UNDRIVEN RESPONSE OF DIODE-CAPACITOR INTEGRATOR
The time that it takes for the current to decay by a certain fraction (50 per cent, in this example) is inversely proportional to the current level.

or by a coth if I_1 is less than I_0.

If $I_1 = 0$, the integral equation reduces to

$$I_{\text{out}}(t) = \frac{I_{\text{out}}(0)}{t/\tau_0 + 1}, \tag{8.6}$$

and the output has a half-life of $\tau_0 \equiv AQ_T/I_{\text{out}}(0)$. The half-life is proportional to the change in $1/I_{\text{out}}(t)$ (i.e., $2/I_{\text{out}}(0) - 1/I_{\text{out}}(0)$), because $1/I_{\text{out}}(t)$ decays at a constant rate of $1/AQ_T$. In general, it takes $(n-1)\tau_0$ for the current to decay by a factor of n, or $n\tau_0$ for it decay from $I_{\text{out}}(0)/n$ to $I_{\text{out}}(0)/(2n)$, as shown in Figure 8.4. This behavior, which is characteristic of the diode-capacitor dynamics, arises because the time scale is inversely proportional to the output current.

8.3.3 Response to Spike Trains

For a sequence of current pulses at the input, the general solution (Equation 8.5) reduces to:

$$I_{\text{out}}(t) = \frac{I_{\text{out}}(t_0^-)}{\frac{I_{\text{out}}(t_0^-)}{\text{AQ}_\text{T}}\left(\sum_{j=1}^{n}\exp(-\frac{1}{\text{Q}_\text{T}}\sum_{i=j}^{n}q_i)(t_j - t_{j-1}) + (t - t_n)\right) + \exp(-\frac{1}{\text{Q}_\text{T}}\sum_{i=0}^{n}q_i)}, \tag{8.7}$$

for $t_n \leq t < t_{n+1}$, where $\{t_i\}$, $t_0 < t_1 < t_2, \ldots$, are the times at which the pulses occur, and $\{q_i\}$ are the amounts of charge that each spike supplies. Assuming that each spike adds the same amount of charge q_α to the input, we obtain

$$I_{\text{out}}(t) = \frac{I_{\text{out}}(t_0^-)}{\frac{I_{\text{out}}(t_0^-)}{\text{AQ}_\text{T}}\left(\sum_{j=1}^{n}(1+\alpha)^{j-(n+1)}(t_j - t_{j-1}) + (t - t_n)\right) + (1+\alpha)^{-(n+1)}},$$

where $1+\alpha \equiv \exp(q_\alpha/\text{Q}_\text{T}) > 1$ is the factor by which the output current is multiplied.

We get a less cluttered expression by breaking this equation into two parts:

$$I_{\text{out}}(t) = \frac{I_{\text{out}}(t_n)}{\frac{I_{\text{out}}(t_n)}{\text{AQ}_\text{T}}(t - t_n) + 1}, \qquad t_n \leq t < t_{n+1}; \tag{8.8}$$

$$I_{\text{out}}(t_n) = \frac{I_{\text{out}}(t_0^-)}{\frac{I_{\text{out}}(t_0^-)}{\text{AQ}_\text{T}}\sum_{j=1}^{n}(1+\alpha)^{j-(n+1)}(t_j - t_{j-1}) + (1+\alpha)^{-(n+1)}}. \tag{8.9}$$

A fixed quantity of charge increments the voltage by a fixed amount and thereby multiplies the current by a fixed factor. Hence, the incremental change in the output current caused by a spike is not fixed: it is proportional to the output-current level at the time that the spike occurs. The result that we just obtained does indeed predict that $I_{\text{out}}(t_n^+) = (1+\alpha)I_{\text{out}}(t_n^-)$, as this argument leads us to expect. These multiplicative increments enable the spike's contribution to the output current to be facilitated by spikes that arrive earlier. This effect has been seen in the excitatory postsynaptic potentials recorded from real neurons, where it goes by the name **paired-pulse**

facilitation. Using Equation 8.8, we can show that

$$\Delta I_{\text{out}}(\Delta t) \equiv I_{\text{out}}(t_1) - I_{\text{out}}(t_1^-) = \left(\frac{I_{\text{out}}(t_1)}{I_{\text{out}}(t_1^-)} - 1\right) I_{\text{out}}(t_1^-) = \frac{\alpha I_{\text{out}}(t_0)}{\frac{I_{\text{out}}(t_0)}{\text{AQ}_\text{T}}\Delta t + 1}$$

where $\Delta t \equiv t_1 - t_0$ is the time difference between the two spikes. Hence, the height of the excitatory postsynaptic current produced by the second pulse reaches a maximum of $\alpha I_{\text{out}}(t_0)$, when $\Delta t = 0$, and decays like $1/\Delta t$ thereafter.

If the interspike intervals are all the same, we can sum the geometric series in the denominator of Equation 8.9 and obtain

$$I_{\text{out}}(t) = \frac{I_{\text{out}}(t_0 + nT)}{\frac{I_{\text{out}}(t_0+nT)}{\text{AQ}_\text{T}}(t - (t_0 + nT)) + 1}; \; t_0 + nT \le t < t_0 + (n+1)T,$$

$$I_{\text{out}}(t_0 + nT) = \frac{1}{1/\hat{I}_T + (1/I_{\text{out}}(t_0) - 1/\hat{I}_T)(1+\alpha)^{-n}}; \qquad n = 1, 2, \ldots, \tag{8.10}$$

$$\hat{I}_T \equiv \alpha\frac{\text{AQ}_\text{T}}{\text{T}}, \tag{8.11}$$

where $T \equiv t_j - t_{j-1}$; $j = 1, 2, \ldots$. Now we can see how the integrator responds to a step increase or decrease in the frequency of input pulses: The peak output current levels attained immediately after each spike converge to $\hat{I}_T$ when $(1+\alpha)^{-n} \ll 1$. Hence, the time taken to reach equilibrium is independent of the change in frequency—it depends on only the number of spikes. As the interspike intervals become shorter, the time scales accordingly. This property enables the circuit to match the time that it takes to do the computation to the rate at which information is supplied.

The equilibrium output-current level is proportional to the input frequency $f = 1/T$: It is $\alpha\text{AQ}_\text{T}f$ at the peaks, it is $(1+\alpha)$ times less at the throughs, and the mean level is $\text{A}q_\alpha f$. I obtained this expression for the mean level by integrating $I_{\text{out}}(t)$ over the interval T. Alternatively, intuition tells us that, at equilibrium, the charge that leaks away during the interspike interval T exactly matches the charge supplied by each spike, q_α. And we know that for every charge q that leaks away through the input diode, A times q flows in the output transistor, since the current-mirror integrator has a gain of A. Hence, the mean output current level must be $\text{A}q_\alpha/T$.

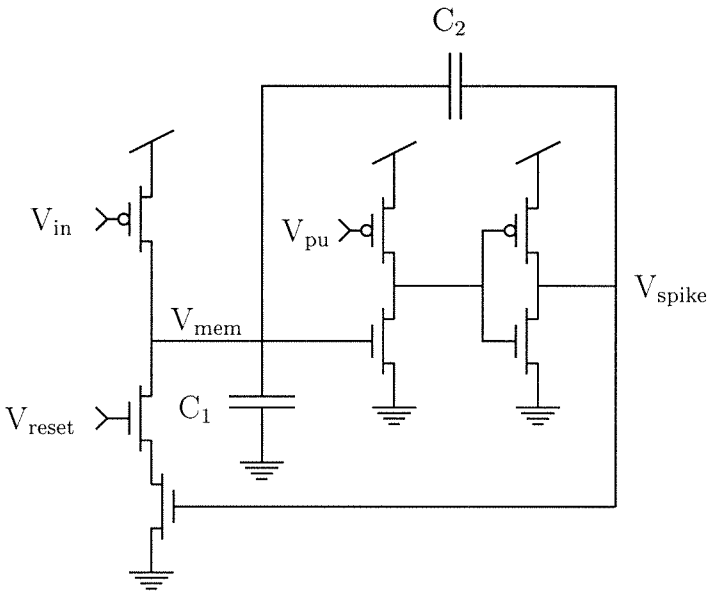


Figure 8.5: Modified Self-Resetting Axon-Hillock Circuit
This pulse-generating circuit is identical to the one described in Mead's monograph [122], except that I have replaced the pullup of the first inverter with a fixed current source to reduce the switching current.

8.4 Axon-Hillock Circuit

The axon-hillock circuit is shown in Figure 8.5. It has a high-gain amplifier with positive feedback around it that models the thresholding and the regenerative action of spike-generating mechanisms at the axon-hillock of a neuron. In addition, there is a reset mechanism that terminates the spike. This circuit is described in detail in Chapter 12 of Mead's monograph [122]. Here, I provide a brief description of the ciruit and reiterate salient points about its behavior that will be handy for analyzing the adaptive neuron circuit.

The two digital invertors provide a noninverting voltage gain of about 100. I have replaced the first CMOS invertor with a pseudo-nMOS invertor with a weak pullup, to reduce the current passed at the switching threshold. This modification reduces power consumption significantly because, as pointed out by Lazzaro [135], the invertor spends most of its time close to threshold—unlike in a conventional digital circuit. The capacitive divider adds about one-tenth of the 5V output-voltage swing back to

the input, providing positive feedback. A switched current sink, implemented by two transistors in series, terminates the spike by discharging the amplifier's input node V_{mem}. One transistor acts as a switch, and the other limits the rate at which the input capacitor is discharged. The switch closes when the output V_{spike} goes high, and we use the bias voltage V_{reset} to adjust the discharge rate and, thereby, set the pulse width.

The axon-hillock circuit makes a fixed-height, fixed-width pulse when the input voltage reaches the switching threshold of the nMOS inverter, which is $\mathrm{V_{dd}} - V_{\mathrm{pu}}$, assuming that the n- and p-type devices are matched. As V_{pu} moves toward $\mathrm{V_{dd}}$, the threshold moves toward ground. However, the threshold should not be set too low, because the delay of the inverter increases as current levels are reduced. A longer delay will increase the pulse width, which is the sum of the repolarization time and the propagation delay of the amplifier. The latter is not negligible if the NMOS inverter's pullup current becomes comparable to the reset current.

Since the input current must charge $\mathrm{C_1}$ and $\mathrm{C_2}$ by $\mathrm{V_{dd}C_2/(C_1 + C_2)}$ to generate a spike, the interpulse interval is given by

$$t_{\mathrm{low}} = \frac{\mathrm{C_2 V_{dd}}}{I_{\mathrm{in}}},$$

as shown in Mead's monograph [122]. The pulse width is given by

$$t_{\mathrm{hi}} = \frac{\mathrm{C_2 V_{dd}}}{I_{\mathrm{reset}} - I_{\mathrm{in}}},$$

assuming that the delay of the inverters is negligible. Hence, the firing frequency is

$$f = \frac{1}{t_{\mathrm{low}} + t_{\mathrm{hi}}} = \frac{1}{\mathrm{C_2} V_{\mathrm{dd}}} \frac{I_{\mathrm{in}}}{1 + I_{\mathrm{in}}/(I_{\mathrm{reset}} - I_{\mathrm{in}})}.$$

It is linearly proportional to I_{in} for $I_{\mathrm{in}} \ll I_{\mathrm{reset}}$. As I_{in} approaches I_{reset}, the pulse width increases, limiting the firing frequency. The circuit fails to reset, and hangs, if I_{in} exceeds I_{reset}. I choose therefore to operate this circuit with I_{reset} over 10 times larger than the maximum expected input current, or $\mathrm{V_{reset}} > 1\mathrm{V}$ for subthreshold

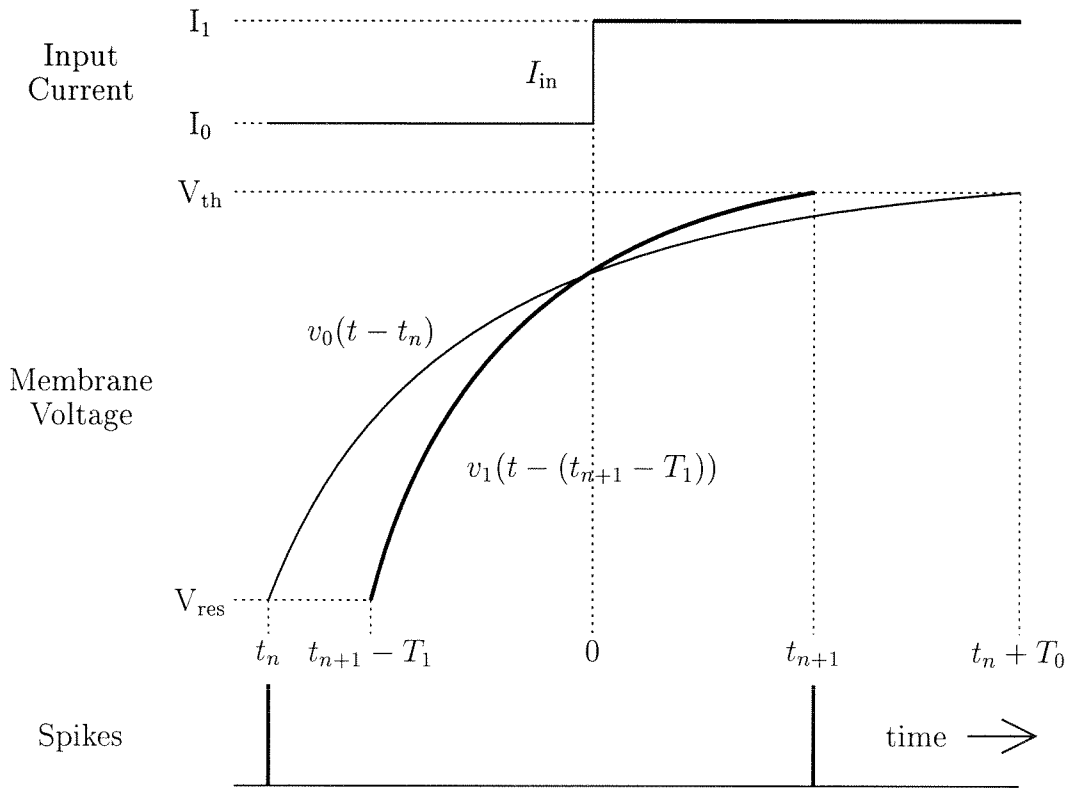


Figure 8.6: MEMBRANE-VOLTAGE TRAJECTORIES
The two curves represent the trajectories for two steady input current levels, I_0 and I_1. The second curve corresponds to the larger current (I_1), and therefore has a steeper slope. These curves are drawn such that they intersect at the time $t = 0$ when the input is stepped from I_0 to I_1. The membrane voltage follows the first curve before the step, and follows the second after the step.

input currents. I set V_{pu} to about 3.5V to obtain a threshold voltage of about 1.5V and a switching current of about 5μA.

8.5 Neuronal Latency and Synchrony

I introduce the techniques that we use to compute the average latency and synchronicity of neuron circuits in this section. We use the axon-hillock circuit to illustrate the procedure and to serve as our benchmark.

We have to determine the time, t_{n+1}, when a neuron fires, after we step its input current from I_0 to I_1 at time $t = 0$, given the previous time that it fired, t_n. The

situation is depicted in Figure 8.6. The trajectory of the membrane voltage during the interspike interval is generally determined by a differential equation of the form

$$\mathrm{C}\frac{dV_{\mathrm{mem}}}{dt} = I_{\mathrm{in}}(t) - I_{\mathrm{K}}(t - t_n) - \mathrm{g}V_{\mathrm{mem}}(t - t_n),$$

where C is the membrane capacitance, g is the membrane conductance, and $I_{\mathrm{K}}(t-t_n)$ is the current shunted across the membrane by active or passive potassium channels. I am ignoring the Na-channels and all the other fast channels because the behavior during the interspike interval is determined largely by the slow processes. Notice that I have set up the origins for $I_{\mathrm{K}}(\Delta)$ and $V_{\mathrm{mem}}(\Delta)$ such that they give the values of the potassium current and the membrane voltage Δ seconds after the last spike occured—rather than at the absolute time t. This equation gives

$$\begin{aligned} \mathrm{C}\int_{t_n}^{t_{n+1}} dV_{\mathrm{mem}} &= \int_{t_n}^{t_{n+1}} (I_{\mathrm{in}}(t) - I_{\mathrm{K}}(t - t_n) - \mathrm{g}V_{\mathrm{mem}}(t - t_n))dt, \\ \Rightarrow \mathrm{Q_{th}} &= \mathrm{I_1}t_{n+1} - \mathrm{I_0}t_n - \int_{t_n}^{t_{n+1}} (I_{\mathrm{K}}(t - t_n) + \mathrm{g}V_{\mathrm{mem}}(t - t_n))dt, \quad (8.12) \end{aligned}$$

since $V_{\mathrm{mem}}(0) = \mathrm{V_{res}}$, $V_{\mathrm{mem}}(t_{n+1} - t_n) = \mathrm{V_{th}}$, and $\mathrm{Q_{th}} \equiv \mathrm{C}(\mathrm{V_{th}} - \mathrm{V_{res}})$. Here, $\mathrm{V_{res}}$ and $\mathrm{V_{th}}$ are the reset voltage and the threshold voltage, respectively, and $\mathrm{Q_{th}}$ is the repolarization charge. This result allows us to compute the neuron's latency, which is simply t_{n+1}, since the step occurs at $t = 0$. The latency is variable because the values of the membrane voltage and the potassium current when the step is applied vary from trial to trial and from neuron to neuron. These state variables are determined by t_n, the time that the last spike occurred, measured relative to the step.

