Transistors and Synapses: Robust, Low Power Analog Circuits in CMOS Radios and the Rabbit Retina

Alyosha Christopher Molnar

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-60
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-60.html

May 17, 2007

Transistors and Synapses:
Robust, Low Power Analog Circuits in CMOS Radios and the Rabbit Retina

by

Alyosha Christopher Molnar

B.S. (Swarthmore College) 1997
M.S. (University of California, Berkeley) 2003

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering-Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Kristofer S.J. Pister, Co-Chair
Professor Frank Werblin , Co-Chair
Professor Jose Carmena
Professor Frederic Theunissen

Spring 2007

The dissertation of Alyosha Christopher Molnar is approved:

Co-chair_____ Date _____

Co-chair_____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

Spring 2006

Transistors and Synapses: Robust, Low Power Analog Circuits in CMOS Radios and the Rabbit Retina

By Alyosha Christopher Molnar

Abstract

Transistors and Synapses:

Robust, Low Power Analog Circuits in CMOS Radios and the Rabbit Retina

by

Alyosha Christopher Molnar

Doctor of Philosophy in Engineering-Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Kristofer S.J. Pister, Co-Chair

Professor Frank Werblin , Co-Chair

Modern silicon integrated circuits and vertebrate nervous systems are the two of the most compact, complex information processing systems presently known. Within these two classes of system, most information processing is handled in discrete (ie digital) form, but analog components appear in both systems as well. In silicon, wireless communication circuits are typically implemented using analog signal processing. In the central nervous system, the retina plays a similar role of interfacing with the physical world, and also employs analog circuitry. In both cases the primary problem is performing analog operations with low power and high robustness in spite of imperfect circuit components. In this dissertation presents transistor circuits that make up a very low power radio, and neural circuits that have been extracted from direct measurement of retinal neurons in the rabbit retina.

Running at a carrier frequency of 900MHz, the radio described here was shown to communicate up to 100 kilobits per second at ranges of 16 meters or more while consuming 1.3mW in transmit mode and 1.2mW in receive. The whole design only requires 4 external components at a cost of less than 1 dollar. Power reduction was achieved by stacking circuits to make maximum use of battery voltage, and using a single high quality inductor to resonate out capacitance on the inputs of each RF block. The receiver makes extensive use of complementary CMOS circuits for robustly high gain at low power. Simple passive switching mixers were also employed, improving the linearity of the system and permitting demodulation of 1.0 picowatt wanted signals in the face of interfering signals as large as 100μW. Part of this design incorporated a new type of current mirror that with just three additional transistors dramatically reduces the required voltage headroom required to maintain a constant current output by more than a factor of 2 with very little cost in terms of current or die area. The circuits presented here represented a new record in terms of performance at low power.

In the retina, bipolar cells are the primary analog feed-forward cells, but also participate in feedback networks thought to generate diverse signaling pathways. Strikingly, while there are at least 10 morphologically distinct classes of bipolar cell, electrophysiological measurements of these cells only showed 4 distinct types of inhibitory feedback. Also striking was the observation that the OFF cells, which receive increased excitation in response to decreases in light level, and make up fully half of the bipolar cell population, receive only one discernable type of inhibition. This inhibition increased with *increased* light intensity (an ON signal) and so acted to

enhance the OFF response across a wide range of time scales. In contrast, ON bipolar cells (representing the other half of the bipolar cell population) received a variety of different types of inhibition. Some received OFF inhibition which acted to enhance responses in a way similar to the OFF system. Others received inhibition from within the ON system which suppressed low frequency signals but carried an apparent delay causing it to actually enhance the response to high frequency inputs. A third class (identified morphologically as rod bipolar cells) received inhibition that from within the ON system which acted exclusively to suppress responses at all frequencies. Thus, the feedback to bipolar cells is asymmetric between the ON and OFF pathways. Furthermore, these different types of inhibition were pharmacologically distinct, employing different types of inhibitory neurotransmitter.

The most common type of inhibition, which we call "crossover" inhibition, acts between the ON and OFF systems, causing each to suppress the other. This type of inhibition also was evident in amacrine and ganglion cells, such that it makes up the most common form of inhibition in the inner retina. Chapter 5 shows that this class of inhibition acts to suppress even-order (rectifying) nonlinearity, which is introduced by most synapses in the retina. This effect is analogous to the way differential circuits used in radios suppress even order nonlinearity in receiver circuits. Pharmacological manipulation shows that the "linear" outputs of the retina, long considered the "default" pathways (as opposed to special-function nonlinear outputs) are in fact only linear because of the activity of crossover inhibition suppressing their nonlinearity. All of these effects have been verified using both direct patch-clamp recordings and computer simulations. Preliminary analysis also

indicates that the precise asymmetry seen in bipolar and amacrine inhibition yields a circuit topology that is optimally stable and resistant to system-wide shifts in excitability.

This work elucidates the inhibitory circuitry that maintains linearity in the primary visual pathways in mammals, and further demonstrates how the retina maintains its robust functionality in the face of inevitable variability in the components of the system. Thus, as in low power radios, the primary problem to be solved in low power analog circuits in the retina is one of reliability in the face of unreliable physical components.

Acknowledgements

I would like acknowledge the patience and very good advice of my co-advisors Kris Pister and Frank Werblin. Neither of them batted an eye at the idea of an RFIC designer going into retinal neurophysiology, and both have been extremely encouraging through the difficult process of setting aside one field of knowledge and picking up another one. Fellow graduate students in Cory 471: Steve Lanzisera, Brian Leibowitz, Mike Scott, Matt Last, Subbu v and Sarah Bergbreiter all provided good for feedback on talks and papers, as well as lots of other subjects having nothing to do with graduate school. I especially want to acknowledge Ben Cook, with whom I had many insightful conversations about low power IC design. Similarly, I want to thank my fellow graduate students in the Werblin Lab. Shelley Fried and Thomas Muench taught me the basics of both the retina and electrophysiology, and along with Tom Russell and Gareth Spor were involved in many brain storming sessions that helped me to understand my data as well as providing insights into the workings of biologists' minds. I especially want to thank Ann Hsueh for generously sharing her hard-won amacrine cell data and numerous trips to get coffee when the long hours in a dark physiology lab got to be much. Finally I would like to thank my wife, Erika, for putting up with all of the ups and downs of graduate school, and my daughter Vivian, who has only been around for the last year of this process but has already contributed more than her fair share of motivation, sleep loss and deleted figures.

Table of contents

Chapter 1

Introduction to Low Power Radios and Retinas

Introduction

In modern integrated circuits and systems, signal processing is increasingly carried out in the digital domain. Nonetheless, certain aspects of signal processing are still performed by analog circuits before digitization. Analog circuits especially find use in detecting and processing signals derived from the physical environment where signal strength may change by many orders of magnitude, and strong, redundant or unwanted signals are present along with signals of interest. In such situations, analog circuits are typically used to amplify, filter and otherwise extract the wanted signal before digitization. Is this division fundamental, or can we expect an eventual digitization of all circuit functions? Simple calculations seem to indicate that for radio

communications, some analog component of receivers at least can be expected to remain for a long time to come. Observations of the nervous system, which is nature's equivalent of an integrated circuit (albeit a large, wet, three-dimensional one) indicate that this division may be inherent: although most of the central nervous system employs action potentials (spikes) which are discrete in nature and have many of the same benefits of digital encoding of signals, the initial processing of sensory signals is performed primarily in the analog domain. The retina, in particular demonstrates this: light signals (which can vary in intensity by at least 12 orders of magnitude) are initially transduced into analog voltage levels and undergo two rounds of synaptic processing and inhibitory feedback before finally being discretized into a spike train for communication down the optic nerve to the brain.

Digital circuits provide a number of clear benefits over their analog equivalents. Significantly better noise rejection and a lack of DC biasing current means that digital circuits, at least in CMOS, tend to require much less power than their analog equivalents. This benefit becomes especially important since digital circuits consume very little power when not performing computations, while most analog circuits will consume bias current whether or not an input signal is present. Digital circuits are also inherently resistant to imperfections in manufacturing: as long as individual logic gates have sufficient gain to maintain reasonable noise margins, a digital circuit will reliably perform computations, even if its components show large variations in voltage threshold, conductance and leakage. These benefits also apply to spiking in neurons: in order to generate an action potential (spike) a neuron need only achieve positive feedback in its membrane, independent of the exact membrane conductance, ionic

concentration, or morphology of the neuron. Similarly, synaptic transmission is simplified in a spiking cell, where a spike either causes neurotransmitter release, or not, with much less sensitivity to the precise molecular components of the synapse. Indeed, in the nervous system, where voltage levels are comparatively small, and components and connectivity are relatively unreliable, the benefits of discrete coding become even more important.

Given the benefits of digital circuits over analog circuits, one may ask two questions. The first is: when, if ever, does it make sense to use analog circuits to process signals? The second is, how does one build reliable, low power analog circuits from components optimized for digital signal processing? In the following chapters, I will explore these questions in the context of two mixed-signal systems: an ultra-low power 900MHz radio receiver and in the rabbit retina.

Although radio receivers and retinas are very different in some ways (retinas detect light in 2-dimensional array, while a radio detects modulated electromagnetic waves at high frequencies) they share some important requirements in the types of signal processing they are required to perform. Retinas and radios also face similar sorts of physical limitations which arise from the underlying silicon and neural components. As a result of these similarities in signal processing requirements and physical limitations, it is not surprising that these systems have similar architectures.

In chapter 2 I will discuss the design of a very low power radio transceiver, and especially the receiver design and testing (the transmitter and local oscillator chain were mostly described previously in my masters thesis [1] and will only be briefly reviewed here). Chapter three describes a new type of current mirror designed

especially for low power/low supply voltage analog systems. This mirror was used in the transceiver described in chapter 2, but would also be broadly useful in analog circuits where efficient use of voltage headroom is at a premium. Chapter 4 describes work done in the rabbit retina, studying the types of inhibition received by bipolar cells. These cells are non-spiking, "analog" neurons which perform early visual processing in the retina, and receive feedback from other bipolar cells via inhibitory interneurons. Chapter 5 describes a particular class of retinal circuit found throughout the retina, including bipolar cells. This circuit is analogous to differential circuits used in silicon analog electronics and, it will be shown, performs one of the common functions of such circuits, i.e. suppressing even order nonlinearity. Chapter 6 describes a variety of modeling work related to our measurements of retinal activity, and seeks to replicate and explain the function of many of the aspects of retinal circuitry described in previous chapters. In particular, many of the most dominant aspects of retinal circuitry, much like integrated circuitry, function not so much to process signals as to maintain them in spite of imperfections and distortions introduced by the components of neural circuits.

The rest of this introductory chapter will compare the basic requirements (specs if you will) physical limitations and architecture of modern integrated radios and the mammalian retina. Specifically different sections of this chapter are intended to provide a brief overview of these traditionally separate fields (radio design and retinal neurophysiology) for specialists in one, but not the other field.

Signal processing requirements on radios and retinas.

Both radios and retinas take very wide-band analog signals from the physical world and convert them to relatively narrower bandwidth digital signals. In this context, both systems are required to detect signals whose amplitudes vary by many orders of magnitude, such that both noise (for weak signals) and saturation (for strong signals) can present problems. Both systems also need to detect weak, informative signals in the presence of much stronger, uninformative signals

The basic function of a radio receiver is to extract a narrow band signal from the full electromagnetic spectrum, and then demodulate that signal to extract information. Since in modern radios, demodulation generally happens in the digital domain, the primary role of the analog receiver is to reduce the number of bits per second required to code the signal to a rate realizable in an analog to digital converter.

The retina's basic function is to detect light levels at a very large number of distinct points and then process that information so that it can be discretized and carried by the optic nerve without significant loss of relevant information. Thus both systems function to process very wide-band analog inputs and extract a meaningful, much lower bit-rate discrete output.

For radios, the requirement of dealing with a wide variety of signal strengths is usually a consequence of one or both ends of a wireless link being mobile. The two radios may be separated by a variable distance, affecting signal strength. In addition, the signal path is vulnerable to increased channel loss due to intervening obstacles (walls, etc) and to fading from multi-path effects. In particular, the weakest signal that a radio receiver can detect sets its ability to maintain a communication link in the

presence of these various sources of attenuation. Thus, a radio receiver must be designed to detect the weakest signal possible given power and cost constraints, yet it must also be able to accurately demodulate much larger signals as well, implying a wide dynamic range. For a typical cellular radio, this dynamic range covers approximately 9.5 orders of magnitude (of power) [2, 3]. And about 2 orders of magnitude of dynamic fading due to changing channel properties.

In the retina, the equivalent requirement is that imaging be possible under a wide variety of lighting conditions. The extremes go from direct mid-day sun on snow through starlight, or a dynamic range of about 12 orders of magnitude. Lighting level can be vastly different within a single visual scene, between, for example full sun and shade. And lighting can change dramatically in a short time: walking out of a building, or when a cloud crosses in front of the sun. Thus, not just sensitivity, but dynamic range is very important in the retina.

The primary form of unwanted signal that radios must be designed to deal with comes in the form of radio signals at frequencies other than that of the wanted signal. These interferers (blockers) tend to set the instantaneous linearity requirements on a radio receiver, since they can occur at the same time as much weaker wanted signals. If such a signal occurs at the same frequency as one's wanted signal, and is much stronger, the wanted signal is typically lost. However, in most narrow-band systems (which includes most systems today) a given wanted or interfering signal is constrained to a much narrower band than the total bandwidth available, and so it is unlikely that they occupy the same frequency band. It is possible for multiple blockers to interact to disrupt reception of the wanted signal, but the most common scenario is

of a single dominant blocker at a different frequency than a weaker wanted signal. Such a dominant blocker will interact with nonlinearities in the receiver to disrupt the wanted signal. Although the total average power in the 0-1GHz band is typically on the order of 1μW[4], cellular radios are specified to handle close-in (ie other cellular signals) interferers of 10 μW, and to deal with far out interferers as large as 1mW [2]

In the retina, there is no equivalent to interferers at other channels, but there is significant redundancy within the visual signal, which if removed can significantly reduce the effective dynamic range required to pass meaningful information about the visual scene. Furthermore, there are multiple features of the visual scene which can be separated and treated is independent signals, but which are superposed and so may interfere with each other. Thus, as in a radio receiver, multiple, linearly independent signals are present at very different magnitudes.

Physical constraints on radios and retinas

Both modern radios and retinas are built from submicron-scale components, dominated by fixed resistance, capacitance and variable conductance. We will now (very briefly) review the basic components of each system, and compare.

CMOS: a primer

Most modern integrated circuits are built in Complementary Metal-Oxide-Semiconductor processes. Basic connectivity is provided by metal wires patterned on 5 or more distinct layers of metal, connected by vertical vias. These wires are insulated from each other and from a conductive silicon substrate by a matrix of silicon oxide (SiO_2). Metal wires have relatively low resistance, such that when

higher resistance is needed, polycrystalline silicon (poly) is used. The poly forms an additional layer of connectivity, and is also insulated by SiO_2.

Thin layers of SiO_2 can be used to generate parallel-plate capacitance between metal and/or silicon layers, allowing for the design of filters, and for other capabilities such as AC-coupling and sampling of signals. Inductors are sometimes built on chip from spiraling wires, especially for radio circuits where they are used to resonate with capacitance and so to generate filters and increase impedance. However, their large physical size and relatively low quality compared to off-chip inductors, limits their use to high-frequency cases, where they are used only sparingly.

Although resistors and capacitors can accomplish many simple circuit functions, nonlinear components are required for basic functions such as signal sampling and multiplication, and transistors in particular are required to generate power gain.

Passive nonlinearity is available in the form of pn-junction diodes, which respond to a voltage across their terminals with a current[5]:

$$I = I_s e^{Vq/kT} \qquad \text{(eq. 1.1)}$$

Where I_s is a baseline current set by the physical dimensions of the diode and the precise structure of its junction. The term kT/q derives from the Boltzman distribution, where k is Boltzman's constant (1.3×10^{-26} J/K), T is absolute temperature (in Kelvin) and q is the charge of an electron. At room temperature (T = 300K), kT/q = 26mV, and is often referred to as the "thermal voltage" V_T. The current flow in a diode reflects the number of electrons and holes with enough energy to overcome a potential barrier set by the geometry and doping of the diode. As this barrier is

lowered by changing the voltage across the diode, the number of available electrons, and so current flow, increases exponentially.

The most useful component on a silicon chip is the transistor. Metal-Oxide-Semiconductor Field-Effect Transistors (MOSFETs) are the most commonly used transistors today, and are arguably the most common structure built by mankind. A MOSFET consists of a gate electrode, capacitively coupled to a region of silicon, the "channel", between two other electrodes, the drain and source. An appropriate voltage introduced at the gate generates a sheet of charge in the channel, changing the channel's conductance. By imposing a voltage difference between the drain and source, a current is generated between them that depends upon the gate voltage even though no DC current flows into or out of the gate. MOSFETS are essentially symmetric structures, such that the definition of drain and source depends upon the bias state of the device, and upon the type of MOSFET. N-channel MOSFETS (NMOS), which use electrons as their primary charge carriers, have their source at the lower voltage and drain at the higher (see Fig1.1b); P-channel MOSFETS (PMOS), which use holes as their primary charge carriers [6], have their drain at the lower voltage and source at the higher (see Fig1.1b).

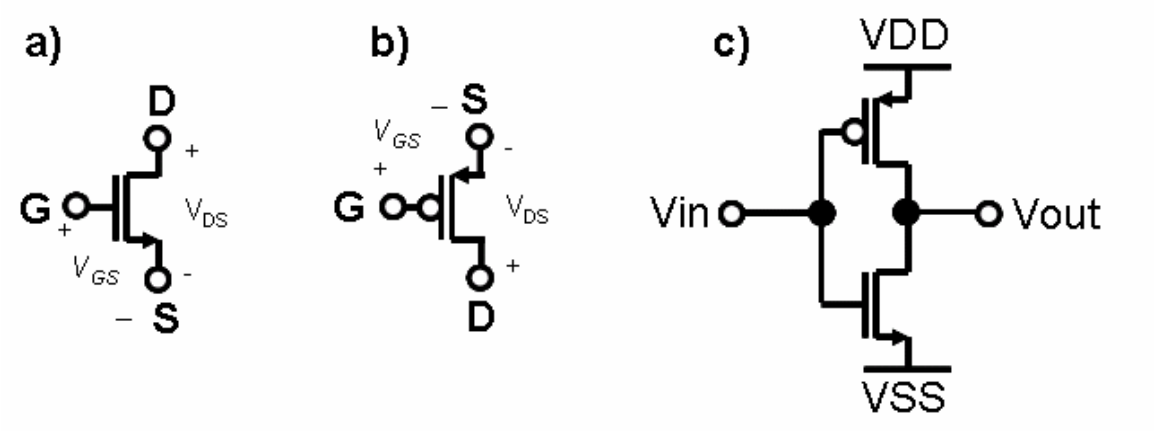


Figure 1.1. Symbols for a) N-channel MOSFETs and b) P-channel MOSFETs, with input voltages V_{GS} and V_{DS} defined, drain current ID is defined as flowing into the drain terminal. c) CMOS inverter constructed from N and P FETs.

The precise relationship between gate voltage, which is typically defined relative to the source (V_{GS}), drain voltage (V_{DS}) and drain current (I_D, which is equal and opposite to source current), depends on the precise operating point of the transistor. Three general states can be defined: saturation, triode and sub-threshold; but note that the transitions between these states are generally smooth.

In saturation, defined as when V_{GS} is greater than a certain threshold voltage (V_{TH}), and V_{DS} is greater than V_{GS}-V_{TH} (this voltage difference is often referred to as V_{DSAT}), I_D can be approximated as:

$$I_D = k\frac{W}{2L}(V_{GS} - V_{TH})^2 \qquad \text{(eq 1.2)}$$

Where k and V_{TH} are set by specific parameters of the manufacturing process, W and L are the physical dimensions of the transistor's gate. Note that in real devices, I_D is

still somewhat dependant upon V_{DS} even in saturation, this relationship is usually linear and approximated by an output resistance r_o.

In triode, V_{GS} is greater than V_{TH}, but V_{DS} is less than V_{DSAT}, in this case, I_D is approximately:

$$I_D = k\frac{W}{L}\left(V_{GS} - V_{TH} - \frac{V_{DS}}{2}\right)V_{DS} \qquad \text{(eq 1.3)}$$

Finally, in subthreshold operation, V_{GS} is less than V_{TH}, and I_D can be approximated as:

$$I_D = I_o\frac{W}{L}e^{V_{GS}\eta/V_T}\left(1 - e^{-V_{DS}\lambda/V_T}\right) \qquad \text{(eq 1.4)}$$

Where η and λ relate to the degree of coupling between gate and channel and drain and channel, and always take values less than one. Note the presence of V_T, the thermal voltage also present in the description of diode behavior, and follows the behavior of a subthreshold device with $V_{GS} = V_{DS}$, and $\eta = \lambda = 1$.

Another class of transistor, the bipolar junction transistor (BJT), approaches the $\eta = \lambda = 1$ limit. No device in a standard silicon process achieves an exponential stronger than the $\eta = 1$ case.