If we know the distribution of spikes times, t_n, either over a set of trials or across the neuronal population, we can compute mean latency. We can also compute the standard deviation, which quantifies the variability in latency due to the dependence on the initial state and provides a measure of the synchrony of the response across the population.

Here, I choose instead to characterize the response by the synchronous firing rate. The **synchronous firing rate** is defined as the instantaneous total number of spikes

per second divided by the number of neurons or the number of trials. In other words, the synchronous firing rate is equal to the probability density function for the distribution of the first spike that occurs after the step. The delay and height of the peak in the probability density function serve as measures of the latency and of the strength of the neuronal responses, respectively. Furthermore, I define the **synchronicity** of the response as the peak synchronous firing rate divided by the neuron's mean firing rate immediately after the step. I believe that the normalized peak synchronous firing rate is the most relevant measure of neuronal response, because it is the most salient quantity from the point of view of the postsynaptic target, which must extract stimulus-triggered activity that is superimposed on random background activity. I denote measures of latency, peak synchronous firing rate, and synchronicity by μ, $\hat{f}$, and λ.

Given the probability density function for spiking, $p_0(t_n)$; $-\mathrm{T}_0 \leq t_n < 0$, for a period of length T_0 preceeding the step, we can compute the probability density function, $p_1(t_{n+1})$; $0 \leq t_{n+1} < \mathrm{T}_1$, for a subsequent period of length T_1 following after the step. Let Δt_n be a time interval that occurs before the step, and let Δt_{n+1} be a time interval that occurs after the step. Let us demarcate the second interval such that any spike that occurs in first interval is followed by a spike that falls in second interval. Then, we must have

$$\Delta t_{n+1} p_1(t_{n+1}) = \Delta t_n p_0(t_n), \tag{8.13}$$

$$\Rightarrow p_1(t_{n+1}) = \frac{dt_n}{dt_{n+1}} p_0(t_n). \tag{8.14}$$

Therefore, the **time-scaling factor**, dt_n/dt_{n+1}, is all that we need to know to calculate the new probability density function. This function tells us exactly how much we should shrink or stretch time intervals as we map them from the period $[0, \mathrm{T}_1]$ before the step to the period $[-\mathrm{T}_0, 0]$ after the step.

We can obtain an expression for the derivative of t_n with respect to t_{n+1} as follows. If we shift forward by dt_n the time, t_n, that the previous spike occurred, we find that the value of the membrane voltage, $V_{\mathrm{mem}}(-t_n)$, at the time that the step occurs,

reduces by

$$DV_{\mathrm{mem}} = \left.\frac{dV_{\mathrm{mem}}}{dt}\right|_{t=0^-} \times dt_n = dt_n(\mathrm{I}_0 - I_{\mathrm{K}}(-t_n) - gV_{\mathrm{mem}}(-t_n))/C.$$

Making this voltage shift is equivalent to sliding the membrane-voltage trajectory after the step to the right by

$$dt_{n+1} = \frac{DV_{\mathrm{mem}}}{\left.\frac{dV_{\mathrm{mem}}}{dt}\right|_{t=0^+}} = DV_{\mathrm{mem}}C/(\mathrm{I}_1 - I_{\mathrm{K}}(-t_n) - gV_{\mathrm{mem}}(-t_n)).$$

Hence,

$$\frac{dt_n}{dt_{n+1}} = \frac{\mathrm{I}_1 - I_{\mathrm{K}}(-t_n) - gV_{\mathrm{mem}}(-t_n)}{\mathrm{I}_0 - I_{\mathrm{K}}(-t_n) - gV_{\mathrm{mem}}(-t_n)}. \tag{8.15}$$

When no time-dependent potassium currents or conductances are present in the membrane, the time-scaling factor is constant; therefore, the shape of the probability density function remains unchanged—its time scale is simply rescaled by the resultant linear relationship between t_{n+1} and t_n. In contrast, time-dependent potassium currents and membrane conductances can produce a variable time-scaling factor and reshape the probability density function. Therefore, by carefully designing the time dependence of the potassium current, we can manipulate the time-scaling factor, so as to reshape the probability density function to reduce the latency and to increase the synchronicity of the response.

For the simple axon-hillock circuit, $I_{\mathrm{K}}(t) = 0$ and $g = 0$, since there is no potassium current and the devices tied to the input are current sources. Hence, Equation 8.12 reduces to

$$t_{n+1} = \mathrm{T}_1 + \frac{\mathrm{T}_1}{\mathrm{T}_0} t_n,$$

where $\mathrm{T}_0 = \mathrm{Q}_{\mathrm{th}}/\mathrm{I}_0$ and $\mathrm{T}_1 = \mathrm{Q}_{\mathrm{th}}/\mathrm{I}_1$ are the interspike intervals for the currents I_0 and I_1 (I am replacing $\mathrm{I}_0/\mathrm{I}_1$ everywhere with $\mathrm{T}_1/\mathrm{T}_0$, for uniformity). Hence, as t_n changes from $-\mathrm{T}_0$ to zero, t_{n+1} changes linearly from zero to T_1, as expected. Now,

using Equations 8.14 and 8.15, the probability density functions are related by

$$p_1(t_{n+1}) = \frac{T_0}{T_1} p_0 \left(\frac{T_0}{T_1} (t_{n+1} - T_1) \right). \tag{8.16}$$

If the spikes are uniformly distributed initially, then $p_0(t_n) = 1/T_0$, and we obtain $p_1(t_{n+1}) = 1/T_1$. Since $p_1(t_{n+1})$ is also uniform, the uniform distribution will persist indefinitely—no matter how many steps occur in the input current. Hence, the spikes occur completely randomly, and the synchronous firing rate simply is equal to the firing frequency of the neuron, $1/T_1$. Since there is no peak in the probability density function, we have to choose a point in the distribution to obtain the latency by some other criteria. The midpoint seems to be the most reasonable choice since it corresponds to the mean delay. So we assign the simple integrate-and-fire neuron a latency of $\mu = T_1/2$, a peak synchronous firing rate of $\hat{f} = 1/T_1$, and a synchronicity of $\lambda = 1$.

These results suggest that we can estimate the onset time and the strength of a particular stimulus by determining the subset of spikes in the neuronal population triggered by that stimulus and computing the peak synchronous firing rate for this ensemble. For the simple axon-hillock circuit, there is no peak in the synchronous firing rate, and we simply use the constant synchronous firing rate, f. The latency is inversely proportional to f, and the input current level is proportional to f. Hence, the time that the stimulus occurred is given by $\langle t_i \rangle - 1/(2f)$, where $\langle t_i \rangle$ is the mean spike arrival time, and the input current supplied by the stimulus is given by $C_2 V_{dd}/f$. We investigate how these measures of latency and synchronicity are affected by introducing calcium-dependent potassium channels in Section 8.6.

8.6 Calcium-Dependent Potassium Channels

To obtain firing-rate adaptation, I connected the diode-capacitor integrator around the axon-hillock circuit, as shown in Figure 8.2; the integrator's output current is subtracted from the input current. The current source tied to the input of the integrator

is turned on when V_{spike} is high. Thus, a miniscule amount of charge, determined by V_{quanta} and by the pulse width, is dumped on the capacitor for each spike, causing a small increment in V_{Ca}. This feedback path models the calcium-dependent potassium channels that produce firing-rate adaptation.

In this section, I analyze the behavior of the neuron with firing-rate adaptation in place. I start by looking at the steady state behavior, and I show that there are two distinct regimes of operation, depending on the parameter values chosen. Then, I obtain a relationship between the interspike interval and the input current for a step input, and derive the probability of firing a spike, given the distribution of times at which the previous spikes occurred. Finally, I use this probability distribution to calculate the latency and synchronicity of the adaptive neuron. Before we proceed with the analysis, we must set $\text{C}_4 = 0$ to eliminate membrane time-constant adaptation. That mechanism is dealt with in a more complete analysis that is outside the scope of this thesis.

With $\text{C}_4 = 0$, the system of differential equations that describes the circuit (Equations 8.2 and 8.3) reduces to

$$\text{C}_{\text{mem}} \frac{dV_{\text{mem}}}{dt} = I_{\text{in}} - I_{\text{K}} - \text{Q}_{\text{th}} \delta(V_{\text{mem}} - \text{V}_{\text{th}}), \tag{8.17}$$

$$\frac{\text{Q}_{\text{T}}}{I_{\text{K}}} \frac{dI_{\text{K}}}{dt} = \text{q}_\alpha \delta(V_{\text{mem}} - \text{V}_{\text{th}}) - \frac{1}{\text{A}} I_{\text{K}}, \tag{8.18}$$

where $\text{C}_{\text{mem}} = \text{C}_1 + \text{C}_2$, $\text{C}_{\text{Ca}} = \text{C}_3$, and $\text{Q}_{\text{T}} = \text{C}_{\text{Ca}}\text{U}_{\text{T}}/\kappa$. Following the same procedures we used to obtain Equations 8.12 and 8.6, we integrate these equations over the interspike interval (t_n, t_{n+1}) to obtain

$$\text{Q}_{\text{th}} = \text{I}_1 t_{n+1} - \text{I}_0 t_n - \text{AQ}_{\text{T}} \ln\left(\frac{I_{\text{K}n}}{\text{AQ}_{\text{T}}}(t_{n+1} - t_n) + 1\right), \tag{8.19}$$

$$I_{\text{K}}(t) = \frac{I_{\text{K}n}}{\frac{I_{\text{K}n}}{\text{AQ}_{\text{T}}}(t - t_n) + 1}, \tag{8.20}$$

when the input current is stepped from I_0 to I_1 at time $t = 0$. That is, the step occurs within the interval (t_n, t_{n+1}), and these spike times are measured relative to the step.

I_{Kn} is the integrator's output current at the time that the last spike before the step occurred. Substituting the expression for $I_K(t)$ given by the second equation into Equation 8.12 and integrating yields the natural-log term in the second equation.

Our next goal is to solve Equation 8.19 for t_n as a function of t_{n+1}, and of the step height $I_1 - I_0$. Then we can differentiate the result to obtain the time-scaling factor, which gives us the probability density function for spiking. Before we proceed, however, it is instructive to consider the steady-state behavior.

8.6.1 Effect on Steady-State Behavior

For a constant input current, I_0=I_1=I_{in}, Equation 8.19 becomes

$$Q_{th} = I_{in}\Delta_n - AQ_T \ln\left(\frac{I_{Kn}}{AQ_T}\Delta_n + 1\right), \tag{8.21}$$

where $\Delta_n \equiv t_{n+1} - t_n$ is the interspike interval. When adaptation is complete, the interspike intervals become equal and we have $I_{Kn} = \alpha AQ_T/\Delta_n$, according to Equation 8.11. Hence, substituting this expression for I_{Kn} into Equation 8.21, we obtain

$$\Delta_n = \frac{Q_{th} + Aq_\alpha}{I_{in}} = \gamma\frac{Q_{th}}{I_{in}} \tag{8.22}$$

(remember that $q_\alpha = Q_T \ln(1 + \alpha)$). Amazingly, this result predicts a linear relationship between spike frequency and input current. It becomes obvious why this relationship holds if we observe that, in steady state, the input current must supply the charge $Q_{th} = C_2V_{dd}$ to C_1 and C_2, and supply the charge Aq_α removed by the integrator during the interspike interval, Δ_n, where q_α is the quantity of charge added to the integration capacitor by each spike. Notice that firing-rate adaptation reduces the firing rate by a factor of

$$\gamma \equiv 1 + \frac{Aq_\alpha}{Q_{th}}; \tag{8.23}$$

I call this parameter the **firing-rate adaptation attenuation factor**.

It is important to know how $I_K(t)$ varies relative to $I_{in}(t)$. Two distinct regimes of operation are feasible. During the $(n + 1)$th interval:

1. $I_{\mathrm{K}}(t) > I_{\mathrm{in}}$ for $t_n < t < u_n$ and $I_{\mathrm{K}}(t) < I_{\mathrm{in}}$ for $u_n < t < t_{n+1}$. That is, the packet of charge added to V_{Ca} is large enough to boost I_{K} above I_{in} immediately after a spike occurs, and there is therefore a net outward current that discharges V_{mem}. Then I_{K} decays, and eventually it becomes less than I_{in} at $t = u_n$; now, there is a net inward current that charges V_{mem}, and eventually it reaches threshold at $t = t_{n+1}$.

2. $I_{\mathrm{K}}(t) < I_{\mathrm{in}}$ for $t_n < t < t_{n+1}$. That is, the packet of charge is so small that I_{K} never exceeds I_{in}. In that case, the net input current is always inward, and V_{mem} starts recharging immediately after the spike occurs.

Equations 8.19 and 8.22 are valid over both regimes of operation, as long as the current mirror's output device remains saturated. The second regime is the preferred mode of operation, because V_{mem} stays close to the threshold, making the latency shorter and less variable.

To guarantee that the circuit operates in regime 2, we need to show only that it is in this regime at equilibrium. That is sufficient because any subsequent increases in the input cannot make I_{K} exceed I_{in}. For the equilibrium condition, I_{K} is less than I_{in} for all t if $I_{\mathrm{K}n}$ is less than I_{in}. Expressing $I_{\mathrm{K}n}$ in terms of the interspike interval Δ_n, and expressing Δ_n itself in terms of I_{in}, we find that

$$I_{\mathrm{K}n} = \frac{(\gamma - 1)\alpha}{\gamma \ln(1 + \alpha)} I_{\mathrm{in}}. \tag{8.24}$$

Therefore, the circuit operates in regime 2 if

$$\frac{\ln(1 + \alpha)}{\alpha} > 1 - \frac{1}{\gamma}.$$

Since $\ln(1 + \alpha) < \alpha$ and $\lim_{\alpha \to 0} \ln(1 + \alpha)/\alpha = 1$, the acceptable values for γ become larger as α becomes smaller. It should come as no surprise that smaller values of α make the circuit more likely to operate in regime 2.

Given a value of α, we can use the preceding inequality to obtain a limit on the

values of γ :

$$\gamma < \frac{\alpha}{\alpha - \ln(1+\alpha)}. \tag{8.25}$$

Smaller values of γ constrain operation to regime 2 because of the concommitant reduction in A (see Equation 8.23) which produces faster decay rates (see Equation 8.6), allowing larger values of α to be used. For $\alpha = 0.1$, which is about the largest value we would use, we must have $\gamma < 21.3$. Since α is so small, we can use the truncated Taylor series approximation for the log:

$$\ln(1+x) = x - \tfrac{1}{2}x^2 + R_3(x); \tag{8.26}$$

$$R_3(x) < \tfrac{1}{3}x^3 \quad \text{for } x > 0. \tag{8.27}$$

This approximation gives a conservative upper bound for γ if we drop $R_3(\cdot)$: $\gamma < 2/\alpha$. For $\alpha = 0.1$, this expression gives 20, compared to the exact upper limit of 21.3. The desire to operate in regime 2 imposes a serious tradeoff on the circuit's performance. If we want a large adaptation-attenuation factor, we must use a small charge quantum, making the number of spikes required to adapt excessively large.

8.6.2 Effect on Time-Scaling Function

We are now in a position to compute the adaptive neuron's time-scaling function. But first, we must solve Equation 8.19 for t_n. We can rewrite that equation as

$$\mathrm{Q_{th}} = \mathrm{I_1} t_{n+1} - \mathrm{I_0} t_n - \mathrm{A Q_T} \ln\left(\alpha \frac{t_{n+1} - t_n}{T_{\mathrm{K}n}} + 1\right), \tag{8.28}$$

where $T_{\mathrm{K}n} \equiv \alpha \mathrm{A Q_T}/I_{\mathrm{K}n}$. When adaptation is complete, the interspike interval, $\Delta_n \equiv t_{n+1} - t_n$, becomes equal to $T_{\mathrm{K}n}$ and Equation 8.24 gives the relationship between $I_{\mathrm{K}n}$ and the steady input I_{in}. But Δ_n is always less than $T_{\mathrm{K}n}$ during the course of adaptation. Consequently, α is an upper bound on the first term in the argument of the log. Hence, we can use the Taylor series approximation for the log (Equation 8.26).

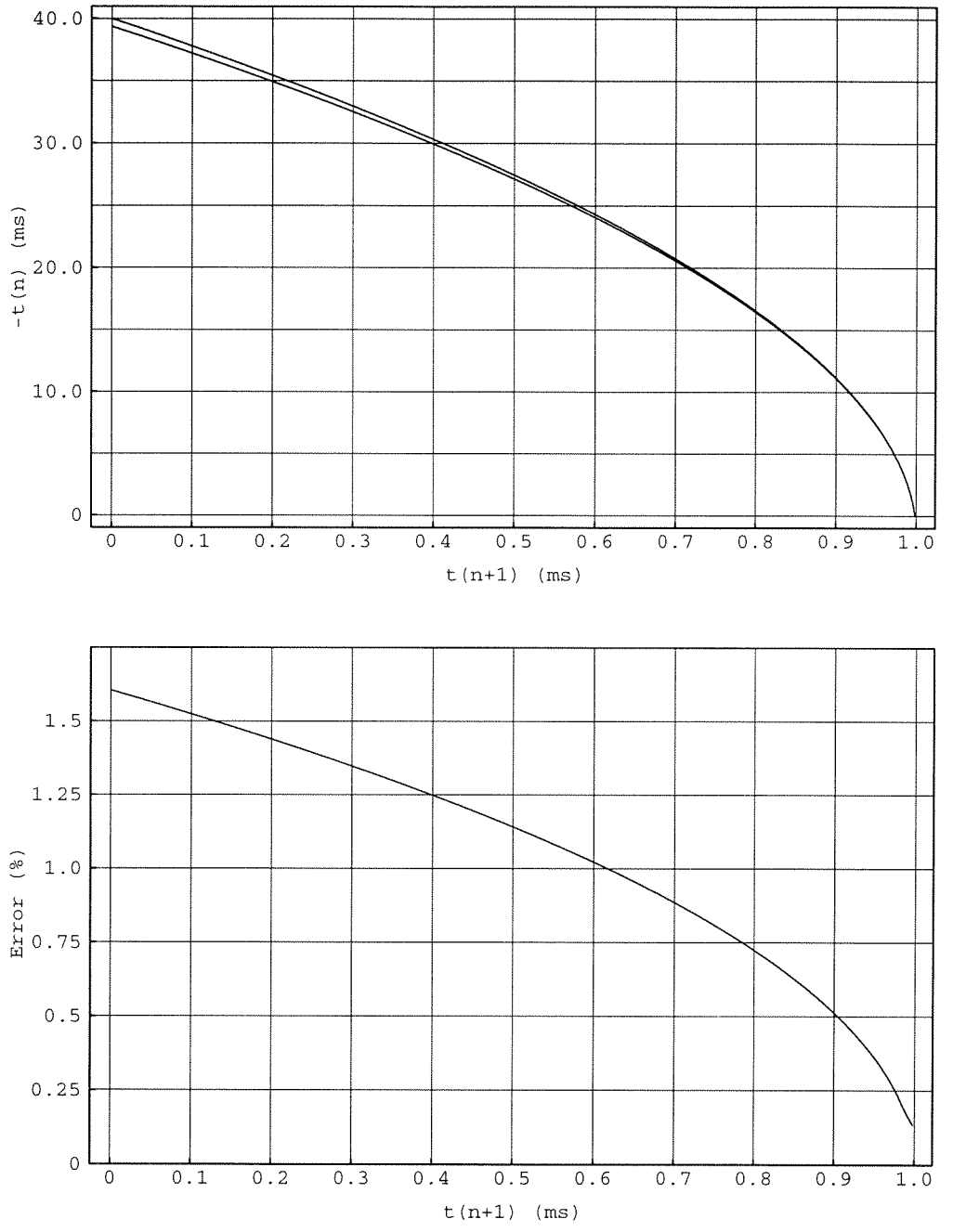


Figure 8.7: Spike Timing Relative to Input Step

Here, $-t_n$ is the time at which the step occured, minus the time at which the last spike occurred; t_{n+1} is the time at which the circuit responds by emitting a spike minus the time at which the step occurred; that is, it is the latency of the neuron. The top plot shows the exact relationship between t_n and t_{n+1} (higher curve), which I obtained by solving Equation 8.28 numerically, and the approximate solution (lower curve) given by Equation 8.30. The bottom plot shows the error in the approximation. This model includes only firing-rate adaptation.

The result is

$$\frac{\mathrm{Q_{th}}}{\alpha \mathrm{AQ_T}} = \frac{\mathrm{I}_1 t_{n+1} - \mathrm{I}_0 t_n}{\alpha \mathrm{AQ_T}} - \frac{t_{n+1} - t_n}{T_{\mathrm{K}n}} + \frac{\alpha}{2}\left(\frac{t_{n+1} - t_n}{T_{\mathrm{K}n}}\right)^2 - R_3(t_{n+1} - t_n),$$

where

$$R_3(t_{n+1} - t_n) < \frac{\alpha^2}{3}\left(\frac{t_{n+1} - t_n}{T_{\mathrm{K}n}}\right)^3.$$

Dropping $R_3(\cdot)$ and rearranging terms, we obtain the following quadratic equation in $t_{n+1} - t_n$:

$$\frac{\alpha}{2}\left(\frac{t_{n+1} - t_n}{T_{\mathrm{K}n}}\right)^2 + \left(\frac{T_{\mathrm{K}n}\mathrm{I}_0}{\alpha \mathrm{AQ_T}} - 1\right)\frac{t_{n+1} - t_n}{T_{\mathrm{K}n}} + \frac{(\mathrm{I}_1 - \mathrm{I}_0)t_{n+1} - \mathrm{Q_{th}}}{\alpha \mathrm{AQ_T}} = 0.$$

We can now solve this equation for $t_{n+1} - t_n$, and, use the solution to calculate t_n, given t_{n+1}. First, however, we can simplify the equation by replacing $\alpha \mathrm{AQ_T}$ with $T_{\mathrm{K}n} I_{\mathrm{K}n}$, using the defintion of $T_{\mathrm{K}n}$. Thus, we obtain

$$\frac{\alpha}{2}\left(\frac{t_{n+1} - t_n}{T_{\mathrm{K}n}}\right)^2 + \epsilon\left(\frac{t_{n+1} - t_n}{T_{\mathrm{K}n}}\right) + (\epsilon + 1)\left(\left(\frac{\mathrm{I}_1}{\mathrm{I}_0} - 1\right)\frac{t_{n+1}}{T_{\mathrm{K}n}} - \frac{\mathrm{Q_{th}}}{T_{\mathrm{K}n}\mathrm{I}_0}\right) = 0,$$

where $\epsilon \equiv \mathrm{I}_0/I_{\mathrm{K}n} - 1$. Assuming that the circuit has had time to adapt to I_0, we can express $I_{\mathrm{K}n}$ in terms of I_0 using Equation 8.24, and obtain an expression for ϵ :

$$\epsilon = \frac{\gamma \ln(1 + \alpha)}{(\gamma - 1)\alpha} - 1. \tag{8.29}$$

Finally, we solve the quadratic equation and obtain

$$\frac{t_{n+1} - t_n}{T_{\mathrm{K}n}} = \sqrt{\left(\frac{\epsilon}{\alpha}\right)^2 + \frac{2(1 + \epsilon)}{\alpha}\left(\frac{1}{\gamma} - \left(\frac{\mathrm{I}_1}{\mathrm{I}_0} - 1\right)\frac{t_{n+1}}{T_{\mathrm{K}n}}\right)} - \frac{\epsilon}{\alpha}. \tag{8.30}$$

I have replaced $\mathrm{Q_{th}}/T_{\mathrm{K}n}\mathrm{I}_0$ with $1/\gamma$ since $T_{\mathrm{K}n}$ is equal to $\gamma \mathrm{Q_{th}}/\mathrm{I}_0$ when adaptation is complete.

Figure 8.7 compares the approximate solution with the exact solution that we obtained by solving Equation 8.28 numerically. As you can see, the approximation is within 2 per cent for the value of α used, which was 0.05. The other parameter values

are $Q_{th} = 0.1pC$, $Q_T = 15fC$, $\gamma = 40$, which gave $A = 5329$ and $\epsilon = 8.239 \times 10^{-4}$. The input current was stepped from $I_0 = 100pA$ to $I_1 = 200pA$, which gave $I_{Kn} = 99.92pA$, and $T_{Kn} = 40.0ms$. Thus, the interspike interval is 40ms after the circuit adapts to the 100pA input current. If the step occurs 40ms after the last spike happened ($-t_n = 40ms$), the neuron responds with 0 latency, because its membrane voltage is at the threshold. If the last spike happened earlier ($-t_n < 40ms$), the latency will be longer, because the neuron is further away from threshold. The latencies range from 0 to 1ms—40 times smaller than the initial interspike interval, since $\gamma = 40$ and $(I_1 - I_0)/I_0 = 1$.

After the step occurs, the membrane voltage moves rapidly because all the extra current goes to charge the membrane capacitance. The membrane voltage spends more time on this rapid trajectory as $-t_n$ gets smaller. Hence, smaller values of $-t_n$ map to smaller intervals on the t_{n+1} axis. This compression explains the increasing slope of the curve as $-t_n$ gets smaller and t_{n+1} gets larger. It also means that the longer latencies are much more probable, assuming that the spikes were uniformly distributed initially. Indeed, the probability of firing after the step goes like dt_n/dt_{n+1}, as I showed mathematically in Section 8.5.

The error in the approximation gets larger for large values of $-t_n$ because this approximation essentially models the $1/(t+1)$ decay of the potassium current with a constant slope, as we replaced the integral of the current with a quadratic. This linear-decay model works well if the difference between the input current and the potassium current is large compared to the amount by which the potassium current changes during the interspike interval—a condition that certainly holds after the step occurs, provided $I_1 - I_0 \gg I_0/\gamma$. The larger t_n gets, the smaller t_{n+1} gets, and the less time we spend with a large difference between the input current and the potassium channel current; hence, the less accurate the approximation gets.

A few sanity checks are in order at this point. When $t_{n+1} = 0$, we expect $t_n = -T_{Kn}$. Equation 8.30 gives

$$-\frac{t_n}{T_{Kn}} = \sqrt{\left(\frac{\epsilon}{\alpha}\right)^2 + \frac{2(1+\epsilon)}{\gamma\alpha}} - \frac{\epsilon}{\alpha}.$$

Solving Equation 8.29 for $1/\gamma$, and using the Taylor series approximation for the log given in Equation 8.26, we find that

$$\frac{2(1+\epsilon)}{\gamma\alpha} = 2\frac{(1+\epsilon)\alpha - \ln(1+\alpha)}{\alpha^2}, \tag{8.31}$$

$$= 1 + \frac{2\epsilon}{\alpha} - \frac{2R_3(\alpha)}{\alpha^2}, \tag{8.32}$$

Replacing $R_3(\cdot)$ with its upper limit, substituting in the expression for t_n, and using a first-order Taylor series approximation for the square-root, gives

$$\begin{aligned} -\frac{t_n}{T_{\mathrm{K}n}} &> \sqrt{\left(\frac{\epsilon}{\alpha}+1\right)^2 - \frac{2}{3}\alpha} - \frac{\epsilon}{\alpha}, \\ &\simeq 1 - \frac{1}{3}\frac{\alpha}{\frac{\epsilon}{\alpha}+1}. \end{aligned}$$

Thus, Equation 8.30 yields $-t_n/T_{\mathrm{K}n}$ close to 1 for $t_{n+1} = 0$. The error is less than 1.640 per cent for the parameter values given; the actual error was 1.605 per cent. The error grows as ϵ decreases; this dependence is consistent with the argument given in the previous paragraph.

On the other hand, when $t_n = 0$, we have to set $t_n = 0$ in the left-hand side of Equation 8.30, and to solve the resulting quadratic equation to find the corresponding value of t_{n+1}. The result is

$$\frac{\hat{t}_{n+1}}{T_{\mathrm{K}n}} = \frac{1+\epsilon}{\alpha}\left(\sqrt{\left(\frac{\mathrm{I}_1}{\mathrm{I}_0} - \frac{1}{1+\epsilon}\right)^2 + \frac{2\alpha}{\gamma(1+\epsilon)}} - \left(\frac{\mathrm{I}_1}{\mathrm{I}_0} - \frac{1}{1+\epsilon}\right)\right). \tag{8.33}$$

This value is important because it is the largest value that t_{n+1} can have and therefore it determines the range over which Equation 8.30 is valid. It is also the interspike interval immediately after the step, and therefore the firing rate of the neuron is $1/\hat{t}_{n+1}$. Since $\alpha/(1+\epsilon) \simeq 2/\gamma$, the second term in the argument for the square root is negligible if $\mathrm{I}_1/\mathrm{I}_0 - 1/(1+\epsilon) \gg 2/\gamma$. In that case, we can use the first-order

approximation for the square-root and obtain

$$\frac{\hat{t}_{n+1}}{T_{\mathrm{K}n}} \simeq \frac{\mathrm{I}_0}{\mathrm{I}_1 - \mathrm{I}_0/(1+\epsilon)} \frac{1}{\gamma}. \tag{8.34}$$

This estimate is reasonable because we would expect the new interspike interval to be inversely proportional to the extra current, which is equal to $\mathrm{I}_1 - I_{\mathrm{K}n} = \mathrm{I}_1 - \mathrm{I}_0/(1+\epsilon)$, and to be γ times less than $T_{\mathrm{K}n}$, because no adaptation has taken place. We must be careful, however: When $\epsilon = 0$, this expression for $\hat{t}_{n+1}$ gives $\hat{t}_{n+1} - t_n = 0$ when used in Equation 8.30, and that result is wrong; the full-blown expression gives the right answer.

Proceeding with the derivation, we differentiate Equation 8.30 to obtain the time scaling factor, and we use the latter to compute the probability density function, which is also the synchronous firing rate:

$$p_1(t_{n+1}) = \frac{dt_n}{dt_{n+1}} p_0(t_n) = \left(1 + \frac{\frac{1+\epsilon}{\alpha}\left(\frac{\mathrm{I}_1}{\mathrm{I}_0} - 1\right)}{\sqrt{\left(\frac{\epsilon}{\alpha}\right)^2 + \frac{2(1+\epsilon)}{\alpha}\left(\frac{1}{\gamma} - \left(\frac{\mathrm{I}_1}{\mathrm{I}_0} - 1\right)\frac{t_{n+1}}{T_{\mathrm{K}n}}\right)}} \right) \frac{1}{T_{\mathrm{K}n}}. \tag{8.35}$$

This approximate expression for the probability density function is compared with the exact result obtained by numerically solving and differentiating Equation 8.28 in Figure 8.8. The error is -3 per cent or less over almost the entire range. We use the same parameter values and a Δt of 0.1μs to compute the derivative of the exact solution numerically. The synchronous firing rate starts out at 526Hz, rising steadily initially and then much more rapidly as we approach the maximum delay $\hat{t}_{n+1}$; it reaches a maximum of 12.1KHz when $t = \hat{t}_{n+1}$. Compare the synchronous firing rate with the neuron's firing rate immediately after the step, which is $1/\hat{t}_{n+1}$ or 1KHz.