The operating point a given transistor is used in depends strongly upon the function required of it. For simple switching circuits (such as sampling circuits) transistors typically are switched between open (deep subthreshold) and closed (deep triode with V_{DS} close to zero) states. In such operation, the transistor can be approximated as a variable resistor, whose resistance depends upon V_{GS}. For real switching operation, V_{GS} must be driven over a wide enough range that R_{ON} is small enough to permit

sufficient conduction for the application, while R_{OFF} is large enough to prevent leakage. The ratio R_{OFF}/R_{ON} is strictly limited to less than

$$\frac{R_{OFF}}{R_{ON}} > e^{(V_{ON}-V_{OFF})/V_T} \qquad \text{(eq 1.5)}$$

(the case of $\eta = \lambda = 1$) For amplification, either saturation or subthreshold operation are typically used. However, for a given sized device, transconductance ($gm = dI_D/dV_{GS}$) is always greater in saturation than in subthreshold operation. In situations where speed is an issue, greater transconductance is preferable for a given size device, since a given device size corresponds to a roughly constant load capacitance, and transconductance sets the rate at which that capacitance can be charged or discharged:

$$\frac{dV_{out}}{dt} = V_{in}\frac{gm}{C} \qquad \text{(eq 1.6)}$$

Thus, in high speed applications, transistors are typically operated in saturation. However, in low power applications where speed is less important, subthreshold operation makes more sense: if one calculates the ratio of gm to I_D, one finds the maximum to be achieved in the subthreshold domain where

$$\frac{gm}{I_D} = \frac{\eta q}{kT} = \frac{\eta}{V_T} \qquad \text{(eq 1.7)}$$

For simple voltage and power gain, the basic structure in CMOS is a complementary push-pull stage. This structure, shown in figure 1.1c, is often referred to as a "CMOS inverter", because, when driven by a saturating input, it generates a saturated output of the opposite polarity. This simple circuit represents the basic building block of most digital systems. Basic operation of a CMOS inverter is derived from the complimentarity of N-and P-MOSFETs. For an input voltage close to the negative rail

(usually ground) the NMOS enters deep subthreshold operation, while the PMOS is typically in deep triode, pulling the output close to the positive rail (referred to as "VDD"). For an input voltage close to the positive rail the PMOS enters deep subthreshold operation, while the NMOS is in deep triode, pulling the output close to the negative rail. Thus a saturating input close to one rail generates a saturated output close to the other rail. At a mid-level input voltage, both transistors are on and generate equal drain currents. If this current is shifted slightly, one of these currents dominates, for example, if the input voltage is slightly increased, the current through the NFET increases, while the PFET current decreases, this current interacts with the output resistances of the two devices and generates a shift in output voltage significantly larger than the input shift, but in the opposite direction of the input. Thus the circuit has significant voltage gain. This gain is fundamental to digital, switching operation, since it restores saturated binary signals in the presence of input attenuation. If this gain is too low, the difference between "one" and "zero" states becomes difficult to distinguish, leading to ambiguity in the signal being coded. Thus, at low voltage supplies, the capacity to generate gain is the critical requirement for function. CMOS inverters are also useful for generating high gain at low power in analog circuits, and were used extensively in this capacity in the design of the low power radio described in chapter 2.

The maximum DC voltage gain possible in a single stage is limited by the strongest nonlinearity available to a silicon device. An ideal transistor conducts current according to

$$I_s e^{Vin \cdot q/kT}\left(1 - e^{-Vout \cdot q/kT}\right) \qquad \text{(eq 1.8)}$$

corresponding to a bipolar junction transistor or MOSFET in weak inversion with perfect coupling between gate and channel.

Maximum dc gain at low power supply (ie battery) voltages is achieved by using a complementary push-pull stage (ie a CMOS inverter). Using transistors described by equation 1 in such a stage provides a maximum gain of:

$$A_V = \frac{gm_N + gm_P}{g_{DSN} + g_{DSP}} \qquad \text{(eq 1.9)}$$

That is, the transconductance ($gm = dI_D/dV_{GS}$) of the stage divided by its output conductance ($g_{DS} = dI_D/dV_{DS}$), which for the optimal bias point ($V_{GSN} = V_{GSP} = V_{DSN} = V_{DSP} = V_{DC}/2$) gives:

$$gm_N = gm_P = \frac{I_s}{V_T} e^{V_{DC}/2V_T}\left(1 - e^{-V_{DC}/2V_T}\right) \qquad \text{(eq 1.10)}$$

and

$$g_{DSN} = g_{DSP} = \frac{I_s}{V_T} e^{V_{DC}/2V_T}\left(e^{V_{DC}/2V_T}\right) \qquad \text{(eq 1.11)}$$

which combine to give the voltage gain as $Av = 1 - e^{V_{DC}q/2kT}$ [7]

Thus, voltage gain depends exponentially on the ratio of DC supply voltage (V_{DC}) to the "thermal voltage" $V_T = kT/q$ which is approximately 26mV at room temperature. A voltage gain greater than one is required by analog circuits, and by digital circuits in order to maintain the difference between one and zero states in the face of attenuation and noise. Furthermore, a DC supply voltage greater than $V_T \cdot \ln(2) = 35mV$ is required for a voltage gain agnitude greater than one at room temperature. Significantly higher DC voltages are required for more robust gains (for example, $|A_v| > 2$ requires $V_{DC} >$ 55mV).

Real transistors never achieve nonlinearity as strong as that assumed in equation 1.8, such that supply voltages closer to 100mV are typically required for DC gain in CMOS[7]. Furthermore, robust circuits generally require gains significantly greater than 1 to overcome noise and loss. Thus, the lowest supply voltages reported for integrated digital circuits are about 180mV[8], and maximum efficiency (in a power per operation sense) requires supply voltages closer to 400mV. Similarly, useful analog circuits typically operate with supply voltages greater than 400mV[9].

The Boltzman distribution also sets the lower limit on signal voltage for efficient signal processing. For example, rectifying circuits are a fundamental component of demodulators in communication circuits. For simple AM radios, a rectifier is used to detect the signals amplitude, for more advanced radios, the multiplication (mixing) operation (discussed below) is based upon even-order nonlinearity, and thus, upon rectification. At the low voltage limit, a rectifier generates an output proportional to the input voltage squared. The strongest such nonlinearity will be generated by a pair of transistor driven differentially, and generates an output current proportional to:

$$Isig = I_{DC} \cosh\left(Vin \big/ 2V_T \right) \qquad \text{(eq. 1.12)}$$

Which can be approximated, for small signals by its Taylor series:

$$Isig = I_{DC}\left(1 + \frac{Vin^2}{8{V_T}^2}\right) \qquad \text{(eq. 1.13)}$$

For a given input voltage, then, the output (squared) part of the signal will be generated with a current efficiency of

$$\frac{Vin^2}{Vin^2 + 8V_T{}^2}.$$

Thus, to achieve an efficiency of 10% requires a signal level of 26mV (see Fig 1.2). Since in many communication systems, received signals are significantly smaller than 1mV, this limitation on even-order nonlinearity requires either a great deal of gain before demodulating a signal, or that a second, strong local signal be generated to cross-modulated with the received signal. Either approach will inherently consume power.

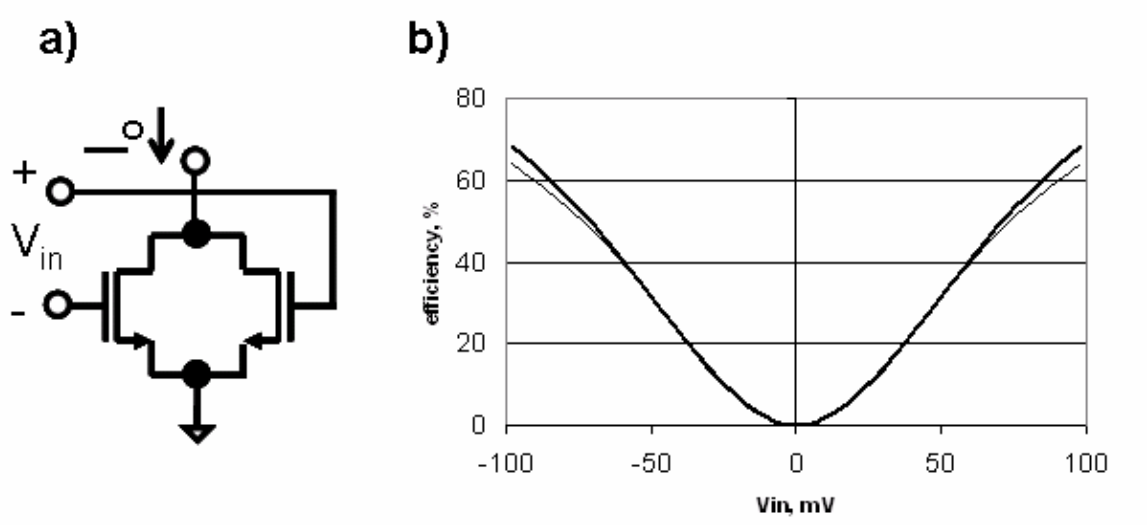


Figure 2.1. a) Simple even order nonlinearity circuit for converting a differential input Vin, to an output current proportional to Vin^2. b) Ideal efficiency curve as a function of Vin.

Basic cellular neuroscience: a primer

The basic unit of any neural circuit is the neuron. Each neuron is a cell complete with nucleus containing DNA, metabolic machinery which breaks down sugar to make ATP, a cell membrane and cytoplasm. Electrical signals are conducted by the cytoplasm, which contains various organic and inorganic ions suspended in aqueous solution. The cytoplasm is insulated from the extracellular solution (which is also conductive) by the cell membrane, which is made up of an oily lipid bilayer which does not pass ions. Thus, the cytoplasm takes the place of wires and the membrane the place of insulation. Since the cytoplasm is resistive and membrane is capacitive, basic RC filtering is to be expected in cellular electrical conduction. Additional, larger resistance is supplied by a class of membrane protein structures called gap junctions which form pores between neurons and so couple them electrically [10].

Active transmission of signals within and between neurons begins with ion gradients. These gradients are maintained by active protein pumps which consume metabolic energy (in the form of ATP) and transport ions across the membrane. The most well understood pump is the sodium-potassium pump, which for each ATP consumed, transports three sodium ions out of the cell and two potassium ions into the cell [10]. As a result, there is much more potassium inside the cell, and much more sodium outside. These ionic gradients have energy associated with them, which in turn can be treated as a voltage associated with each ionic species, set by the Nernst equation:

$$V_{ion} = \frac{kT}{qZ_{ion}} \ln\left(\frac{[ion]_e}{[ion]_i}\right) \text{ [11]} \qquad \text{(eq 1.14)}$$

That is, the voltage associated with a given ion's gradient is set by V_T, the charge of the ion (Z_{ion}) and the natural logarithm of the ratio of extracellular to intracellular concentration.

Ions cannot pass the lipid membrane itself, but may conduct through specific protein pores called ion channels. These channels are structured such that they selectively pass some ions but not others, based upon those ions' charge and size. Thus ion channels provide selective leakage pathways (and so act like resistors) for specific ionic species, and so may be considered to be in series with the specific voltage associated with each ion. A simple electrical model of the neuron is shown in figure 1.3 with representative voltage numbers for each ion (based on solutions used in [12]). One important thing to note is that the selective conductances are variable, and it is through this variability that the actual voltage across the membrane is controlled. Another thing to note is that the three primary ions, potassium, sodium and chloride, span a voltage range of only about 140mV; less than the absolute minimum supply voltage ever reported in a silicon circuit, and more than a factor of 3 less than the minimum efficient supply voltage. This raises the question of how neural circuits can function at so much lower a supply voltage than silicon circuits.