8.6.3 Effect on Latency and Synchronicity

Finally, we can write expressions for the latency, μ, and the synchronicity, λ, with firing-rate adaptation. By definition, the latency is given by the delay of the peak in the synchronous firing rate and the synchronicity is equal to the ratio between the

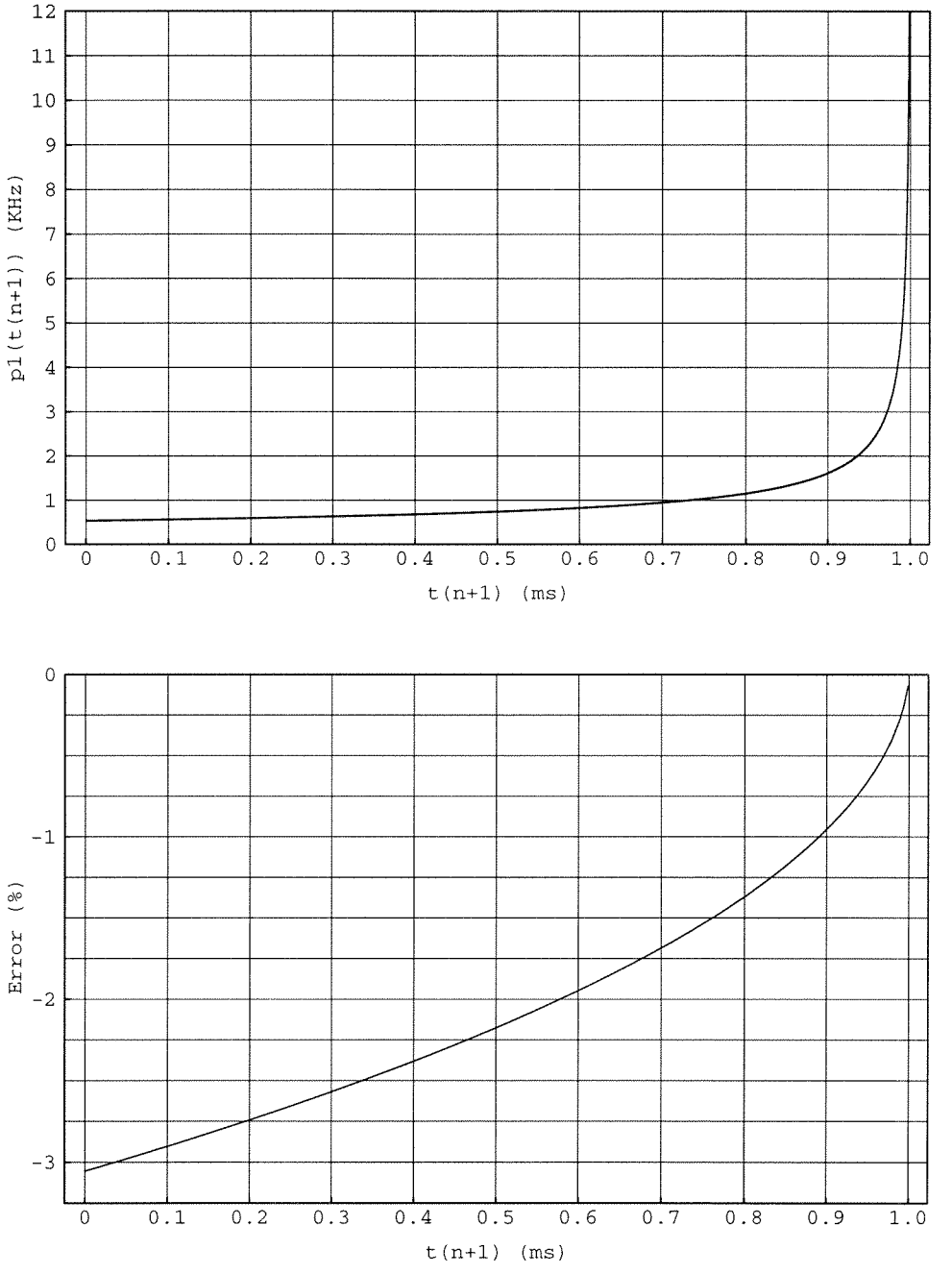


Figure 8.8: Firing Probability Distribution

Here, t_{n+1} is the time at which the spike occurs relative to the input step. We calculate the probability density function by multiplying the prior distribution by the time-scaling factor dt_n/dt_{n+1} obtained by differentiating the curves in Figure 8.7. The results plotted here are for a uniform prior distribution. The values obtained analytically from our approximate solution (Equation 8.35) and numerically from the exact solution (Equation 8.28) are plotted in the top graph; the two curves are virtually indistinguishable. The error is plotted in the bottom graph. This model includes only firing-rate adaptation.

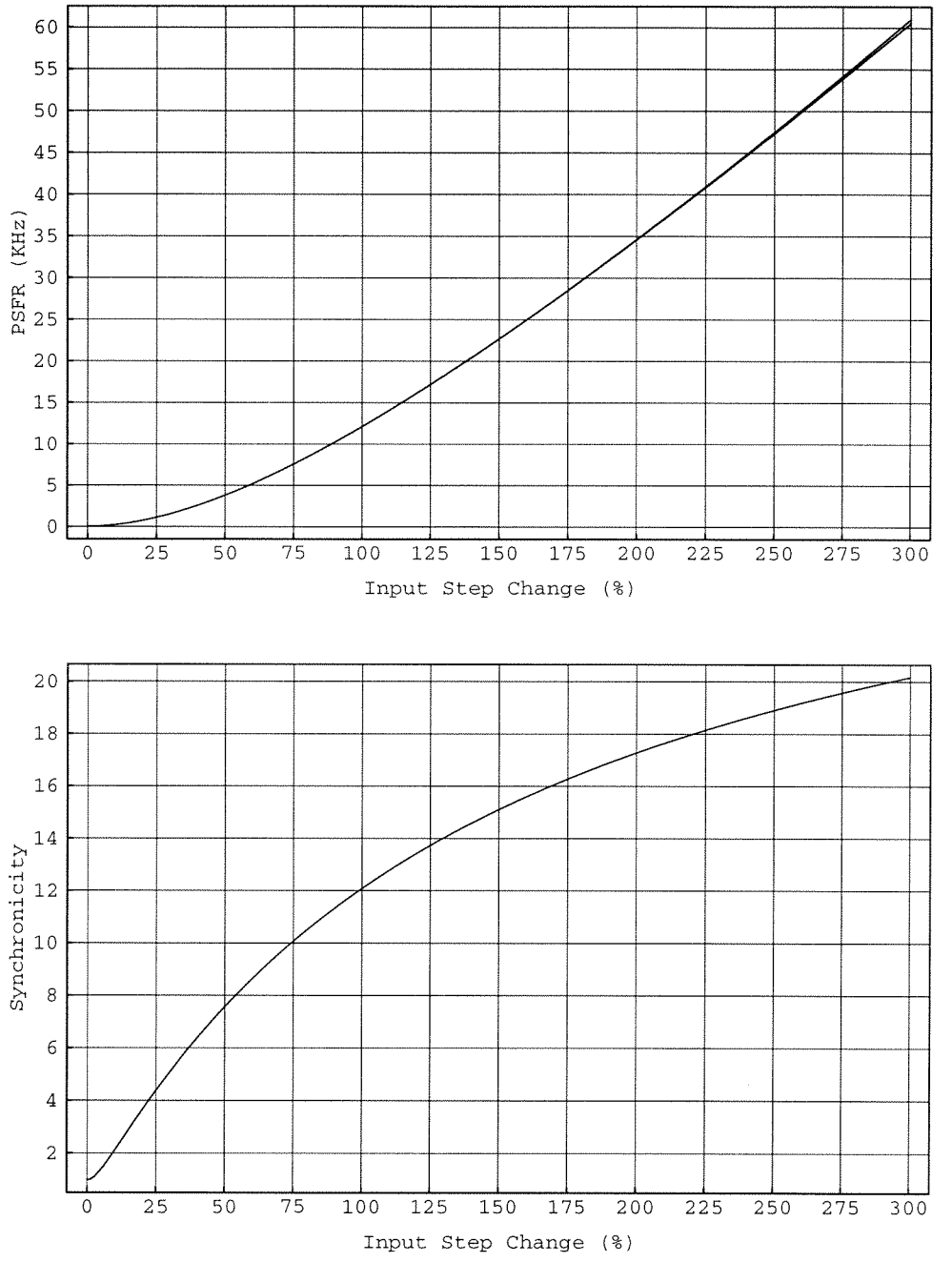


Figure 8.9: PEAK SYNCHRONOUS FIRING RATE AND SYNCHRONICITY
The synchronous firing rate is the total instantaneous firing rate divided by the number of neurons or the number of trials; it is equal to the probability density function. Therefore, the peak synchronous firing rate is obtained from the peak in the probability density function shown in Figure 8.8. The graph on the bottom shows the synchronicity of the response, which is defined as the ratio of the peak synchronous firing rate to the firing rate of the neuron immediately after the step. This model includes only firing-rate adaptation.

peak synchronous firing rate and the neuron's firing rate immediately after the step. Hence, using the approximate value for $\hat{t}_{n+1}$ given by Equation 8.34 to compute μ, we have

$$\mu \equiv \hat{t}_{n+1} \simeq \frac{\mathrm{I}_0}{\mathrm{I}_1 - \mathrm{I}_0/(1+\epsilon)} \frac{T_{\mathrm{K}n}}{\gamma}, \tag{8.36}$$

$$\lambda \equiv p_1(\hat{t}_{n+1})\hat{t}_{n+1} = \left(1 + \frac{\frac{1+\epsilon}{\alpha}\left(\frac{\mathrm{I}_1}{\mathrm{I}_0} - 1\right)}{\frac{\hat{t}_{n+1}}{T_{\mathrm{K}n}} + \frac{\epsilon}{\alpha}}\right) \frac{\hat{t}_{n+1}}{T_{\mathrm{K}n}}. \tag{8.37}$$

For the parameter values that we are using, $2/\gamma$ is 5 per cent and so the approximation $\mathrm{I}_1/\mathrm{I}_0 - 1/(1+\epsilon) \gg 2/\gamma$, holds for a percentage change of 25 per cent or more in the input current. For λ, we used Equation 8.30 to obtain a simpler expression for the square-root term in the denominator, and then set $t_{n+1} = \hat{t}_{n+1}$ and $t_n = 0$.

Substituting the approximate expression for $\hat{t}_{n+1}/T_{\mathrm{K}n}$ given by Equation 8.34 in the expression for λ, we obtain

$$\lambda = \frac{1}{\gamma\left(\frac{\mathrm{I}_1}{\mathrm{I}_0} - \frac{1}{1+\epsilon}\right)} + \frac{\frac{1+\epsilon}{\alpha}\left(\frac{\mathrm{I}_1}{\mathrm{I}_0} - 1\right)}{1 + \frac{\epsilon\gamma}{\alpha}\left(\frac{\mathrm{I}_1}{\mathrm{I}_0} - \frac{1}{1+\epsilon}\right)}.$$

We can neglect the first term if $\mathrm{I}_1/\mathrm{I}_0 - 1/(1+\epsilon) \gg 2/\gamma$. The second term displays two distinct behaviors. If ϵ is zero the synchronicity increases linearly with the change in the input current over the entire range, and the peak synchronous firing rate follows a square law. If ϵ is nonzero, this behavior is observed for only small changes in the input current (i.e., $\mathrm{I}_1/\mathrm{I}_0 - 1/(1+\epsilon) \ll \alpha/\epsilon\gamma$). For large changes, the synchronicity plateaus at $(1+\epsilon)/\epsilon\gamma$, and the peak synchronous firing rate changes linearly. For the parameter values that we are using, the transition occurs at 31.9 per cent, and the synchronicity attains a maximum value of 30.4.

The dependencies of the peak synchronous firing rate and the synchronicity of the response on the step amplitude are shown in Figure 8.9 for the same parameter values given previously. As expected, the peak synchronous firing rate follows a square-law for changes less than 30 per cent and then becomes linear for larger changes. Hence, the synchronicity increases initially, and then plateaus.

Compared to the integrate-and-fire neuron, the adaptive neuron with potassium-dependent calcium channels displays much shorter latencies and much higher synchronicities. For a 100–per-cent change in the input, the latency is γ times less than the interspike interval before the step, compared with only 4 times less for the integrate-and-fire neuron. It is a factor of 4 for the integrate-and-fire neuron because the latency is one-half of the interspike interval after the step increase in the input current. And the interspike interval after the step is, in turn, half the interspike interval before the step, since the current has doubled. Since γ can be much larger than 4 if extremely small values of α are acceptable (remember that $\gamma < 2/\alpha$ if the circuit operates in regime 2), the adaptive neuron can have much shorter latencies that the integrate-and-fire neuron. The synchronicity is also much higher for the adaptive neuron because longer delays are much more probable. For a 100–per-cent change in the input, the synchronicity is approximately $1/\alpha$ for $\epsilon \ll \alpha/\gamma$. In this case as well we can obtain high synchronicities if we use small values of α.

8.7 Test Results

In this section, I present measurements from the adaptive neuron circuit, with and without membrane–time-constant adaptation. To turn off time-constant adaptation, I inserted a device in series with the integrator's output transistor, between the drain of the output device and the membrane-voltage node (see Figure 8.2). By operating this device as a current buffer (also called a cascode), I could isolate the membrane-voltage node from the calcium-integration node, and could thus turn off time-constant adaptation. By turning on this device hard, I could short the drain of the output device to V_{mem}, and could thus turn on time constant adaptation.

Figures 8.10 and 8.11 show the response of the adaptive neuron to a step change in input current, with and without membrane–time-constant adaptation, respectively. These figures demonstrate the integration of current pulses by the diode-capacitor integrator, and the progressive reduction in spike frequency, as the integrator's output current increases. When time-constant adaptation is turned on, the spike frequency

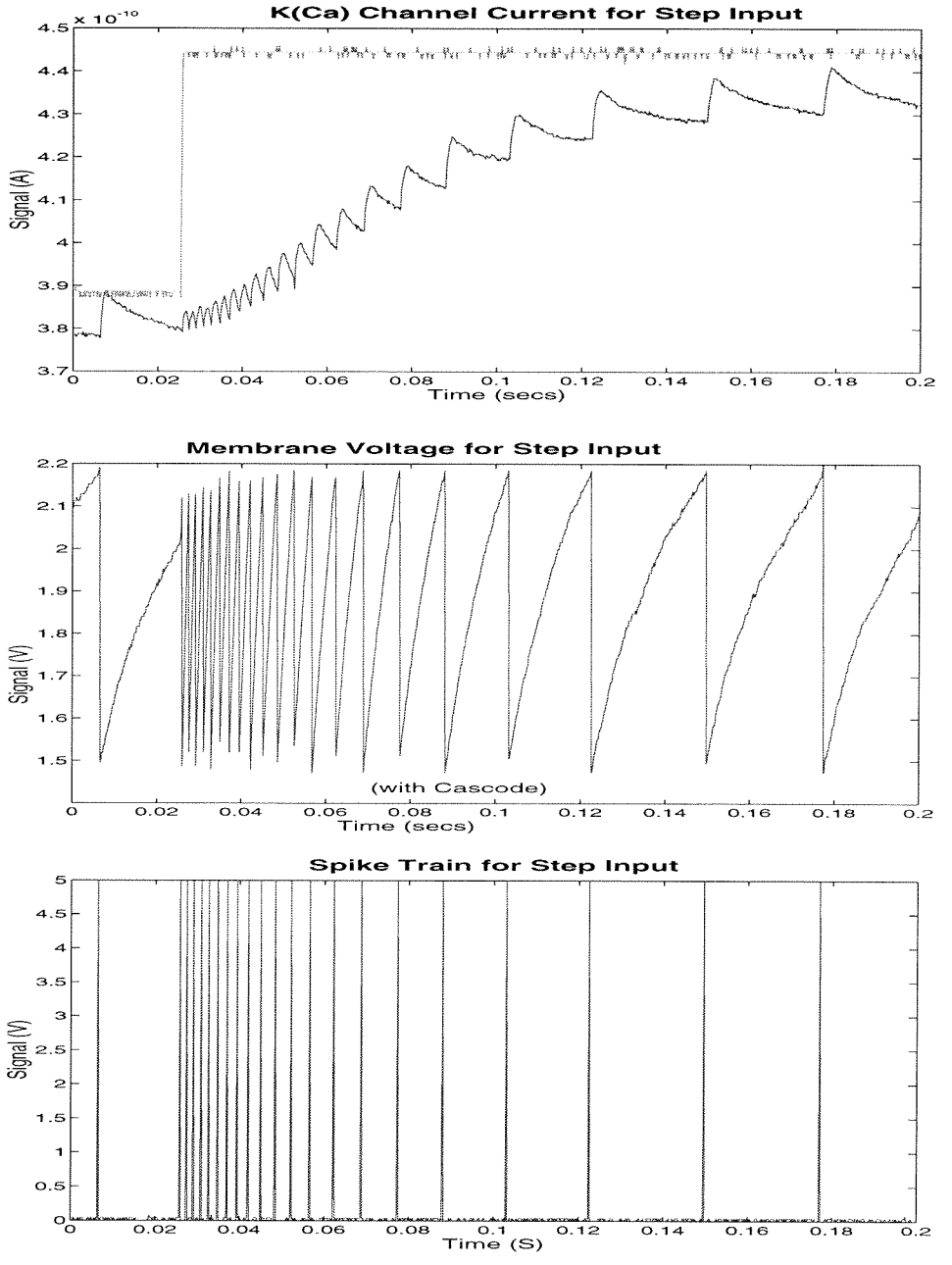


Figure 8.10: Adaptive Neuron's Step Response 1

Top: The neuron's input current and the integrator's output current. Middle: The neuron's membrane voltage ramping up between the reset (1.5V) and threshold levels (2.2V). The difference between the input current and the integrator's output ramps up the membrane voltage as the excess current charges the input capacitance. Bottom: The neuron's spike train. A spike is generated each time that the membrane voltage reaches threshold, and the membrane voltage is reset immediately afterward. Time-constant adaptation was turned off.