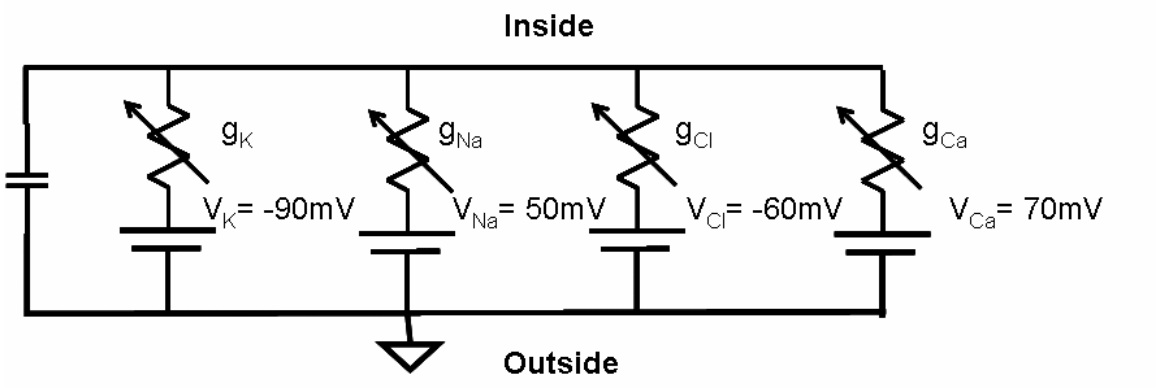


Figure 1.3 Basic electrical model of a neuron. Each battery/conductance pair corresponds to a given ion: K = Potassium, Na = Sodium, Cl = Chloride, Ca = Calcium.

The first thing is to understand how neurons sense their own internal voltage state. This is largely accomplished through voltage gated ion channels. Voltage sensing entails the movement of protein segments across the membrane. This movement is driven by the electric field across the membrane exerting force upon amino acid residues which contain a fixed charge. This movement of segments results in a conformational change in the structure of the channel, opening or closing it to permit or block the flow of ions. The probability of a given channel being open also derives from the Boltzman distribution:

$$p(O) = \frac{1}{1+\exp\left(\frac{V_{th} - nV_{mem}}{kT/q}\right)} \text{[11]} \qquad \text{(eq 1.15)}$$

Where Vth now reflects the relative stability of the open and closed states, and n reflects the number of fixed charges associated with the transition. Thus the conductance associated with a population of identical ion channels is described by:

$$g(V) = \frac{g_o N}{1 + \exp\left(\frac{V_{th} - nV_{mem}}{kT/q}\right)} \qquad \text{(eq 1.16)}$$

Where go is the conductance of a single, open channel, and N is the number of channels involved.

In a semiconductor diode, the Boltzman distribution applies to independent electrons, but in a voltage gated channel, it applies to the channel protein as a whole, such that multiple charges may be involved. As a result, the energy distribution in voltage form involves an exponential with a voltage constant of less than kT/q, typically by a factor of 4 or more, reflecting the number of fundamental charges crossing the membrane[11]. Thus, the underlying nonlinearity is four or more times stronger than in a silicon diode, as shown in Figure 1.4. In addition, because the opening and closing is independent of the voltage associated with the ion involved, effectively negative conductances are possible; that is as voltage increases, less current can flow instead of more. The result is positive feedback, and gain. This kind of gain is especially important for generating and maintaining the discrete impulses (action potentials) that communicate information over long distances in the nervous system.

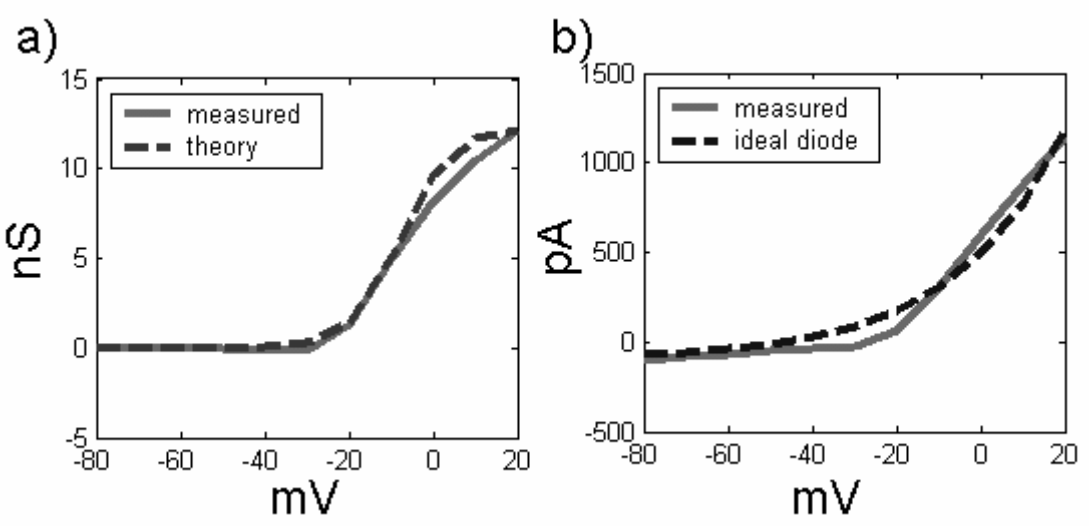


Figure 1.4 a) Measured conductance (solid green) of potassium current in a retinal bipolar cell, and Boltzman distribution for n = 4, amplitude scaled to match (dashed red). b) Measured I-V curve of the same bipolar cell (solid) and ideal semiconductor diode (again scaled). Note the much sharper corner in the membrane curve.

This feedback takes the form of an ion-specific conductance, and so generates a current proportional to that conductance (set by equation 1.16) and the "driving force" associated with the channel. This driving force is set by the difference between the cell's voltage when the ion channels are closed (the "resting potential", V_{rest}) and the reversal potential of the ion(s) that the channel is specific to (V_{ion}). This current interacts with the total conductance of the membrane to modify the cell's voltage according to simple voltage division:

$$V(g) = \frac{V_{ion} - V_{rest}}{1 + g(V)/g_L} \qquad \text{(eq 1.17)}$$

The total feedback, voltage gain is:

$$F = \frac{dV}{dg}\frac{dg}{dV} \qquad \text{(eq 1.18)}$$

Which expands to:

$$F = \frac{(V_{rest} - V_{ion})nqg_o}{kTg_L} \frac{e^{nV/V_T}}{\left(1 + (1 + g_o/g_L)e^{nV/V_T}\right)^2} \qquad \text{(eq 1.19)}$$

Where and g_L is the conductance of the membrane not due to the ion channels (and we have assumed Vth = 0). Solving for the maximum gain possible for a given n and driving force V_{ion}-V_{rest} (equivalent to battery voltage in electronics) gives:

$$F = \frac{(V_{rest} - V_{ion})nq}{2kT} \qquad \text{(eq 1.20)}$$

Which for n = 4 implies that a gain greater than one may be achieved for driving voltages as low as 13mV. Actual neurons typically use much larger driving voltages close to 140mV [11] to generate strong positive feedback when generating action potentials. It is interesting to note that the ratio between ideal minimum driving voltage and actual driving voltage (140mV/13mV~11) is about the same as that for digital transistor circuits (400mV/35mV~12).

Voltage gated ion channels are also a critical part of active synapses. Here voltage gated calcium channels allow an influx of calcium ions according to the voltage of a presynaptic (input) cell. Thus increase in calcium concentration triggers a biochemical cascade causing vesicles (small bubbles of membrane separate from the cell membrane) to merge with the membrane and release their contents. In a synapse, the release happens on a portion of the membrane close to an adjacent (post synaptic) cell and the vesicles release neurotransmitter, which is typically a small molecule similar to an amino acid. The post synaptic membrane contains receptors sensitive to

that neurotransmitter, as shown in Figure 1.5. Typically these receptors are either ion channels in their own right, and so open in the presence of neurotransmitter, or are coupled, via a second-messenger cascade, to ion channels which may either open or close in the presence of neurotransmitter.

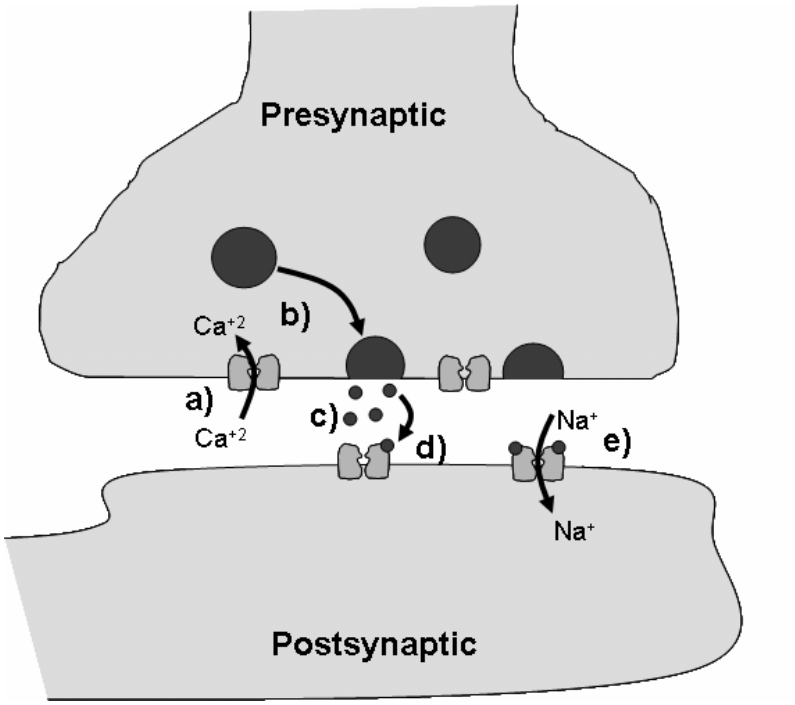


Figure 1.5. Basic chemical synaptic transmission. a) Increased presynaptic voltage opens voltage-gated calcium channels allowing calcium to enter the presynaptic cell. b) Calcium causes vesicles to fuse with cell membrane. c) Neurotransmitter inside the vesicle is released, diffusing across the synaptic cleft. d) Neurotransmitter binds to postsynaptic receptors causing (e) them to open and pass ions into the post synaptic cell.

Typically it has been found that rate of vesicle release is proportional to calcium concentration [11], and so may be thought of as proportional to the number of calcium channels open, following equation 1.16. Reception follows the Hill equation:

$$R([NT]) = \frac{[NT]^m}{1+[NT]^m} \quad \text{(eq 1.21)}$$

With a Hill coefficient, m, of either 1 or 2 (indicating the number of molecules that must bind to trigger opening). Thus, the postsynaptic conductance now follows a similar curve to that of a voltage gated ion channel, concatenated with the Hill equation:

$$R(V_{pre}) = \frac{\lambda^m}{\left(1+\exp\left(\frac{-nV_{pre}}{kT/q}\right)\right)+\lambda^m} \quad \text{(eq 1.22)}$$

where λ is a proportionality constant that relates the probability of calcium channel opening to neurotransmitter release. This in turn can be concatenated with voltage division, such as in equation 1.17 to yield the voltage transfer across the synapse:

$$V_{post} = \frac{V_{ion} - V_{rest}}{1 + Ng_o R(V_{pre})/g_L} \quad \text{(eq 1.23)}$$

Once again we can calculate the maximum voltage gain, as a function of Vion-Vrest:

$$A_V = \frac{(V_{ion} - V_{rest})nmq}{2kT} \quad \text{(eq 1.24)}$$

Thus synapses are capable of generating gain for very small driving forces, potentially below 10mV.

As a result, of the concatenation of cooperative binding and the calcium curve, synapses regularly perform meaningful (nonlinear) computations on signals of less than 10mV [13]. Typical driving forces for post-synaptic ion channels are on the order of 60mV (based on [12]).