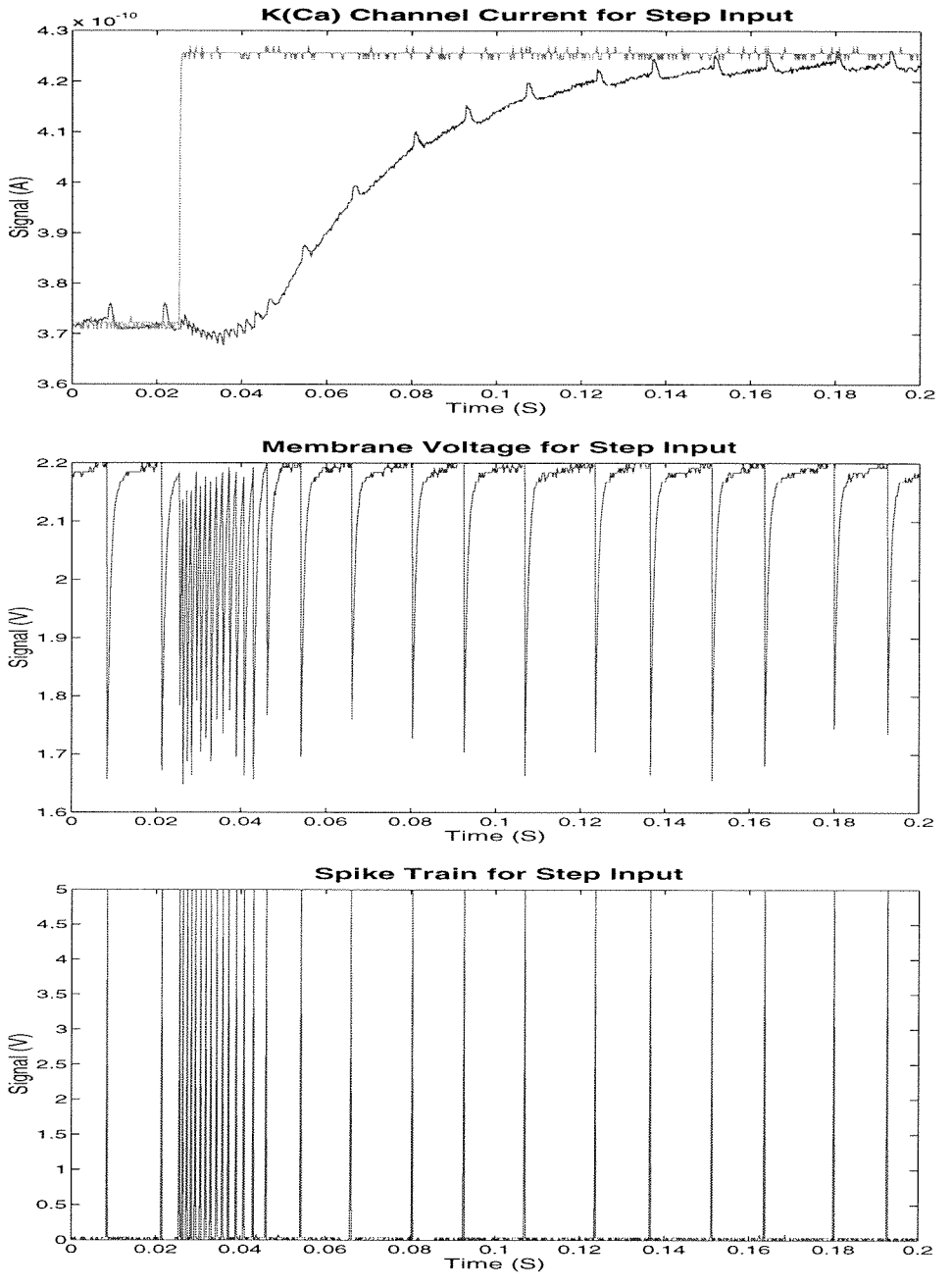


Figure 8.11: Adaptive Neuron's Step Response 2

This figure shows traces of the input current (top), the integrator's output current (top), the membrane voltage (middle), and the spike train (bottom), it is just like Figure 8.10, except that membrane–time-constant adaptation was turned on. The membrane voltage increases rapidly immediately after it is reset, and another spike is generated immediately if there is excess input current. Thus, a clump of spikes is generated in response to the step, and the circuit adapts more abruptly. In steady state, when there is no excess current, the membrane voltage homes in on the threshold slowly.

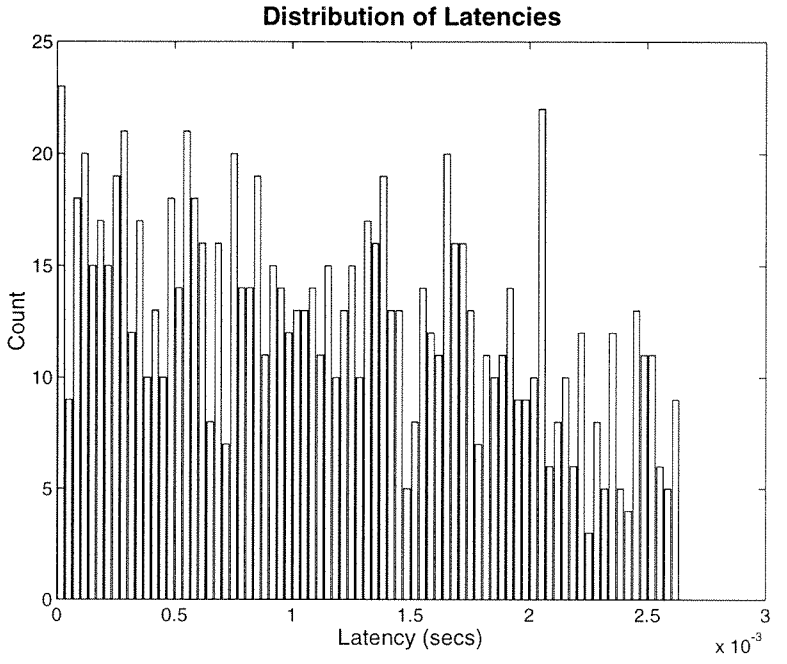
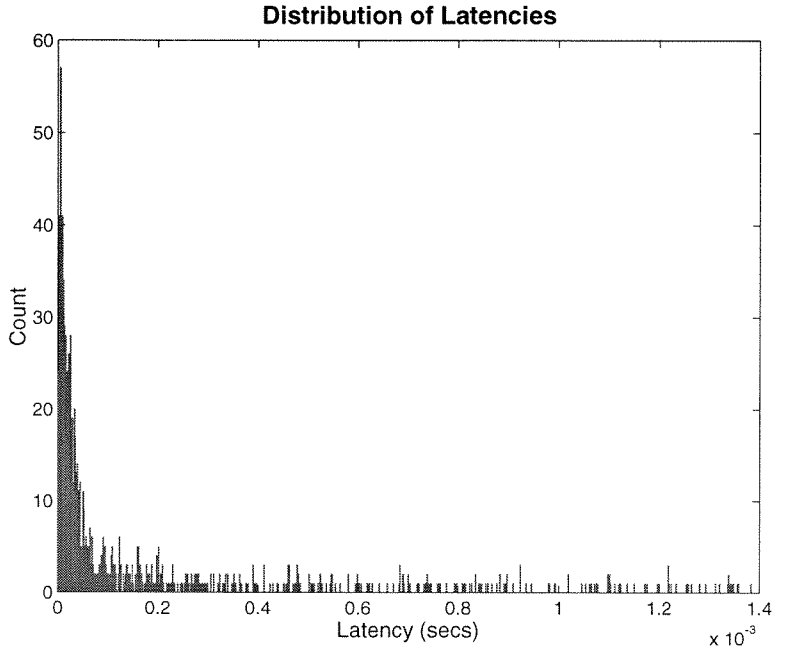


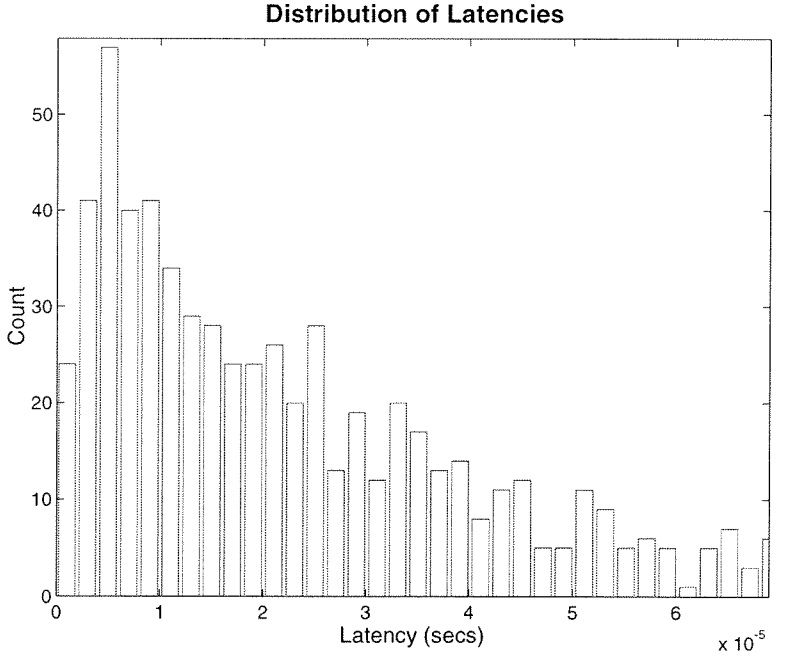
Figure 8.12: ADAPTIVE NEURON'S LATENCY AND SYNCHRONICITY 1
I measured the distribution of latencies by performing 1000 trials and plotting a histogram of the results. There are 80 bins, each of size 33.3μs; the longest latency is 2.63ms. The spikes are distributed more or less uniformly, with a slight tendency toward shorter latencies. For a uniform distribution, each bin would have 12.5 spikes. The mean latency was roughly 1.3ms and the synchronicity was close to 1. Membrane–time-constant adaptation was turned off.

decreases abruptly.

Figures 8.12 and 8.13 show the distribution of latencies for the adaptive neuron, with and without membrane–time-constant adaptation, respectively. These figures demostrate that spikes are more or less uniformly distributed when only spike-frequency adaptation is present, and the distribution is skewed heavily toward shorter latencies when both spike-frequency and time-constant adaptation are present. In the former case, the bias toward longer latencies predicted by the theory (see Figure 8.8) is not evident in the data. There are two reasons for this discrepancy: (1) the large bins that I used tend to smear out sharp peaks; and (2) the current buffer does not completely eliminate capacitive coupling between the membrane-voltage node and the calcium-integration node, so there is residual time-constant adaptation. This



(a)



(b)

Figure 8.13: Adaptive Neuron's Latency and Synchronicity 2
(a) The distribution of latencies with time-constant adaptation turned on. (b) The distribution around the peak. I performed 1000 trials, and I distributed the results among 700 bins, each of size 2μs; the longest latency was 1.4ms. Each bin would have 1.43 spikes if the spikes were distributed uniformly. However, the distribution is skewed heavily toward extremely short latencies. The latency of the peak is 6μs, and the synchronicity is 39.4.

conclusion is supported by the progressive decrease in slope in the membrane voltage trajectories (see Figure 8.10). To model the behavior accurately, we need a complete theory that includes time-constant adaptation

8.8 Discussion

I proposed and studied a simple adaptive spiking neuron circuit in this chapter. I introduced three simple circuit elements to model the biophysics of voltage- and calcium-dependent potassium channels:

1. A diode-capacitor integrator models the accumulation and buffering of intracellular calcium.

2. Capacitive coupling between the membrane-voltage node and the calcium-integration node models the fast voltage dependence of the potassium channels.

3. A single transistor, with its gate tied to the calcium-integration node models the potassium-channel population.

I analyzed the effects of these mechanisms, with emphasis on spike timing, and compared my theoretical predictions with experimental measurements. I characterized spike-timing precision by measuring how much time the neuron takes to respond to a step change in its input by firing a spike. I measured the distribution of these firing times over several trials, and defined the latency as the position of the peak in the distribution, and the synchronicity as the height of the peak, normalized by the height of the uniform distribution. For the same average steady-state firing rate, the calcium dependence and the voltage dependence of the potassium channels improved the adaptive neuron's latency and synchroncity, compared with a simple integrate-and-fire model.

Calcium-dependent potassium channels improved latency by attenuating the steady-state firing rate, which could be over 40 times less than the firing rate immediately after the step. With the shorter interspike intervals immediately after the step, we

achieved shorter latencies without paying the price of a corresponding increase in steady-state firing rate. The calcium dependence had no effect on synchronicity, since it more or less preserved the uniform firing-probability distribution of the integrate-and-fire neuron.

Adding voltage-dependence improved synchronicity as well as latency by repolarizing the membrane rapidly, such that the membrane voltage was just shy of the threshold voltage most of the time. The high likelihood of finding the membrane voltage just below threshold reshaped the firing-probability distribution, skewing it heavily toward shorter latencies. As a result, the mean latency dropped from 0.7ms to 6μs, and the synchronicity shot up from 1 to 39.

These results call into question several common notions about how neurons encode information. Neurobiologists generally believe that the mean firing rate is a valid measure of the efficacy of a neuron in producing a response in its target. Furthermore, if the target neuron listens to several neurons, they obtain the net effect by summing their mean firing rates. For such linear summation to be valid, the postsynaptic neuron must smooth out fluctuations in firing rates, or the presynaptic neurons must fire at uniform rates and in an uncorrelated fashion. My measurements invalidate both assumptions, and are in agreement with more recent physiological studies [3].

First, neurons are exquisitely sensitive to small changes in their input, and can generate a spike in response to these changes in less than 1 millisecond. Consequently, instead of smoothing out variations in their inputs, they amplify these variations.

Second, the latencies are much shorter than the interspike interval, and so the instantaneous firing rate that the target neuron observes when several spike trains converge in its dendritic tree may be much higher than you would expect from simply summing the individual rates. For example, if the presynaptic neurons fire once every 10ms, on average, but all the spikes happen to occur during the first 1ms, then the instantaneous rate will be 10 times higher. In fact, the synchronicity tells us exactly how many times higher the instaneous firing rate is than the rate obtained by linear summation.

Neurons can use synchronicity to amplifiy their firing rates. We overlook this mechanism completely when we use mean firing rates and ignore spike timing.

Chapter 9 Neuromorphic VLSI: A Retina on a Chip

In this final chapter, I describe a retinomorphic vision chip that uses neurobiological principles to perform all four major operations found in biological retinae:

1. Continuous sensing for detection
2. Local automatic gain control for amplification
3. Spatiotemporal bandpass filtering for preprocessing
4. Adaptive sampling for quantization

All four operations are performed right on the focal plane, at the pixel level.

The first—and only—attempt to integrate these four operations was made by Misha Mahowald. The pixel that she designed, which is described in her monograph [4], used continuous sensing for detection, logarithmic compression for amplification, temporal highpass filtering for preprocessing, and a simple integrate-and-fire neuron for quantization. My work improves on, and extends, Mahowald's pioneering research in three ways:

1. By using local gain control for amplification, I extend the dynamic range without sacrificing sensitivity; logarithmic compression, in contrast, trades sensitivity for dynamic range.
2. By using a spatiotemporal bandpass for preprocessing, I cut out wideband spatial and temporal noise; highpass filtering, in contrast, amplifies high-frequency signals with poor signal-to-noise ratios.
3. By using an adaptive neuron for quantization, I increase the sampling rate—and reduce the latency—without increasing the average firing rate; a simple

Pixel-Level Operations	Pixels	Area (μm^2)	$L(\mu m)$	Area (L^2)	Devices
Detection (CCD [97])	962×654	5.05×5.55	—	—	1
Amplification (CMD [136])	660×492	7.3×7.6	0.4*	85.5	1
Amplification (APS [98])	256×256	20×20	0.9	492.8	5
Filtering and LAGC [131]	230×210	39.6×43.8	1.2	1204.5	12
Quantization (Σ-Δ [105])	64×64	60×60	1.2	2500	22
Quantization (PFM [137])	10×10	104×104	1.6	4225	17
Retinomorphic	64×64	106×98	2.0	2597	32

Table 9.1: TRENDS IN IMAGER DESIGN
L is the minimum gate length. CCD – charge-coupled device; CMD – charge-modulation device; APS – active pixel sensing; LAGC – local automatic gain control; Σ–Δ – sigma–delta modulation; PFM – pulse-frequency modulation. *Estimated.

integrate-and-fire neuron, in contrast, must maintain a high steady-state firing rate to sample high-frequency signals.

Like Mahowald's chip, my retinomorphic chip includes a random-access time-division multiplexed communication channel that reads out asynchronous pulse trains from a 64- $\times$ 64-pixel array in the imager chip. The communication channel transmits these spike trains to corresponding locations on a second chip that has a 64×64 array of integrators. Both chips are fully functional.

9.1 Smart-Pixel Arrays

The primary difference between retinomorphic imagers and conventional ones is that they perform—at the pixel level—all four operations listed in the introduction. The migration of more sophisticated signal processing down to the pixel level is driven by shrinking feature sizes in CMOS technology, which allows higher levels of integration to be achieved. The representative examples listed in Table 9.1 illustrate this trend; the pixel area, normalized by the square of the minimum gate length, is plotted versus the number of transistors in Figure 9.1.

If this trend continues as the feature sizes shrink, it will be possible to design pixels with 10 transistors that are 13 μm per side, and pixels with 35 transistors that

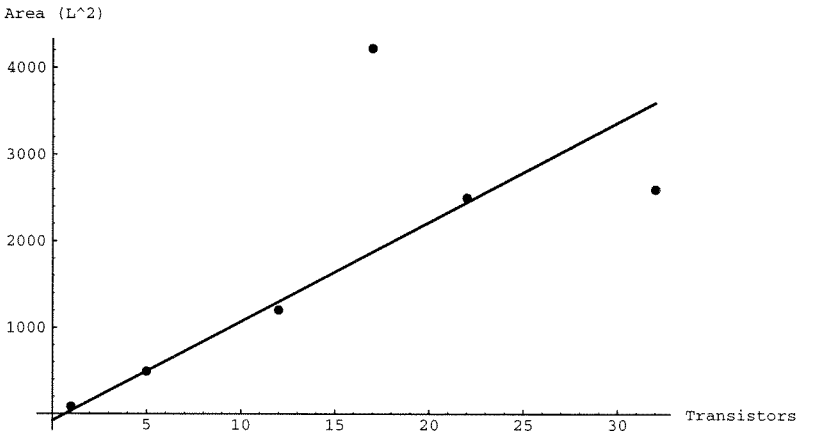


Figure 9.1: Scaling of Pixel Area

The pixel area is more or less proportional to transistor count. The points plotted are for the imagers listed in Table 9.1; the line is a linear-regression fit. The fit indicates that the pixel area increases by $115L^2$ for each additional transistor. Area is measured in units of the minimum gate length (L) squared.

are 25 μm per side, in a 0.4 μm process. As the size of the active devices becomes small compared to the sensor area—which is typically about 5 μm per side—it will become cost effective to shrink the detector area and to use lenselet arrays to focus the light [138], freeing up that area for additional image-processing functions. Shrinking the sensor along with the active devices would enable the scaling trend to continue unabated. Hence, it should be feasible to build a 730- $\times$ 730-pixel imager with 10 transistors per pixel, or a 380- $\times$ 380-pixel imager with 35 transistors per pixel, on a 1cm square die, with just over 5 million transistors in today's state-of-the-art 0.4μm CMOS process. In comparison, the human fovea has only about 200×200 cones, but the density is much higher: these cones occupy an area of just 0.6 mm $\times$ 0.6 mm!

It has been clear for over 20 years that, given the rate at which feature sizes are shrinking, CMOS technology will give us many more transistors than we know what to do with [139, 140]. At present, we are on the verge of the billion-transistor 1Gb DRAM chip, and there is a dire need for new pixel-parallel architectures to take advantage of the increasing numbers of transistors available [141]. I describe a retinomorphic vision system that addresses this need by mimicking biological sensory

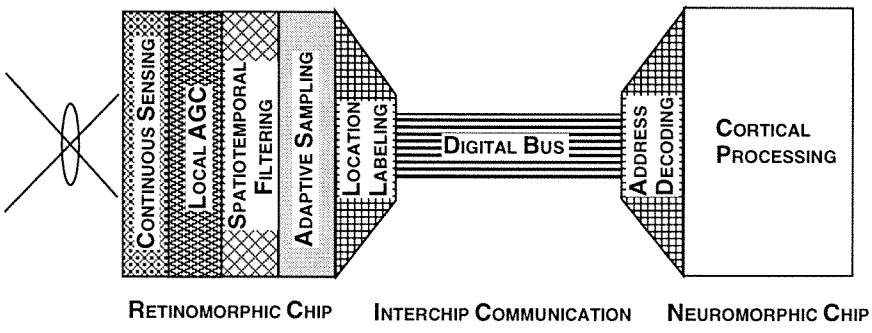
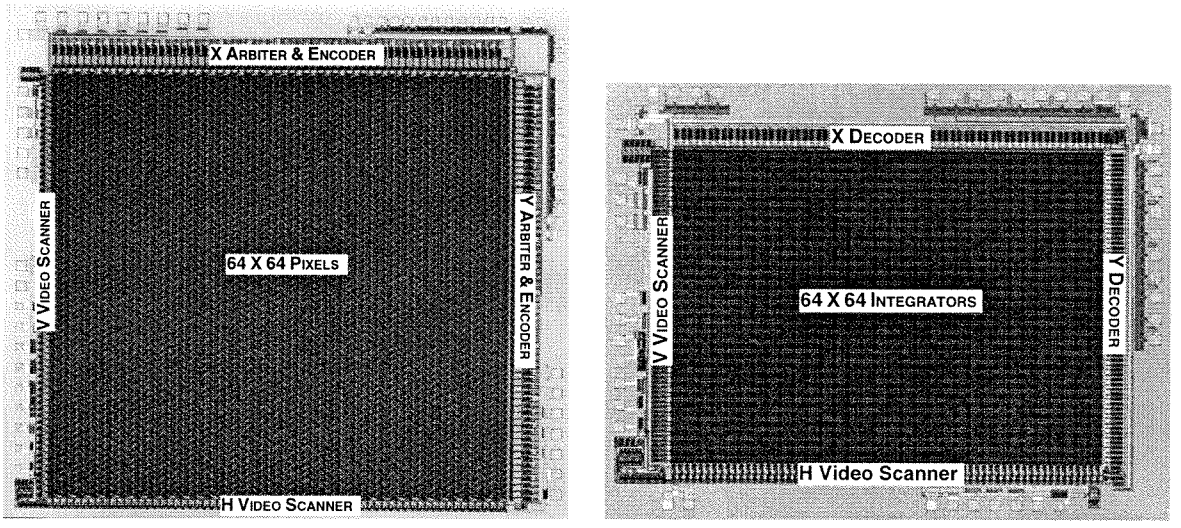


Figure 9.2: RETINOMORPHIC SYSTEM CONCEPT
The retinomorphic chip acquires, amplifies, filters, and quantizes the image. All these operations are performed at the pixel level. The interchip communication channel reads out asynchronous digital pulses from the pixels by transmitting the location of pulses as they occur. A second neuromorphic chip decodes these address events, and recreates the pulses.

systems. My work was inspired by the pioneering work of Mahowald and Mead [142].

In particular, the **retinomorphic approach** uses the system architecture and neurocircuitry of the nervous system as a blueprint for building integrated, low-level, vision systems—systems that are retinomorphic in a *literal* sense. Morphing of biological wetware into silicon-based hardware results in sensory systems that maximize information uptake from the environment, while minimizing redundancy in their output; that achieve high levels of integration, by performing several functions within the same structure; and that offer robust system-level performance, by distributing computation across several pixels.

The retinomorphic system described in this chapter consists of two chips: a focal-plane image processor and a postprocessor with a two-dimensional array of integrators. The system concept is shown in Figure 9.2. Both chips are fully functional; specifications and die photos are shown in Table 9.2 and in Figure 9.3. I describe the retinomorphic pixel design in Section 9.2, and present test results from the complete two-chip neuromorphic system in Section 9.3. The communication channel used to transmit the pulse trains from chip to chip is described in detail a forthcoming paper; brief descriptions aee already available [7, 143]. My concluding remarks are presented



(a) (b)

Figure 9.3: Die Micrographs of Retinomorphic Focal-Plane Processor and Postprocessor
(a) Retinomorphic focal-plane processor. The core of this chip is a 64×64 array of pixels arranged on a hexagonal grid. Pixels generate pulses and communicate the occurence of these pulses by signalling on the column and row request lines. The arbiters ensure that pulses are read out of the array one by one, in an orderly manner, by selecting one pixel at a time with the column and row select lines. The encoders generate the addresses of the selected row and column; this pair of binary words uniquely identifies the location of the pulse. (b) Postprocessor. The core of this chip is a 64×64 array of diode-capacitor integrators. We can feed short current pulses to any integrator in the array by supplying its row and column addresses to the decoders. We use the scanners (shift registers) to read out analog currents from the array for display on a video monitor.

in Section 9.4.

9.2 A Retinomorphic Pixel

I designed the pixel circuit shown in Figure 9.4 using the retinomorphic approach; it senses, amplifies, filters, and quantizes the visual signal. In general terms, this retinomorphic pixel operates as follows.

The transducer is a vertical bipolar transistor; its emitter current is proportional to the incident-light intensity [99]. Two current-spreading networks [5, 132, 129, 128]

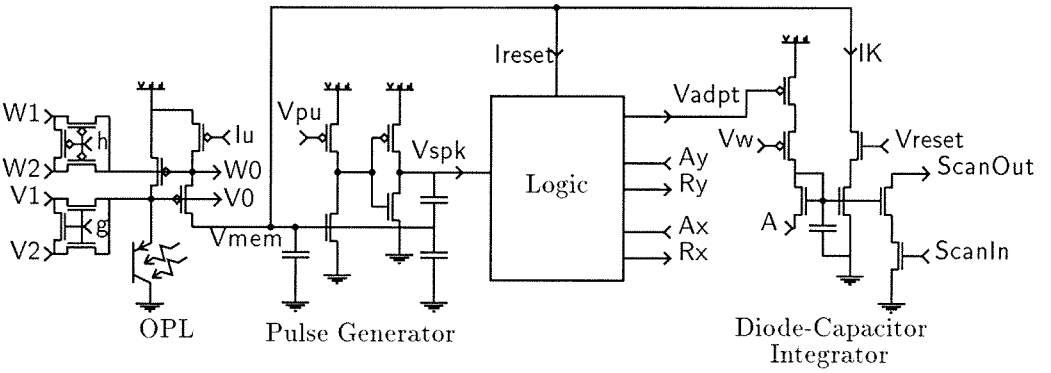


Figure 9.4: RETINOMORPHIC PIXEL

The OPL (outer-plexiform–layer) circuit performs spatiotemporal bandpass filtering and local automatic gain control using two current-spreading networks. It receives input the bipolar phototransistor tied to node V0, which produces a current that is proportional to the light intensity. Nodes V0 and W0 are connected to their six nearest neighbors on a hexagonal grid by the delta-connected transistors, as shown in Figure 9.5. The outer-plexiform–layer circuit's output current is converted to pulse frequency by the pulse generator. The logic circuit communicates the occurrence of a pulse (Vspk) to the chip periphery using the row and column request and select lines (Ry/Ay and Rx/Ax), turns on Ireset to terminate the pulse, and takes Vadapt low, to feed a current pulse to the integrator; the logic circuit is described elsewhere [7, 143]. The integrator's output current (IK) is subtracted from the input to the pulse generator; the device in series with the integrator's output, whose gate is tied to a fixed bias Vreset, is used to isolate the integrator from the rapid volatges swings that occur at Vmem when the membrane voltage is reset after a spike occurs. The two series-connected transistors on the right are used to scan out the integrator's output for display on a video monitor.

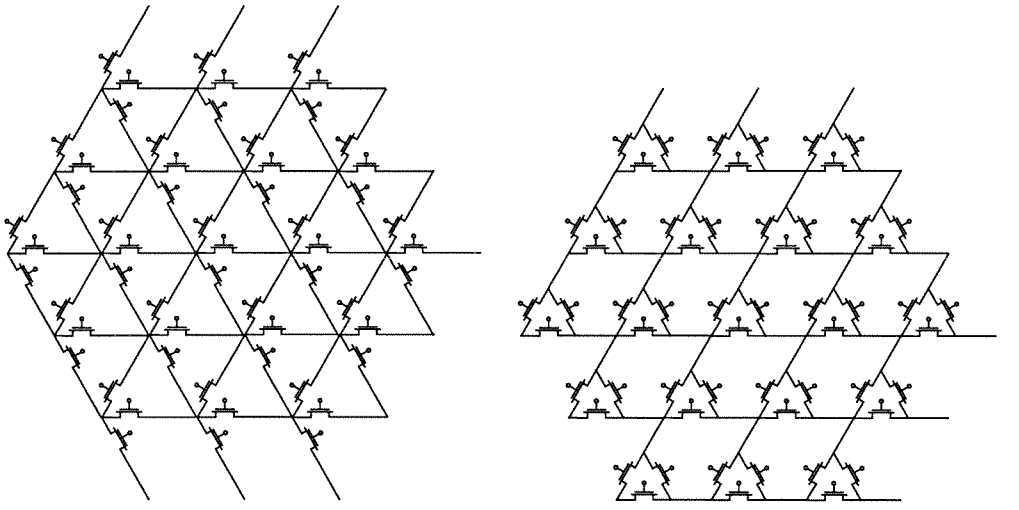


Figure 9.5: Tiling Hexagonal Grids

(a) Star elements. (b) Delta elements. The star-based network requires wires running along three axes, whereas the delta-based network used in the retinomorphic pixel requires wires running along only two axes. Thus, the delta-based hexagonal grid is no more complicated than is the more traditional square grid, and yet it achieves a 33-percent improvement in peak sampling frequency, for pixels of equal area.

diffuse the photocurrent signals over time and space, as described in Chapter 6. The first layer (node V0) excites the second layer (node W0), which reciprocates by inhibiting the first layer. The result is a spatiotemporally bandpass-filtered image [109, 107, 144], as described in Chapters 4 and 7. The second layer computes a measure of the local light intensity, and feeds back this information to the input layer, where the intensity information is used to control light sensitivity. The result is local automatic gain control [5], as described in Chapter 7.

A pulse generator converts current from the excitatory layer into pulse frequency. The diode-capacitor integrator computes a current that is proportional to the short-term average of the pulse frequency; this current is subtracted from the pulse generator's input. The difference becomes larger as the input changes more rapidly, so pulses are fired more frequently. Hence, the more rapidly the input changes, the more rapidly the pulse generator fires. The result is adaptive quantization in time, as described in Chapter 8.

Adding a charge quantum to the integration capacitor produces a multiplicative change in current—due to the exponential current–voltage relation in subthreshold. Hence, the larger the current level, the larger the step size. The result is adaptive quantization in amplitude, as described in Chapter 8. I also use the diode-capacitor integrator in the postprocessor to integrate the pulses, and to reconstruct the current level that was encoded into pulse frequency.

9.3 Overall System Performance

The images shown in Figure 9.6 demonstrate the effects of bandpass filtering and of local automatic gain control. These data are from the OPL (outer plexiform layer) chip described in [5]; images of the same scenes acquired with a CCD camera are included for comparison [100].[1] Bandpass filtering removes gradual changes in

[1]CCD camera specifications: COHU Solid State RS170 Camera (142320), auto iris, gamma factor enabled, 512 × 480 pixels, 8-bit gray-level outputs. Lens specifications: COSMICAR TV lens, ES 50mm, 1:1.8. This comparison, and the face-recognition studies, were done in collaboration with Frank Eeckman, Joachim Buhman, and Martin Lades of the Lawrence Livermore National Laboratories, Livermore, California.

Figure 9.6: CCD Camera Versus Retinomorphic Imager

The CCD camera (top row) has global automatic gain control (AGC), whereas the retinomorphic imager (bottom row) has local AGC and performs bandpass filtering. Three different lighting conditions are shown: Light source on both sides of the subject, to the left, and to the right.

intensity and enhances edges and curved surfaces. It also reduces the variance of the amplitude distribution by mapping uniform areas to the center of the output range (gray level). Local AGC extends the dynamic range by boosting the gain in the dark parts of the scene. Thus, the retinomorphic chip picks up information in the shadows, whereas the output of the CCD camera is zero throughout that region.

Unfortunately, the retinomorphic chip's output is noisier in the darker parts of the image, due to the space constant decreasing with increasing gain. When the space constant decreases, wideband salt-and-pepper noise is no longer attenuated, because the cutoff frequency shifts upward. This noise arises from the poor matching among the small ($4L \times 3.5L$; where L, the minimum feature size, is 2μm) transistors used, which dominates shot noise in the photon flux at the intensity levels that I used. Nevertheless, when the retinomorphic imager replaced the CCD as the front-end of a face-recognition system, the 90- $\times$ 90-pixel OPL chip improved the recognition rate from 72.5 percent to 96.3 percent, with 5 percent false positives, under variable illumination [100].