The increased nonlinearity of ion channels in biological membranes suggests that there are aspects of neuronal processing that simply cannot be replicated as efficiently in

bulk silicon circuits. It is likely that molecular electronics, a field presently in its infancy, can develop devices that replicate the cooperative properties of ion channels and synapses and thereby permit processing of smaller voltage signals that function using less power than is possible in bulk semiconductor circuits.

Synapses have one additional property, which is that they act to shape responses in time, that is, they have memory. The transition of receptors between closed and open states can be modeled by a simple 2-state model, where binding of neurotransmitter drives receptors from a closed state to an open state, and open channels revert to closed with a constant probability. This model can be well described by a linear time-varying differential equation:

$$\dot{x} = k_1 u(t)(1 - x) - k_2 x \qquad \text{(eq 1.25)}$$

Where x is the proportion of open channels, u(t) is the input (proportion of receptors that have bound the required number of neurotransmitter molecules) and k_1 and k_2 are rate constants. While this model is sufficient for some types of receptor, it fails to account for the basic dynamics of many receptors. These receptors typically show a transient response even to constant input levels. This can be modeled by including a third, inactivated state, such that receptors channels that are open have a finite probability of entering this third state, as shown in figure 1.6. Inactivated receptors may be modeled as resetting either by returning to an open state or directly to a closed state (indeed, both paths must be accounted for in a complete model). This three-state model yields a second-order LTV system of differential equations:

$$\begin{aligned} \dot{x} &= k_1 u(t)(1 - x - y) - k_2 x - k_3 x + k_4 x \\ \dot{y} &= k_3 x - k_4 y - k_5 y \end{aligned} \qquad \text{(eq 1.26)}$$

Where y represents the proportion of receptors in an inactivated state, and k3, k4, and k5 represent the rate constants for transitions: **open→inactive**, **inactive→open** and **inactive→closed**, respectively.

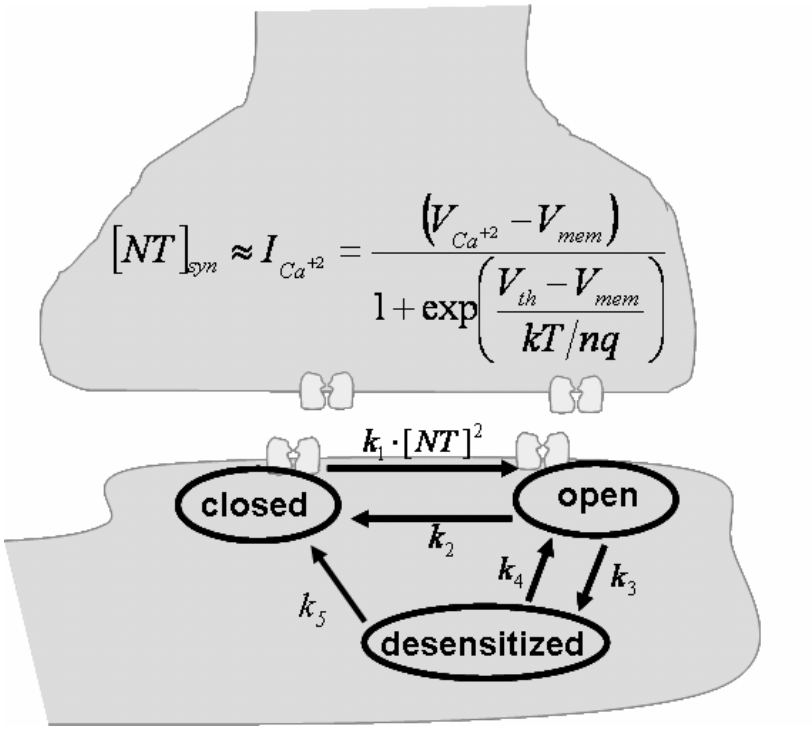


Figure 1.6. Simple model of a chemical synapse: Neurotransmitter concentration is based upon a static nonlinearity, which taken to a power (set by the degree of cooperativity in binding) provides the time-varying coefficient in a 3 state (2-state variable) linear time-varying kinetic model. Note that in steady state, this kinetic model follows the Hill equation.

Thus, a synapse may be reasonably be modeled by a static nonlinearity composed of the calcium channel curve combined with post-synaptic cooperative binding, followed by a 1st or 2nd order LTV system. Some of the implications of this model will be discussed in chapter 6.

Comparing neurons with CMOS, we can roughly treat cytoplasm as equivalent to metal and poly wires, membranes as equivalent to silicon oxide, ion channels as equivalent to diodes (though in some cases the analogy may be better to resonant-tunneling diodes, or other inherently unstable two-terminal devices), and synapses as equivalent to transistors, incorporating some filtering. This last analogy is especially attractive, as synapses, like CMOS transistors, come in two general forms: excitatory synapses (typically mediated by the neurotransmitter glutmate) which open generalized anion channels whose reversal potential is typically about 0mV, and Inhibitory synapses (typically mediated by one of the neurotransmitters glycine or GABA), which open chloride channels whose reversal potential is typically in the range of -60 to -40mV (though higher reversal potentials have been reported in some cell types). These two synapse types can be thought of as corresponding to P- and N-MOSFETs respectively, as shown in Fig 1.7.

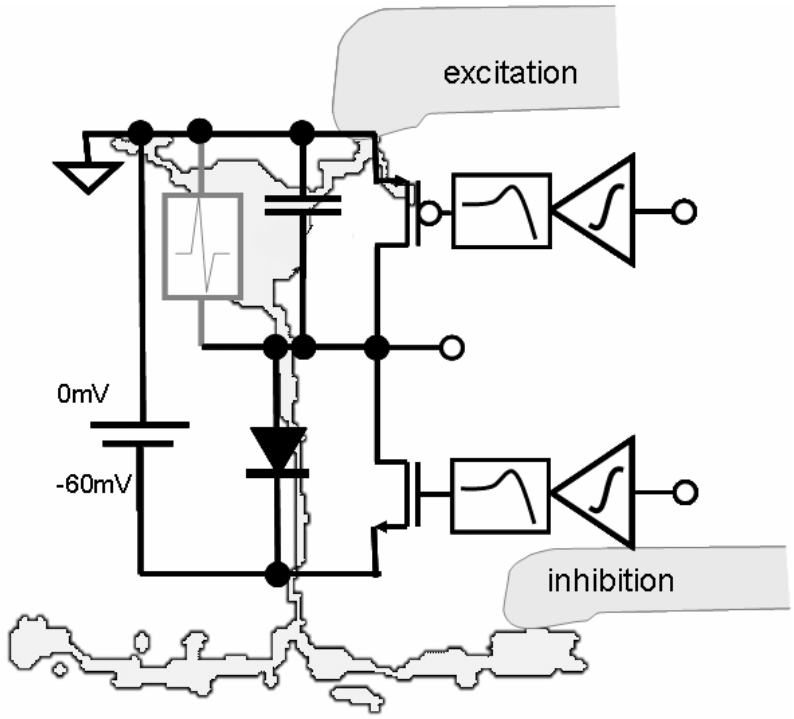


Figure 1.7 Rough circuit analogy for inputs to a neuron: Excitatory inputs act to pull up intracellular voltage, similar to a PFET, inhibitory inputs act to pull down the intracellular voltage, similar to an NFET. Membrane conductances act like a diode ore one-shot.

Architecture comparison: radio receivers vs retinas.

Although in some ways very different, the basic architecture of a radio receiver and that of the retina share some basic features. These architectures will now be dsicussed, and then contrasted and compared.

In both systems, signals are initially transduced and amplified. This is followed by filtering, further amplification (usually variable amplification) and further, more precise filtering, followed by discretization (digitization).

Typical radio receiver architecture:

In radios, the initial transduction of high-frequency signals typically takes the form of a low noise amplifier (LNA) operating at the frequency of the received radio-frequency (RF) signal, followed by a frequency translator block. This frequency translation usually takes the form of a multiplier (mixer) whose other input is driven by a local oscillator (LO) signal at or near the carrier frequency of the wanted signal. These mixers work based upon the trigonometric identity:

$$A(t)\cos(\omega_{RF}t + \varphi(t)) \times \cos(\omega_{LO}t) =$$

$$= \tfrac{1}{2}A(t)\cos((\omega_{RF} - \omega_{LO})t + \varphi(t)) + \tfrac{1}{2}A(t)\cos((\omega_{RF} + \omega_{LO})t + \varphi(t)) \quad \text{(eq. 1.27).}$$

In a receiver, the mixer is usually followed by a low pass filter that acts to suppress the $A(t)\cos((\omega_{RF} + \omega_{LO})t + \varphi(t))$ term, so that the output is $A(t)\cos((\omega_{RF} - \omega_{LO})t + \varphi(t))$. The amplitude and phase of the RF input are preserved, but the carrier frequency is now $|\omega_{RF} - \omega_{LO}| = \omega_{BB}$ instead of ω_{RF}. Note that the absolute value means mixers will down-convert both $\omega_{RF} \pm \omega_{BB}$. That is, both the wanted RF signal and its "image". The image is typically rejected by multiplying the RF signal by both $\cos(\omega_{LO}t)$ and $\sin(\omega_{LO}t)$, and then comparing the relative phase of the outputs. Very simple radios sometimes replace the mixer with a simple rectifier followed by low pass filtering, and use the fact that

$$[A(t)\cos(\omega_{RF}t + \varphi(t))]^2 = \tfrac{1}{2}A(t)^2 + \tfrac{1}{2}A(t)^2\cos(2\omega_{RF}t + 2\varphi(t)) \quad \text{(eq. 1.28)}$$

to detect amplitude modulation (the $A(t)^2$ term)

The RF "front end" generally must be low enough noise to detect the weakest wanted signal possible, while being sufficiently linear that much stronger signals at other frequencies do not cause the front end to saturate. Although the problem of unwanted

signals is usually reduced by a "roofing filter" in front of the LNA, such filters can usually be relied on for only about 20dB of suppression, and usually cannot be frequency tuned, so that in a multi-channel system, the other channels cannot be suppressed, since they may become the wanted channel at any time. In many systems, the front end gain is made programmable, so that it can be reduced in the presence of strong signals, improving linearity at the expense of noise.

Another feature of the front ends of many modern radios is that they take as input the single-ended signal available from most antennas and generate an output that is differential[3, 14]. Thus, the low frequency output of the front end is usually coded by two voltages, whose difference encodes the output signal. This differentiality provides a variety of benefits, including rejection of common mode signals, much better even order nonlinearity, easier bias stabilization, and an effective doubling of signal magnitude. Where single ended signals are recoded as differential depends on the design, and may precede the LNA[9], or be between the LNA and mixer[3, 14],or may be inherent to the structure of either LNA or mixer.

After this initial amplification and down-conversion, the wanted signal is usually either a baseband signal, or at a low intermediate frequency. Baseband signals are typically considered to have a center frequency of 0 Hz, and therefore need to be coded by two parallel signals ("in phase" and "quadrature", referred to as the "I" and "Q" signals) to capture both amplitude and phase information:

$$I = A(t)\cos(\varphi(t))$$

$$Q = A(t)\sin(\varphi(t)) \qquad \text{(eq. 1.29)}$$

Low IF systems do not require I and Q paths, since the intermediate frequency is typically chosen to keep all modulation frequencies positive. However, this provides no image rejection, and so, most low-IF systems still code both I and Q, and then recombine the signals to cancel the opposite (image) sideband.

In both direct conversion and low-IF architectures, the down-conversion is typically followed by a low-pass filter, which suppresses interfering signals that have also been down-converted, but are sufficiently offset from the wanted signal as to fall outside the wanted band. This filtering reduces the linearity requirements of the system, and so is usually followed by more programmable gain. In a direct conversion system, there are usually several interleaved stages of gain and low-pass filtering before digitization. In a low-IF system, there must also be another round of down conversion to baseband, typically after the second gain stage. This is then followed by more filtering and gain, and finally digitization.

The analog portion of a radio receiver is basically just a succession of stages that each reduce the required bandwidth and dynamic range required to maintain signal fidelity, and ease the requirements on succeeding stages. This is accomplished by a combination of frequency translation, which reduces the basic frequency of operation, filtering, which reduces the effective bandwidth of operation, and variable gain which can amplify weak signals relative to noise, or attenuate strong signals to prevent saturation. This process is performed by a series of stages, each of which typically consumes an order of magnitude less power than the previous stage.

Architecture of the retina

The retina's "front end" is its photoreceptors, the rods and cones. Basic transduction is carried out by the photo pigment rhodopsin, which changes configuration when it absorbs a photon of light, and triggers a biochemical cascade leading to current flow across the photoreceptor membrane. This transduction is set up such that increases in light intensity lead to a hyperpolarization of the membrane; that is, the intracellular voltage become more negative. This process involves several stages, each of which amplifies the signal, and each of which can be suppressed by slow biochemical negative feedback[15]. This feedback suppresses baseline responses to sustained light levels (a sort of DC-offset function) and simultaneously reduces transduction gain for incremental signals. This correlation of offset and gain control makes sense in the case of light signals, since most parts of the visual world can be modeled as the reflectance of a surface, which takes values between 0 to 1, multiplied by ambient illumination. Thus, offset, which reflects that illumination, also reflects the amplitude of variations that can be expected. A photoreceptor incorporates the functions of transduction, gain, gain control and filtering in a single cell, and it is the photoreceptor that is responsible for most of the gain and gain control in the retina.

Morphologically, photoreceptors make up the "outer retina" with their photo-sensitive outer segments at the back of the retina (furthest from the cornea), their cell bodies making up the outer nuclear layer, and their outputs forming synapses in the outer plexiform layer (see figure 1.8)

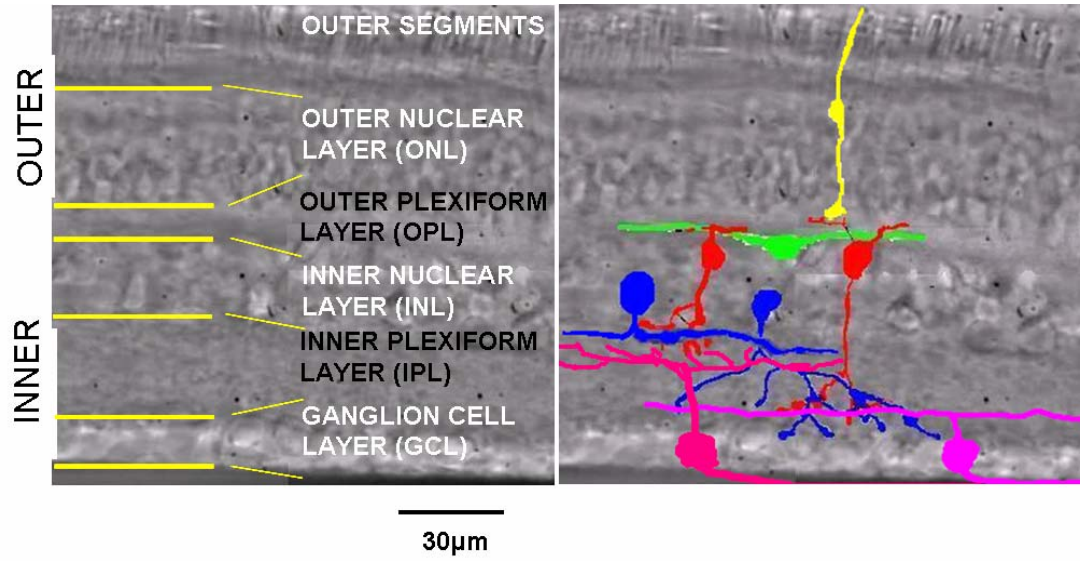


Figure 1.8. Basic retinal structure: layers and their names are shown on the left, typical cell morphologies are shown on the right. Light is detected by the outer segments of photoreceptors (yellow) whose cell bodies reside in the outer nuclear layer. The resulting electrical signal is carried to the outer plexiform layer, where synapses are made onto horizontal cells (green) and OFF and ON bipolar cells (red; OFF on the left, ON on the right). Bipolar cell bodies reside in the outer portion of the inner nuclear layer. Bipolar cells project axons into the inner plexiform layer, where they form synapses onto inhibitory amacrine cells (blue) and spiking ganglion cells, which carry information down the optic nerve to the brain.

Photoreceptors form excitatory synapses onto, and receive inhibitory feedback from a class of neurons called horizontal cells. These cells integrate inputs over a wide area of retina, and are electrically coupled to one another, such that they are well modeled by a network of resistively coupled nodes. Horizontal cells effectively low-pass filter cone signals in space and time, and then negatively feed back to cones,

generating an overall band-pass response in the cone's release. This operation can be thought of as a whitening of the visual input in both space and time. Natural scenes tend to contain a lot of autocorrelation (discussed in chapter 6), such that low-frequency terms dominate the spectrum of intensity in both space and time[16]. Thus, by band-pass filtering the signal, redundant signals are suppressed, and the effective required dynamic range in the cone synapse is reduced.

Cones also provide excitatory synaptic input to bipolar cells. This input is in the form of glutamate binding to receptors on the dendrites of the bipolar cells. Bipolar cells express multiple types of glutamate receptors, which fall into two general categories: ionotropic receptors, such as AMPA and kainate receptors, and a metabotropic glutamate receptor know as mGluR6. Ionotropic receptors open in the presence of glutamate, preserving the sign of the cone response and so depolarize their bipolar cells for *decrements* in light intensity. Bipolar cells which show this response are called OFF bipolar cells, since they depolarize when light turns off. mGluR6 receptors act on membrane voltage indirectly by triggering a second messenger cascade inside the cell which leads to the closing of ion channels when glutamate in present in the synapse. Thus, these cells will act to invert the sign of the cone response, and so will depolarize the cell in response to *increments* in light intensity. These cells are known as ON bipolar cells.

There are generally thought to be approximately 10 distinct bipolar cell types, divided roughly equally between ON and OFF types [17]. These cells' bodies inhabit the outer half of the inner nuclear layer, and extend axons to the inner plexiform layer (IPL). Interestingly, OFF bipolar cells' ramify exclusively in the outer half of the IPL,

while ON bipolar cells ramify in the inner half of the IPL. Both types of cells make glutamaturgic, ionotropic (sign preserving) synapses onto amacrine and ganglion cells. Amacrine cells are almost exclusively inhibitory, releasing glycine and GABA[18-27], and forming synapses onto bipolar, ganglion and other amacrine cells. There are at least 30 distinct morphologies of amacrine cell [28], including relatively narrow, diffusely stratified cells and wider field, monostratified cells.

Ganglion cells, meanwhile, form the output of the retina, with their axons making up the optic nerve. There are at least 14 distinct types of ganglion cell [29], split roughly into ON and OFF types, as well as at least two types of ON-OFF cells[30]. Ganglion cells' dendrites generally ramify narrowly in a thin substratum of the IPL, reading out from a subset of bipolar cells. Thus ON Ganglion cells co-stratify with ON bipolar cells in the ON sublamina (roughly the inner 60%) of the IPL, while OFF cells stratify in the OFF sublamina (the outer 40%), and the ON-OFF cells stratify in both.

Ganglion cells' dendrites spread over a wider area of the retina than bipolar cells, and, as a result, ganglion cells respond to stimuli over a wider region of the retina. Because different types of ganglion cell have different dendritic spreads, their outputs code for different sizes of stimuli. In addition, different types of ganglion cell have different temporal responses, with some responding to steps in light with very transient responses, and others responding in a much more sustained, sluggish fashion. Thus, different types of ganglion cell code for different aspects of the visual scene[12] [30]. For simple ON and OFF cells, this recoding can be thought of, roughly as a bank of band-pass filters in space and time; Fig 1.9 shows step responses in space and time for 5 types of OFF ganglion cell. Putative morphology for each type is also shown,

showing that larger dendritic spread tends to correspond to more blurred representation of edges. Also shown is an extraction of the rough region of space-time frequency space occupied by each cell type, based upon their step responses (this will be discussed in greater detail in chapter 6). It can be seen that these responses are only partially redundant, and cover a range of frequencies fairly completely. By coding for only a subset of the total visual input, ganglion cells reduce the total bandwidth each has to handle, and so reduce the dynamic range requirements on each pathway. This reduction is especially important since each output is separately discretized into a spike train, and so reduced in its data handling capability.

Thus, one might conceive of the mammalian retina as taking visual input, gain-controlling it in response to basic intensity levels, whitening the resultant signal in space and time, and then dividing the signal into 5 or so sub-bands before discretization. Several additional phenomenon should also be noted however. The first is that many of the outputs are not linear, making a frequency-space description inadequate (and leaving the question of what kind of general description *would* be adequate). This nonlinearity is most obvious in the ON-OFF cells which receive inputs from both ON and OFF bipolar cells. These cells show a full-wave rectified response to simple light stimuli. A large subset of these cells, called ON-OFF directionally selective ("DS") cells actually code for directional movement in one of 4 cardinal directions, and are largely indifferent to the brightness structure of the thing moving[31]. Indeed, all of the ganglion cells in the retina show some rectification, in that their peak firing rates are much larger than ½ their baseline firing rates. However, while many cells can be well modeled by a linear filter followed by a simple half-

wave rectification (these are generally referred to as being linear), other cells are not well modeled this way, indicating nonlinearities present presynaptic to the final spiking mechanism, this will be discussed in greater detail in chapter 5.