The output of the postprocessor—after image acquisition, analog preprocessing, quantization, interchip communication, and integration of charge packets in the receiver's diode-capacitor integrators—is shown in Figure 9.7. The sparseness of the output representation is evident.

When the windmill moves, neurons at locations where the intensity is increasing (white region invading black) become active; hence, the leading edges of the white vanes are more prominent. These neurons fire more rapidly as the speed increases because they are driven by the temporal derivative. The time constant of the receiver's diode-capacitor integrator is intentionally set shorter than that of the sender, so temporal integration occurs at only high spike rates. This mismatch attenuates low-frequency information, and results in an overall highpass frequency response that eliminates the fixed-pattern noise and enhances the imager's response to motion. The mean spike rate was 30Hz per pixel, and the two-chip system dissipated 190mW at this spike rate.

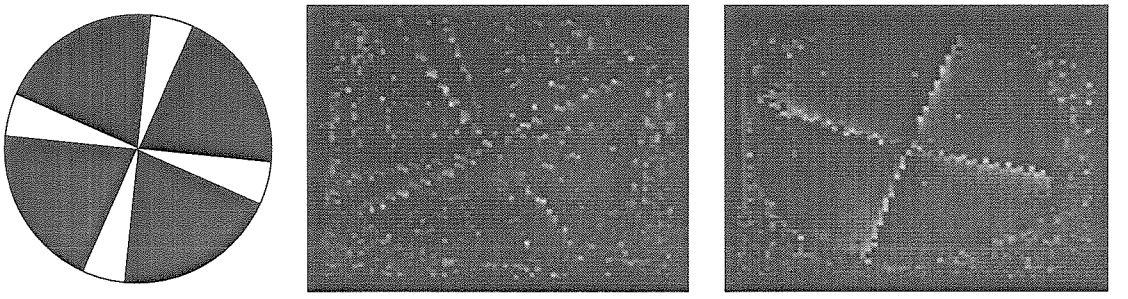


Figure 9.7: VIDEO FROM POSTPROCESSOR CHIP
These video frames were culled from a sequence showing real-time temporal integration of pulses by the postprocessor chip. A rotating windmill pattern (left) was presented to the retinomorphic chip. The response of the postprocessor was captured for slow (middle) and fast (right) rates of counterclockwise rotation.

9.4 Discussion

I have described the performance of a retinomorphic imager. This VLSI chip embodies four principles of retinal operation.

First, the imager adapts its gain locally to extend its input dynamic range without decreasing its sensitivity. The gain is set to be inversely proportional to the local intensity, discounting gradual changes in intensity and producing an output that is proportional to contrast [5]. This adaptation is effective because lighting intensity varies by six decades from high noon to twilight, whereas contrast varies by at most a factor of 20 [6].

Second, the imager bandpass filters the spatiotemporal visual signal to attenuate low-frequency spatial and temporal signals, and to reject wideband noise. The increase in gain with frequency, for frequencies below the peak, matches the $1/f^2$ decrease in power with frequency for natural image spectra, resulting in a flat output power spectrum. This filtering improves information coding efficiency by reducing correlations between neighboring samples in space and time. It also reduces the variance of the output, and makes the distribution of activity sparse.

Third, the imager adapts its sampling rate locally to minimize redundant sampling of low-frequency temporal signals. In the face of limited communication resources

and energy, this sampling-rate adaptation has the additional benefit of freeing up the bandwidth of the communication channel, which is dynamically reallocated to active pixels, allowing higher peak sampling rates and shorter latencies to be achieved [7].

Fourth, the imager adapts its step size locally to trade resolution at high contrast levels, which rarely occur, for resolution at low contrast levels, which are much more common. The proportional step size in the adaptive neuron, which results in a logarithmic transfer function, matches an exponentially decaying amplitude probability density, making all quantization intervals equiprobable. Hence, it maximizes the expected number of signals that can be discriminated, given their probability of occurrence.

For independent samples, information is linearly proportional to bandwidth, and is logarithmically proportional to the signal-to-noise ratio [8]. We increase bandwidth by making the receptors smaller and faster, so that they can sample more frequently in space and time. As an unavoidable consequence, they integrate over a smaller volume of space–time, and therefore the signal-to-noise ratio degrades. There is therefore a reciprocal relationship between bandwidth and noise power (variance) [9]. Since their goal is to maximize information, biological sensory systems aggressively trade off signal-to-noise ratio for bandwidth, operating at ratios close to unity [10, 9].

With this optimization principle in mind, I have proposed compact circuit designs that realize local AGC, bandpass filtering, and adaptive quantization at the pixel level. The overriding design constraints are to whiten the signal, thus making samples independent; to minimize the pixel size, and capacitance, thus making sampling more dense and more rapid; and to minimize power consumption, thus making it possible to acheive very large-scale integration. Hence, all circuits use minimal-area devices and operate in subthreshold, where the transconductance per unit current is maximum.

I realized extremely compact and robust implementations by modeling these circuits closely after their biological counterparts [5, 11], thus reproducing their structure as well as their function.

In earlier chapters, I described and analyzed three limitations faced by these simple circuit designs:

1. Attenuating low frequencies by using a high-gain receptor-to-HC synapse (ratio of $g_{hc}/g_{h0} \equiv 1/\epsilon_h$) results in temporal instability. To break this tradeoff, we must regulate the gain dynamically.

2. Controlling the gain by changing the receptor-to-receptor coupling strength compromises the receptive field size. To decouple these parameters, we must change one of the synaptic strengths (transconductances, g_{ch} or g_{hc}) proportionally.

3. Attenuating the firing rate by using an integrator with a long time constant results in extremely slow adaptation, because we must use a small charge quantum to avoid sending the integrator's output above the input level. To adapt more rapidly, and fire fewer spikes in the process, we must adapt the time-constant.[2]

9.5 Conclusions

Taking inspiration from biology, I have described an approach to building machine-vision systems that perform sophisticated signal processing at the pixel level. These retinomorphic systems are adaptive to their inputs, and thereby maximize their information-gathering capacity and minimize redundant information in their output data stream.

Their optimization principles are radically different from those that drive the design of conventional video cameras. Video cameras are designed to reproduce any arbitrary image to within a certain worst-case error tolerance, whereas biological systems exploit the statistical properties of natural spatiotemporal signals, giving up worst-case performance to get better average-case performance.

Optimizing average-case performance maximizes the discrimination ability of biological systems. Consequently, biomorphic systems promise superior solutions for

[2]The adaptive neuron's membrane time-constant adaptation mechanism was turned off for the system experiments because it made the neurons highly sensitive to transitions on the row and column lines. We must isolate the analog circuits well before we can take advantage of time-constant adaptation.

human-made systems that perform perceptive tasks—such as face recognition and object tracking—energy efficiently.

Specification	Imager	Postprocessor
Technology	2-μm 2-poly 2-metal p-well	
Number of pixels	64×64	
Pixel size (L^2)	53×49	31.5×23
Transistors/pixel	32	8
Die size (mm^2)	8.1×7.4	5.1×4.0
Supply	5 V	
Dissipation (0.2 MS/s)	230 mW (total)	
Throughput	2 MS/s	

Table 9.2: SPECIFICATIONS OF TWO-CHIP RETINOMORPHIC SYSTEM
L is the minimum feature size, which was 2μm for this process; S/s is samples per second.

Bibliography

[1] D. E. Goldman. Potential, impedance, and rectification in membranes. *J. Gen. Physiol.*, 1943.

[2] A. L. Hodgkin and B. Katz. The effect of sodium ions on the electrical activity of the giant axon of the squid. *J. Physiol.*, 1949.

[3] Z. F. Mainen and T. J. Sejnowski. Relaibility of spike generation in neocortical neurons. *Science*, 268:1503–1506, 1995.

[4] M Mahowald. *An Analog VLSI Stereoscopic Vision System.* Kluwer Academic Pub., Boston, MA, 1994.

[5] K Boahen and A Andreou. A contrast-sensitive retina with reciprocal synapses. In J E Moody, editor, *Advances in Neural Information Processing*, volume 4, San Mateo CA, 1991. Morgan Kaufman.

[6] W A Richards. A lightness scale for image intensity. *Appl. Opt.*, 21:2569–2582, 1982.

[7] K. A. Boahen. Retinomorphic vision systems ii: Communication channel design. In *ISCAS 96: IEEE Int. Symp. Circ. & Sys*, volume Supplement, pages 14–17, Piscataway, NJ, May 1996. IEEE Circ. & Sys. Soc., IEEE Press.

[8] C E Shannon and W Weaver. *The Mathematical Theory of Communication.* Univ. Illinois Press, Urbana IL, 1949.

[9] R S Softky. Fine analog coding minimizes information transmission. *Neural Networks*, 8(5), 1995.

[10] R R de Ruyter van Steveninck. The rate of information transfer at graded-potential synapses. *Nature*, 379:642–645, Feb 1996.

[11] Kwabena A Boahen. The adaptive neuron and the diode-capacitor integrator. *In preparation.*

[12] D.A. Baylor, T.D. Lamb, and K.W. Yau. Responses of retinal rods to single photons. *J. Physiol.*, 288:613–634, 1979.

[13] D.A. Baylor, B.J. Nunn, and J.L. Schnapf. The photocurrent, noise, and spectral sensitivity of rods of the monkey *macaca fascicularis*. *J. Physiol.*, 357:575–607, 1984.

[14] J.L. Schnapf and D.A. Baylor. How photoreceptor cells respond to light. *Sci. American.*, 256:40–47, 1987.

[15] R. Shapley and C. Enroth-Cugell. Visual adaptation and retinal gain controls. In N. Osborne and G. Chader, editors, *Progress in Retinal Research, Vol. 3*, pages 263–346. Pergamon Press: Oxford, 1984.

[16] D. Trachina, J. Sneyd, and I.D. Cadenas. Light adaptation in turtle cones: Testing and analysis of a model of phototransduction. *Biophys. J.*, 60:217–237, 1991.

[17] R. Shapley, E. Kaplan, and Purpura P. Contrast sensitivity and light adaptation in photoreceptors or in the retinal network. In R. Shapley and D.M. Lam, editors, *Contrast Sensitivity*, pages 103–116. The MIT Press: Cambridge MA, 1993.

[18] H.B. Barlow. Dark and light adaptation: Psychopysics. In D. Jameson and L.M. Hurvich, editors, *Handbook of Sensory Physiology, Vol. II/4: Visual Psychophysics*, pages 1–28. Springler-Verlag: Berlin, 1972.

[19] H.R. Blackwell. Contrast threshold of the human eye. *J. Opt. Soc. Am.*, 36:624–643, 1946.

[20] P. Sterling, E. Cohen, M. Freed, and R.G Smith. Microcircuitry of the on-beta ganglion cell in daylight, twilight, and starlight. *Neurosci. Res. (Suppl.)*, 6:269–285, 1987.

[21] D. C. Burr. Motion smear. *Nature*, 284:164–165, 1980.

[22] G. Westheimer and S. P. McKee. Perception of temporal order in adjacent visual stimuli. *Vision Res.*, 17:887–892, 1977.

[23] M. Fahle. Figure–ground discrimination from temporal information. *Proc. R. Soc. Lond. B*, 254:199–203, 1993.

[24] C. Enroth-Cugell and J.G. Robson. The contrast sensitivity of the retinal ganglion cells of the cat. *J. Physiol.*, 187:517–552, 1966.

[25] B.B. Boycott and H. Wässle. The morphological types of ganglion cells of the domestic cat's retina. *J. Physiol.*, 240:397–419, 1974.

[26] A.G. Leventhal, R.W. Rodieck, and B. Dreher. Retinal ganglion cell classes in cat and old world monkey: Morphology and central projections. *Science*, 213:1139–1142, 1981.

[27] J.E. Dowling. *The retina: an approachable part of the brain.* Harvard University Press, Cambridge, MA, 1987.

[28] D.M. O'Malley and R.H. Masland. Co-release of acetylcholine and gaba by a retinal neuron. *Invest. Opthalmol. (Suppl)*, 29:273–, 1988.

[29] J.E. Dowling. Retinal neuromodulation: The role of dopamine. *Vis. Neurosci.*, 7:87–97, 1991.

[30] E.C.G.M Hampson, D.I. Vaney, and R. Weiler. Dopaminaergic modulation of gap junction permeability between amacrine cells in mammalian retina. *J. Neurosci.*, 12:4911–4922, 1992.

[31] R.W. Rodieck. The primate retina. *Comp. Primate Biol.*, 4:203–278, 1988.

[32] H. Kolb. Organization of the outer plexiform layer of the primate retina: Electron microscopy of golgi impregnated cells. *Phil. Trans. R. Soc. Lond. (Biol.)*, 258:261–283, 1970.

[33] R. Nelson, A.V. Lützow, H. Kolb, and P. Gouras. Horizontal cells in the cat retina with independent dendritic systems. *Science*, 189:137–139, 1975.

[34] K.W. Stell. Horizontal cell axons and axon terminals in goldfish retina. *J. Compar. Neurol.*, 159:503–520, 1975.

[35] K.W. Stell and D.O. Lightfoot. Color-specific interconnections of cones and horizontal cells in the retina of the goldfish. *J. Compar. Neurol.*, 159:503–520, 1975.

[36] H. Spekreijse and A.L. Norton. The dynamic characteristics of color-coded s-potentials. *J. Gen. Physiol.*, 56:1–15, 1970.

[37] S.R. Cajal. *La retina des vertébrés.* In R.W Rodieck's *The vertebrate retina* [55], 1893.

[38] J.L. Polyak. *The Retina.* Univ. Chicago Press: Chicago, 1941.

[39] H. Kolb, B.B. Boycott, and J.E. Dowling. A second type of midget bipolar in the primate retina. *Phil. Trans. R. Soc. Lond. (Biol.)*, 255:177–184, 1969.

[40] H. Vongersdorff, E. Vardi, G. Matthews, and P. Sterling. Evidence that vesicles on the synaptic ribbon of retinal bipolar neurons can be rapidly released. *Neuron*, 16(6):1221–1227, 1996.

[41] R. Raomirotznik, A.B. Harkins, G. Buchsbaum, and P. Sterling. Mammalian rod terminal — architecture of a binary synapse. *Neuron*, 14(3):561–569, 1995.

[42] E. Raviola and N.B. Gilula. Intramembrane organization of specialized contacts in the outer plexiform of the retina. *J. Cell Biol.*, 65:192–222, 1975.

[43] D.I. Vaney. Patterns of neuronal coupling in the retina. In N. Osbourne and G. Chader, editors, *Progress in Retinal and Eye Research*, volume 13, chapter 12, pages 301–355. Pergamon Press, Oxford, 1994.

[44] S.W. Kuffler. Discharge patterns and functional organization of mammalian retina. *J. Neurophysiol.*, 16:37–68, 1953.

[45] R.W. Rodieck. Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision Res.*, 5:583–601, 1965.

[46] H. Kolb. The morphology of the bipolar cells, amacrine cells and ganglion cells in the retina of the turtle *pseudemys scripta elegans. Phil. Trans. R. Soc. Lond. (Biol.)*, 298:355–393, 1982.

[47] H. Kolb, I. Perlman, and R. A. Normann. Neural organization of the reina of the turtle *mauremys caspica:* a light microscope and golgi study. *Vis. Neurosci.*, 1:47–72, 1988.

[48] H.J. Wagner and E. Wagner. Amacrine cells in the retina of a teleost fish, the roach *(rutilus rutilus)*: A golgi study on differentiation and layering. *Phil. Trans. R. Soc. Lond.(Biol.)*, 321:263–324, 1988.

[49] J. Ammermüller and R. Weller. Correlation between electrophysiological responses and morphological classes of turtle amacrine cells. In R. Weiler and N.N. Osborne, editors, *Neurobiology of the Inner Retina*, volume H31 of *NATO ASI*, pages 117–132. Springler-Verlag, Berling, 1989.

[50] D.I. Vaney. The mosaic of amacrine cells in the mammalian retina. In N.N. Osbourne and G.J. Chader, editors, *Progress in Retinal Research*, volume 9, chapter 2, pages 49–100. Pergamon Press, Oxford, 1990.

[51] B.G. Cleland and W.R. Levick. Brisk and sluggish concentrically organized ganglion cells in the cat's retina. *J. Physiol.*, 240:421–456, 1974.

[52] B.G. Cleland and W.R. Levick. Properties of rarely encountered types of ganglion cells in the cat's retina and an overall classification. *J. Physiol.*, 240:457–492, 1974.

[53] J. Stone and Y. Fukuda. Properties of cat retina ganglion cells: A comparision of w-cells with x- and y-cells. *J. Neurophysiol.*, 37:722–748, 1974.