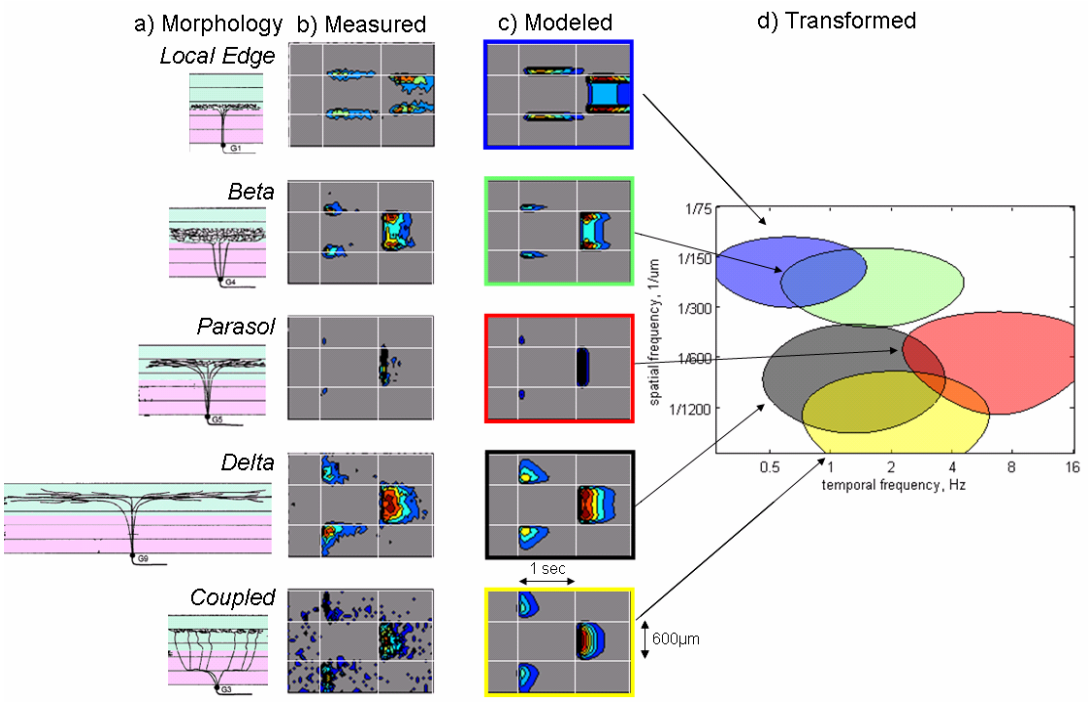


Figure 1.9 Example OFF ganglion cell responses. a) Sketches of morphology of 5 types of OFF ganglion cell (from[29]) b) Spiking responses to a step in space and time (generated by flashing a single 600µm square at different offsets, data from [30]). c) modeled results assuming an impulse response described by a difference of Gaussians in space, and a difference of exponentials in time. d) The Fourier transform of these impulse functions in space and time.

Other nonlinear phenomenon seen in many cells include contrast gain control and saccadic suppression. Both of these phenomenon act to reduce the dynamic range requirements on the output of the retina. Contrast gain control refers to the

observation that rapidly changing signals with high variance have their gain reduced, and low variance signals see their gain increased [32, 33]. This change in gain makes sense in the context of changing lighting conditions where, for example, direct sunlight leads to high contrast between light and shadow, whereas indirect light leads to much lower contrast. Both scenarios must be handled by the same output cells; contrast gain control reduces this dynamic range mismatch. Saccadic suppression takes the form of a very transient ON-OFF inhibition to many ganglion cells, and occurs when a sufficiently large area of the retina sees a synchronized change in brightness[34]. This inhibition is thought to suppress retinal responses when the eye changes orientation rapidly, as it does during a saccade. This suppression makes sense since saccades are driven by the organism itself, and do not reflect meaningful changes in the visual world, yet because a saccade will generate massive changes in the local inputs to the retina, it would be expected to generate a massive retinal response.

The basic array of band-pass responses in ON and OFF systems could be explained in the context simple feedforward pathways involving an initial difference of Gaussians, modeling the cone-horizontal cell interaction, followed by variable synaptic and spiking dynamics and different sized ganglion cell dendritic fields. However, the more complex interactions underlying saccadic suppression and directional selectivity require inputs from amacrine cells to ganglion cells[31, 35]. Amacrine input can also play a role in shaping and refining the basic spatio-temporal filtering underlying the different ganglion cell outputs. Thus, for example, wide-field monostartified amacrine cells can suppress wide-field (low spatial frequency)

responses. To really understand the signal processing of the retina, one must also understand the effects of amacrine inputs to bipolar and ganglion cells, and indeed to other amacrine cells. It is these interactions that will be the focus of chapters 4, 5 and 6.

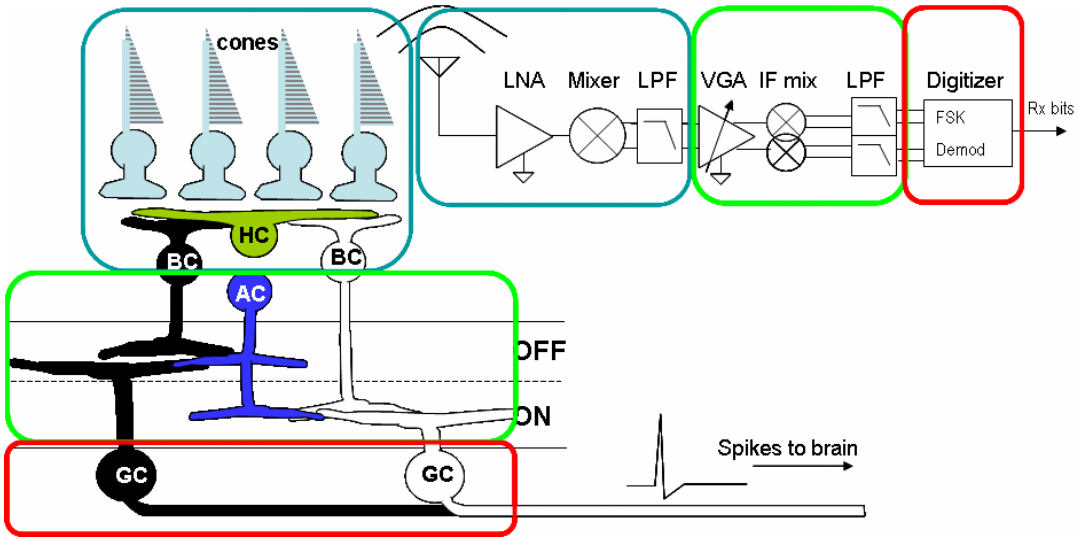


Figure 1.10. Comparison of typical radio receiver and retina. Blue is the front end/outer retina, handling transduction, low-noise gain and initial filtering, and conversion from single ended to differential signaling (back=OFF, white=ON). Green is intermediate frequency analog processing/ inner retinal processing including more variable gain, division into multiple pathways, and more filtering. Red is discretization into spikes in the retina, into bits in the radio.

Comparing the architecture of these two analog systems, the vertebrate retina, and radio receivers, reveals some important similarities, illustrated in figure 1.10: first both start with gain and transduction to signals at frequency bands more easily handled by low power circuits. This front end tends to be the most power-hungry component

of each system. This is followed by a 1st, general filter. The outputs of this first stage are coded as two parallel complementary signals, and then further gained up (variably) and filtered. The signals are then further diversified and filtered into still narrower bands before finally being discretized.

References

[1] A. Molnar, "An Ultra-Low-Power 900MHz Radio Transmitter For Wireless Sensor Networks," UC Berkeley, Berkeley Dec 2003.

[2] ETSI, "Digital cellular telecommunications system (Phase 2+): Radio transmission and reception (GSM 05.05 version 5.4.1)," 2 ed, E. 910, Ed., 1997.

[3] R. Magoon, A. Molnar, G. Hatcher, J. Zachan, and W. Rhee, "A single-chip quad-band (850/900/1800/1900MHz) direct-conversion GSM/GPRS RF transceiver with integrated VCOs and Fractional-N synthesizer," *IEEE JOURNAL OF SOLID-STATE CIRCUITS,* vol. 37, pp. 1710-1720, December 2002.

[4] J. Rabaey: Personal communication, 2003.

[5] P. Grey and R. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3 ed. New York: Wiley & Sons, 1993.

[6] S. M. Sze, *Physics of Semiconductor Devices*, 2 ed. Dehli: Wiley & Sons, 1981.

[7] J. D. Meindl, "Low Power Microelectronics: Retrospect and Prospect," *PROCEEDINGS OF THE IEEE,* vol. 83, pp. 619-635, 1995.

[8] A. Wang and A. Chandrakasan, "A 180-mV Subthreshold FFT Processor Using a Minimum Energy Design Methodology," *IEEE JSSC,* vol. 40, pp. 310-319, 2005.

[9] B. Cook, A. Berne, A. Molnar, S. Lanzisera, and K. Pister, "Low Power 2.4GHz Transceiver with Passive RX front-end and 400mV supply," *IEEE JSSC,* vol. 41, pp. 2757-2766, Dec 2006.

[10] M. J. Zigmond, F. E. Bloom, S. C. Landis, J. L. Roberts, and L. R. Squire, *Fundamentral Neuroscience*: Academic Press, 1999.

[11] B. Hille, *Ion Channels of Excitable Membranes*, 3rd ed. Sunderland Massachusettes: Sinauer Associates, Inc., 2001.

[12] B. Roska and F. Werblin, "Vertical interactions across ten parallel, stacked representations in the mammalian retina," *Nature,* vol. 410, pp. 583-7, Mar 29 2001.

[13] F. S. Werblin and J. E. Dowling, "Organization of the retina of the mudpuppy, Necturus maculosus. II. Intracellular recording," *J Neurophysiol,* vol. 32, pp. 339-55, May 1969.

[14] A. Molnar, R. Magoon, G. Hatcher, J. Zachan, W. Rhee, M. Damgaard, W. Domino, and N. Vakilian, "A single-chip quad-band (850/900/1800/ 1900MHz) direct-conversion GSM/GPRS RF transceiver with integrated

VCOs and Fractional-N synthesizer," in *ISSCC*, San Francisco, 2002, pp. 232, 233.

[15] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walker, *Molecular Biology of The Cell*, 4 ed. New York: Garland Science, 2002.

[16] A. v. d. Schaaf and J. H. v. Hateren, "Modelling the Power Spectra of Natural Images: Statistics and Information," *vision research,* vol. 36, pp. 2759-2770,, 1996.

[17] M. A. MacNeil, J. K. Heussy, R. F. Dacheux, E. Raviola, and R. H. Masland, "The population of bipolar cells in the rabbit retina," *J Comp Neurol,* vol. 472, pp. 73-86, Apr 19 2004.

[18] S. A. Bloomfield and D. Xin, "A comparison of receptive-field and tracer-coupling size of amacrine and ganglion cells in the rabbit retina," *Vis Neurosci,* vol. 14, pp. 1153-65, Nov-Dec 1997.

[19] J. Bolz, P. Thier, T. Voigt, and H. Wassle, "Action and localization of glycine and taurine in the cat retina," *J Physiol,* vol. 362, pp. 395-413, May 1985.

[20] R. Boos, H. Schneider, and H. Wassle, "Voltage- and transmitter-gated currents of all-amacrine cells in a slice preparation of the rat retina," *J Neurosci,* vol. 13, pp. 2874-88, Jul 1993.

[21] M. A. Freed, Y. Nakamura, and P. Sterling, "Four types of amacrine in the cat retina that accumulate GABA," *J Comp Neurol,* vol. 219, pp. 295-304, Sep 20 1983.

[22] U. Grunert and H. Wassle, "Immunocytochemical localization of glycine receptors in the mammalian retina," *J Comp Neurol,* vol. 335, pp. 523-37, Sep 22 1993.

[23] U. Greferath, F. Muller, H. Wassle, B. Shivers, and P. Seeburg, "Localization of GABAA receptors in the rat retina," *Vis Neurosci,* vol. 10, pp. 551-61, May-Jun 1993.