[54] J.H. Caldwell, N.W. Daw, and H.J. Wyatt. Effects of picrotoxin and strychnine on rabbit retinal ganglion cells: lateral interactions for cells with more complex receptive fields. *J. Physiol.*, 276:277–298, 1978.

[55] R.W. Rodieck. *The vertebrate retina: Principles of structure and function.* W.H. Freeman: San Francisco, 1973.

[56] G.D. Guiloff, J. Jones, and H. Kolb. Organization of the inner plexiform layer of the turtle: An electron microscopic study. *J. Comp. Neurol.*, 272:280–292, 1988.

[57] E.V.Jr. Famiglietti and H. Kolb. Structural basis of ON- and OFF-center responses in retinal ganglion cells. *Science*, 194:193–195, 1976.

[58] R. Nelson, Famiglietti. E.V.Jr., and H. Kolb. Intracellular staining reveals the different levels of stratification for on- and off-center ganglion cells in cat retina. *J. Neurophysiol.*, 41:472–483, 1978.

[59] E.V.Jr. Famiglietti, A. Kaneko, and M. Tachibana. Neuronal architecture of on and off pathways to ganglion cells in the carp retina. *Science*, 198:1267–1269, 1977.

[60] R. Nelson and H. Kolb. Synaptic patterns and response properties of bipolar and ganglion cells in the cat. *Vis. Res.*, 23:1183–1195, 1983.

[61] E.V. Famiglietti and H. Kolb. A bistratified amacrine cell and synaptic circuitry in the inner plexiform layer of the retina. *Brain Res.*, 84:293–300, 1975.

[62] H. Kolb and R. Nelson. Functional neurocircuitry of amacrine cells in the cat retina. In A. Gallego and P. Gouras, editors, *Neurocircuitry of the Retina: A Cajal Memorial*, pages 215–232. Elsevier, New York, 1985.

[63] E. Strettoi, E. Raviola, and R.F. Dacheux. Synaptic connections of the narrow-field, bistratified rod amacrine cell (aii) in the rabbit retina. *J. Comp. Neurol.*, 325:152–168, 1992.

[64] J.E. Dowling and B.B. Boycott. Organization of the primate retina: Electron microscopy. *Pro. R. Soc. Lond. (Biol.)*, 166:80–111, 1966.

[65] J. Ritcher and S. Ullman. A model for the temporal organization of x- and y-type receptive fields in the primate retina. *Biol. Cybern.*, 43:127–145, 1982.

[66] G. Maguire, P. Lukasiewicz, and F. Werblin. Amacrine cell interactions underlying the response to change in the tiger salamander retina. *J. Neurosci.*, 9:726–725, 1989.

[67] M. Kidd. Electron microscopy of the inner plexiform layer of the retina of the cat and the pigeon. *J. Anat.*, 96:179–188, 1962.

[68] R.W. Rodieck, R.K. Brening, and M. Wantanabe. The origin of parallel visual pathways. In R. Shapley and D.M. Lam, editors, *Contrast Sensitivity*, chapter 8, pages 117–144. The MIT Press, Cambridge MA, 1993.

[69] M. Livingstone and D. Hubel. Segregation of form, color, movement and depth: anatomy, physiology, and perception. *Science*, 240:740–749, 1988.

[70] J. Stone and K.P. Hofmann. Very slow-conducting ganglion cells in the cat's retina: a major new functional type? *Brain Res.*, 43:610–616, 1972.

[71] W.R. Levick and L.N. Thibos. Receptive fields of cat ganglion cells: Classification and construction. *Progress in Retinal Reasearch*, 2:267–319, 1983.

[72] H.R. Matturana, J.Y. Lettvin, W.S. McCullock, and W.H. Pitts. Anatomy and physiology of vision in the frog (*rana pipiens*). *J. gen. Physiol.*, 43 (Suppl. 2):129–171, 1960.

[73] R.W. Rodieck and M. Wantanabe. Survey of the morphology of macaque retinal ganglion cells that project to the pretectum, suoerior colliculus, and parvocellular laminae of the lateral geniculate nucleus. *J. Comp. Neurology*, 338:289–303, 1993.

[74] M.A. Freed and P. Sterling. The ON-alpha ganglion cell of the cat and its presynaptic cell types. *J. Neurosci.*, 8:2303–2320, 1988.

[75] E. Cohen and P. Sterling. Parallel circuits from cones to the on-beta ganglion cell. *Euro. J. Neurosci.*, 4:506–520, 1992.

[76] H. Kolb and R. Nelson. OFF-alpha and OFF-beta ganglion cells in the cat retina: II. neural circuitry as revealed by electron microscopy of HRP stains. *J. Comp. Neurology*, 329:85–110, 1993.

[77] A. Hughes and H. Wassle. The cat optic nerve: Fibre count and diameter spectrum. *J. Comp. Neurol.*, 169:171–184, 1976.

[78] V.H. Perry, R. Oehler, and A. Cowey. Retinal ganglion cells that project to the dorsal lateral geniculate nucleus in the macaque monkey. *Neurosci.*, 12:1101–1123, 1984.

[79] Y. Fukuda and J. Stone. Retina distribution and central projections of y- x- and w-cells of the cat's retina. *J. Neurophysiol.*, 37:722–748, 1974.

[80] H. B. Barlow. In W. A. Rosenblith, editor, *Sensory Communication*. MIT Press, 1961.

[81] Joseph Atick and Norman Redlich. What does the retina know about natural scene. *Neural Computation*, 4(2):196–210, 1992.

[82] J H van Hateren. A theory of maximizing sensory information. *Biol. Cybern.*, 68:23–29, 1992.

[83] Dawei Dong and Joseph Atick. Temporal decorrelation - a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6(2):159–178, 1995.

[84] Michael P Eckert and Gershon Buchsbaum. Efficient coding of natural time-varying images in the early visual system. *Phil. Trans. Royal Soc. Lond. Biol*, 339(1290):385–395, 1993.

[85] Dawei Dong and Joseph Atick. Statistics of natural time-varying scenes. *Network: Computation in Neural Systems*, 6(3):345–358, 1995.

[86] B. A. Wandell. *Foundations of Vision.* Sinauer Associates, Sunderland, MA, 1995.

[87] H. de Lange. Research into the dynamic nature of human fovea-cortex systems wityh intermittent and modulated light ii. *J. Opt. Soc. Am.*, 48:784–789, 1958.

[88] D.H. Kelly. Visual responses to time-dependent stimuli, i: Amplitude sensitivity measurements. *J. Opt. Soc. Am.*, 51:422–429, 1961.

[89] J. G. Robson. Spatial and temporal contrast sensitivity functions of the visual system. *J. Opt. Soc. Am.*, 56:583–601, 1966.

[90] D. H. Kelly. Frequency doubling in visual responses. *J. Opt. Soc. Am.*, 56:1628–1633, 1966.

[91] D.H. Kelly. Motion and vision i: Stabilized images of stationary gratings. *J. Opt. Soc. Am.*, 69:1266–1274, 1979.

[92] D.H. Kelly. Motion and vision ii: Stabilized spatio-temporal threshold surface. *J. Opt. Soc. Am.*, 69:1340–1349, 1979.

[93] L.J. Frishman, A.W. Freeman, J.B. Troy, D.E. Schweitzer-Tong, and C. Enroth-Cugell. Spatiotemporal frequency responses of cat retinal ganglion cells. *J. Gen. Physiol.*, 89:599–628, 1987.

[94] C. Enroth-Cugell, J.G. Robson, D.E. Schweitzer-Tong, and A.B. Watson. Spatio-temporal interactions in cat retinal ganglion cells showing linear spatial summation. *J. Physiol.*, 341:279–307, 1983.

[95] C. Enroth-Cugell and A.W. Freeman. The receptive-field spatial structure of cat retinal y cells. *J. Physiol.*, 384:49–79, 1987.

[96] K. Purpura, D. Tranchina, E. Kaplan, and R.M. Shapley. Light adaptation in the primate retna: Analysis of changes in gain and dynamics of monkey retinal ganglion cells. *Vis. Neurosci.*, 4:75–93, 1990.

[97] K Fujikawa, I Hirota, H Mori, T Matsuda, M Sato, Y Takamura, S Kitayama, and J Suzuki. A 1/3 inch 630k-pixel it-ccd image sensor with multi-function capability. In John H. Wuorinen, editor, *Digest of Technical Papers*, volume 38 of *IEEE International Solid-State Circuits Conference*, pages 218–219, San Francisco, CA, 1995.

[98] A Dickinson, B Ackland, E El-Sayed, D Inglis, and E R Fossum. Standard cmos active pixel image sensors for multimedia applications. In William Dally, editor, *Proceedings of the 16th Conference on Advanced Research in VLSI*, pages 214–224, Chapel Hill, North Carolina, 1995. IEEE Press, Los Alamitos CA.

[99] C Mead. A sensitive electronic photoreceptor. In H. Fuchs, editor, *1985 Chapel Hill Conference on VLSI*, pages 463–471, Rockville MD, 1985. Computer Science Press, Inc.

[100] J Buhman, M Lades, and Eeckman F. Illumination-invariant face recognition with a contrast sensitive silicon retina. In J D Cowan, G Tesauro, and J Alspector, editors, *Advances in Neural Information Processing*, volume 6, San Mateo CA, 1994. Morgan Kaufman.

[101] B Sakman and 0 D Creutzfeldt. Scotopic and mesopic light adaptation in the cat's retina. *Pflúgers Archiv fúr die gesamte physiologie*, 313:168–185, 1969.

[102] T Delbruck and C Mead. Photoreceptor circuit with wide dynamic range. In *Proceedings of the International Circuits and Systems Meeting*, IEEE Circuits and Systems Society, London, England, 1994.

[103] D J Field. Relations between statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am.*, 4:2379–2394, 1987.

[104] C Jansson, I Per, C Svensson, and R Forchheimer. An addressable 256 × 256 photodiode image sensor array with 8-bit digital output. *Analog Integr. Circ. & Sig. Proc.*, 4:37–49, 1993.

[105] B Fowler, A E Gamal, and D Yang. A cmos area image sensor with pixel-level a/d conversion. In John H. Wuorinen, editor, *Digest of Technical Papers*, volume 37 of *IEEE International Solid-State Circuits Conference*, pages 226–227, San Francisco, California, 1994.

[106] V Torre, W. G. Owen, and G. Sandini. The dynamics of electrically interacting cells. *IEEE Trans. on Systems Man and Cybernetics*, 13(5):757–765, 1983.

[107] S Ohshima, T Yagi, and Y Funashi. Computational studies on the interaction between red cone and h1 horizontal cell. *Vision Res.*, 35(1):149–160, 1994.

[108] H. A. Beaudot. *The Neural Information Processing in the Vertebrate Retina: A Melting Pot of ideas for Artificial Vision.* Phd thesis, Inst. National Polytechnique de Grenoble, Grenoble, France, 1995.

[109] P.C. Chen and A.W. Freeman. A model for spatiotemporal frequency responses in the x cell pathway of cat's retina. *Vision Res.*, 29:271–291, 1989.

[110] J. J. B. Jack, D. Noble, and R. W. Tsien. *Electric Current Flow in Excitable Membranes.* Clarendon Press, Oxford, England, 2nd edition, 1975.

[111] D.H. Kelly. Theory of flicker and transient responses, i: uniform fields. *J. Opt. Soc. Am.*, 16:534–546, 1971.

[112] D.H. Kelly. Spatial frequency selectivity of the retina. *Vision Res.*, 15:665–672, 1974.

[113] F.W. Campbell and D.G. Green. Optical and retinal factors affecting visual resolution. *J. Physiol. (Lond.)*, 181:576–593, 1965.

[114] R.L. DeValois, H. Morgan, and D.M. Snodderly. Psychophysical studies of monkey vision–iii. spatial luminance contrast sensitivity tests of macaque and human observers. *Vision Res.*, 14:75–81, 1974.

[115] R G Smith. Simulation of an anatomically defined local circuit — the cone-horizontal cell network in cat retina. *Visual Neurosci.*, 12(3):545–561, May-Jun 1995.

[116] W. H. Merigan and J. H. R. Maunsell. How parallel are the primate visual pathways. *Annu. Rev. Neurosci*, 16:369–402, 1993.

[117] H. B. Barlow and W. R. Levick. The mechanism of directionally selective units in rabbit's retina. *J. Physiol.*, 178:477–504, 1965.

[118] B. Hassenstein and W. E. Reichardt. Systemtheoretische analyse der zeit-, reihenfolgen-und vorzeichenauswertung bei der bewegungs-perzeption der russelkafers. *Chlorophanus Z. Naturforsch. B.*, 11:513–524, 1956.

[119] E. H. Adelson and J. R¿ Bergen. Spatiotemporal energy model for the perception of motion. *J. Opt. Soc. Am.*, 2:284–289, 1985.

[120] T. F. Weiss. *Cellular Biophysics.* The MIT Press, Cambridge, MA, 1996.

[121] M. A. Maher. *A Charge-Controlled Model for MOS Transistors.* PhD thesis, California Institute of Technology, Pasadena CA, 1989.

[122] Carver A Mead. *Analog VLSI and Neural Systems.* Addison Wesley, Reading MA, 1989.

[123] Y. P. Tsividis. *The Operation and Modeling of the MOS Transistor.* McGraw-Hill, New York, NY, 1987.

[124] E. A Vittoz. *VLSI Circuits for Telecommunications*, chapter Micropower techniques. Prentice Hall, 1985.

[125] J E Meyer. Mos models and circuit simulation. *RCA Review*, 32:42–63, 1971.

[126] E. A. Vittoz and Fellrath J. Cmos analog integrated circuits based on weak inversion operation. *IEEE J. Solid-State Circ.*, 12:224–231, 1977.

[127] K Boahen. Toward a second generation silicon retina. Technical Report CNS-TR-90-06, California Institute of Technology, Pasadena CA, 1990.

[128] K Bult and G J Geelen. An inherently linear and compact most-only current division technique. *IEEE J. Solid-State Circ.*, 27(12):1730–1735, 1992.

[129] E. A. Vittoz and X. Arreguit. Linear networks based on transistors. *Electronics Letters*, 29:297–299, 1993.

[130] Y. P. Tsividis. Linear, electronically tunable resistor. *Electronics Letters*, 28(25):2303–2305, Dec 1992.

[131] A Andreou and K Boahen. A 48,000 pixel, 590,000 transistor silicon retina in current-mode subthreshold cmos. In *Proc. 37th Midwest Symposium on Circ. and Sys.*, pages 97–102, Lafayette, Louisiana, 1994.

[132] Andreas Andreou and Kwabena Boahen. Translinear circuits in subthreshold mos. *J. Analog Integrated Circ. Sig. Proc.*, 9:141–166, 1996.

[133] M. Tartagni and P. Perona. Computing centroids in current-mode technique. *Electonics Letters*, 29(21):1811–1813, Oct 1993.

[134] M. Kamermans and F. Werblin. Gaba-mediated positive autofeedback loop controls horizontal cell kinetics in the tiger salamander retina. *J. Neurosci.*, 12(7):2451–2463, 1992.

[135] John Lazzaro. Temporal adaptation in a silicon auditory nerve. In D Tourestzky, editor, *Advances in Neural Information Processing 4*, volume 4. Morgan Kaufmannn Pub., 1992.

[136] M Ogata, T Nakamura, K Matsumoto, R Ohta, and R Hyuga. A smart pixel cmd image sensor. *IEEE Trans. Electron Dev.*, 38(5):1005–1010, 1991.

[137] A Mortara, E Vittoz, and P Venier. A communication scheme for analog vlsi perceptive systems. *IEEE Trans. Solid-State Circ.*, 30(6):660–669, 1995.

[138] W B Veldkamp. Wireless focal planes: On the road to amacronic sensors. *IEEE J. Quantum Electronics*, 29(2):801–813, 1993.

[139] B Hoeneisen and C Mead. Fundamental limitations in microelectronics-i: Mos technology. *IEEE J. Solid-State Circ.*, 15:819–829, 1972.

[140] C Mead. Scaling of mos technology to submicrometer feature sizes. *J. VLSI Signal Processing*, 8:9–25, 1994.

[141] K Boahen. Retinomorphic vision systems. In *Microneuro'96: Fifth Int. Conf. Neural Networks and Fuzzy Systems, Laussanne Switzerland*, Los Alamitos, CA, Feb 1996. EPFL/CSEM/IEEE, IEEE Comp. Soc. Press.

[142] M Mahowald and C Mead. The silicon retina. *Scientific American*, 264(5):76–82, 1991.

[143] K Boahen. A retinomorphic vision system. *IEEE Micro Magazine*, 16(5):30–39, Oct 1996.

[144] K Boahen. Spatiotemporal sensitivity of the retina: A physical model. Technical Report CNS-TR-91-06, California Institute of Technology, Pasadena CA, 1991.