[24] J. Jager and H. Wassle, "Localization of glycine uptake and receptors in the cat retina," *Neurosci Lett,* vol. 75, pp. 147-51, Mar 31 1987.

[25] P. D. Lukasiewicz and F. S. Werblin, "The spatial distribution of excitatory and inhibitory inputs to ganglion cell dendrites in the tiger salamander retina," *J Neurosci,* vol. 10, pp. 210-21, Jan 1990.

[26] F. Muller, H. Wassle, and T. Voigt, "Pharmacological modulation of the rod pathway in the cat retina," *J Neurophysiol,* vol. 59, pp. 1657-72, Jun 1988.

[27] H. Wassle, I. Schafer-Trenkler, and T. Voigt, "Analysis of a glycinergic inhibitory pathway in the cat retina," *J Neurosci,* vol. 6, pp. 594-604, Feb 1986.

[28] M. A. MacNeil, J. K. Heussy, R. F. Dacheux, E. Raviola, and R. H. Masland, "The shapes and numbers of amacrine cells: matching of photofilled with Golgi-stained cells in the rabbit retina and comparison with other mammalian species," *J Comp Neurol,* vol. 413, pp. 305-26, Oct 18 1999.

[29] R. L. Rockhill, F. J. Daly, M. A. MacNeil, S. P. Brown, and R. H. Masland, "The diversity of ganglion cells in a mammalian retina," *J Neurosci,* vol. 22, pp. 3831-43, May 1 2002.

[30] B. Roska, A. Molnar, and F. Werblin, "Parallel Processing in Retinal Ganglion Cells: How Integration of Space-Time Patterns of Excitation and Inhibition

Form the Spiking Output," *Journal of Neurophysiology,* vol. 95, pp. 3810-22, June 2006.

[31] S. I. Fried, T. A. Munch, and F. S. Werblin, "Directional selectivity is formed at multiple levels by laterally offset inhibition in the rabbit retina," *Neuron,* vol. 46, pp. 117-27, Apr 7 2005.

[32] K. A. Zaghloul, K. Boahen, and J. B. Demb, "Contrast adaptation in subthreshold and spiking responses of mammalian Y-type retinal ganglion cells," *J Neurosci,* vol. 25, pp. 860-8, Jan 26 2005.

[33] S. M. Smirnakis, M. J. Berry, D. K. Warland, W. Bialek, and M. Meister, "Adaptation of retinal processing to image contrast and spatial scale," *Nature,* vol. 386, pp. 69-73, Mar 6 1997.

[34] B. Roska and F. Werblin, "Rapid global shifts in natural scenes block spiking in specific ganglion cell types," *Nat Neurosci,* vol. 6, pp. 600-8, Jun 2003.

[35] S. I. Fried, T. A. Munch, and F. S. Werblin, "Mechanisms and circuitry underlying directional selectivity in the retina," *Nature,* vol. 420, pp. 411-4, Nov 28 2002.

Chapter 2

A 900MHz Radio Transceiver for "Smart Dust"

Smart dust overview

Wireless sensor networks are a rapidly growing area of research with applications in environmental monitoring, inventory tracking, and other areas. Such networks are made up of small wireless nodes, each consisting of a suite of sensors, some custom analog and digital circuitry [1], a microprocessor [2], a power supply, and a means of communication. Although other methods have been suggested, the primary method of communication is through low-power radio[3-6].

Sensor networks are generally conceived of as being peer-to-peer networks wherein data gathered by one node is transmitted to the end user via a series of hops between nodes. Hence, the range required for nominal radio operation indoors, assuming one

node per room, needs to be about 10 meters through at least one wall. The nodes are expected to be deployed for years, powered by a single battery or scavenged ambient energy. In either case, a mean power consumption of less than 10 μW is required [4]. Given a reasonably low duty cycle on the order of 1%, radio power consumption must be limited to approximately 1 mW.

This chapter will discuss the design of a radio that met this goal operating in the 900MHZ ISM band (902MHz-928MHz). It will start by describing the constraints on this design and the overall approach used to meet them. This is followed by a brief discussion of the system level design of the smart dust transceiver, including the architecture used. The context of the receiver design is provided by a brief review of the transmitter design previously described in [7]. Finally, there is an in-depth discussion of the design of the FSK receiver that was integrated with the transmitter, and made use of the same RF oscillator and frequency dividers as used by that transmitter.

System Design

Ultimately, requirements on power consumption are set by cost. If a network is deployed for a period that is longer than the lifetime of a given node, then this relationship is clear: power consumption sets the rate at which nodes (or at least their batteries) must be replaced, and so the cost of maintaining the network. Thus a given external component is worth including if its inclusion decreases power consumption sufficiently to reduce the rate of replacement below the cost of the component. Based on this trade-off, a single off-chip inductor was justified by its effective reduction in

Oscillator power (discussed below), and that a 32kHz off-chip crystal could be justified by the resulting reduction in duty cycle and bandwidth provided by precise timing and center frequency. Beyond these components, an antenna and a battery are required. Battery selection was made based on cost: Lithium coin cells provide the highest energy density at low per-unit cost, however, these batteries have the drawback of providing a relatively high DC voltage of approximately 3 Volts. The challenge of this design, then was to achieve high efficiency using only this one inductor, one slow clock and a 3 volt supply voltage.

This system was intended to be able to communicate ~100kb/s at a range of ~10m while consuming minimum power. Given that even with an off-chip inductor, the oscillator will consume ~400μW in this process (discussed below), and that antenna drivers are difficult to make better than 50% efficient without multiple inductors, the output power of a 1mW transmitter can be expected to be less than ~300μW. Using an empirical model [8] of indoor signal propagation, link margin, the attenuation between receiver and transmitter will be on the order of:

$$\frac{P_{sense}}{P_{rad}} = \left(\frac{\lambda}{4\pi}\right)^2 \frac{1}{r^{\rho}} \qquad \text{(eq. 2.1)}$$

Where λ is wavelength, r is range in meters, and $2<\rho<4$ is a coefficient modeling signal attenuation in an indoor environment. For the above frequency, range, transmit power and a $\rho = 3.5$ this implies a receiver sensitivity of about 10pW, or -80dBm (power in dBm = 10 log(P/1mW)). In addition, the receiver needed to be insensitive to out-of-band interference from other sources. Exactly how strong a signal must be tolerated is somewhat arbitrary, depending upon the how much interference one needs to be able to handle. This receiver was designed to function with a cellular handset

transmitter in the same room as the receiver, which at 850MHz with a 20dBm output power at 3 meters distant, and assuming a worst case, $\rho = 2$ from eq 2.1, implies a signal of about -22dBm of interference at tens of MHz offset.

General Architecture

The transmitter was very simple, comprising an RF oscillator with completely digital frequency centering and modulation driving a PA with controllable output power. This implementation provides somewhat imprecise modulation, but avoids overhead power required by up-conversion mixers, phase splitters, and other aspects of higher performance, higher power implementations. To minimize power consumption in the receiver, the number of circuits operating at RF frequencies had to be minimized. Thus a low-IF receiver was used, with an IF of 1.7MHz, and power was minimized by using a single-phase RF mixer without image rejection (discussed in greater detail below).

In both of the transmitter and receiver described above, overhead power is dominated by the RF oscillator[9]. Minimizing oscillator power consumption requires maximizing tank impedance, and so inductance and Q, implying either a physically large on-chip inductor or an off-chip inductor. For this design, an off chip inductor provided approximately a factor of 30 improvement over on-chip implementation and so was the clear solution. The oscillator's high Q LC tank was then used to directly drive all other RF circuits, eliminating the need for buffers and reducing current drive requirements by a factor of Q.

The necessary oscillator swing for this design is only about 1.2V peak to peak. This could result in poor efficiency, since the oscillator requires significantly less voltage headroom than the 3 volt battery supplies. This extra headroom was used by stacking the RF oscillator with the other RF and analog blocks and reusing its bias current. Combining stacking with the architecture decisions above results in the block diagram shown in figure 2.1. In both receive and transmit modes, current flows through the oscillator and is then redistributed to the various RF and analog blocks. At the same time, the oscillator core is AC-coupled to the inputs of these same blocks, tuning out their capacitance.

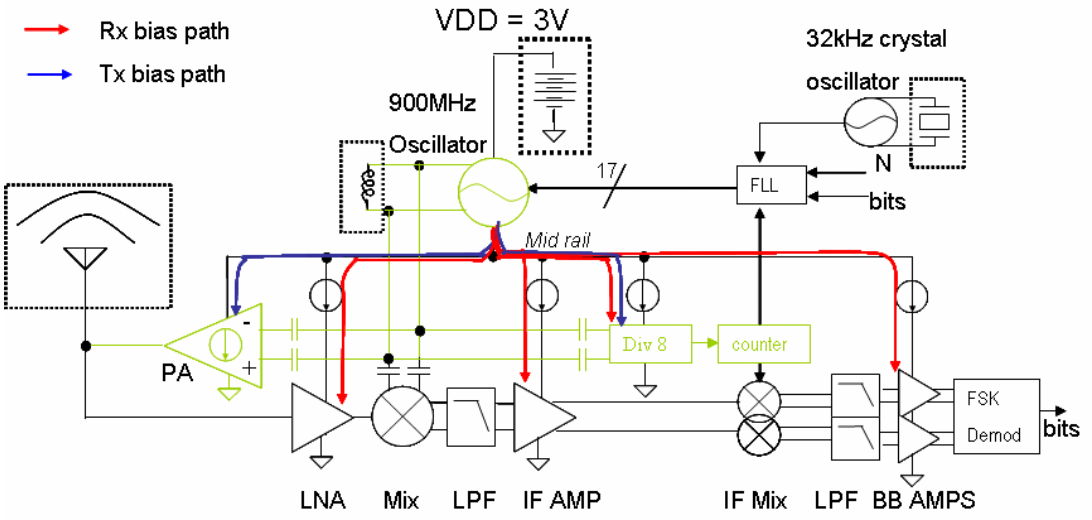


Figure 2.1. Block diagram of low power transceiver. Off chip components are outlined in dashed lines. The 900MHz oscillator is stacked with and distributes its bias current to other analog and RF circuits, and AC-couples its RF output into their inputs. Bias current paths in transmit mode are shown in blue, in receive mode in red. Greened-out blocks were previously discussed in [7].

Local Oscillator

The most critical specification for the oscillator is that loop-gain be greater than one; otherwise the oscillation will not occur. Loop gain is set by the product of tank impedance (equal to ωLQ) and feed-back transconductance, which is set by dc bias current. Thus, to minimize bias current, tank impedance must be maximized. Cost constraints require that it be only one off-chip inductor, and power constraints require maximum gm for a given bias current. These combined requirements favor the use of cross-coupled CMOS inverters to provide active feedback, as shown in figure 2.2. This structure increases gm without extra current by combining the gm of the N- and P-FETs, while providing a natural bias path with only one inductor. For a 15nH inductor in series with two 1.5nH bondwires, and a Q of about 30, the active parts of the circuit needed to provide about 400μS of conductance. To ensure reliable operation, the oscillator was overdriven by a factor of about 2, implying a current of about 400μA. Once up and running, the oscillator hard-switches this current across the resonant tank, giving a swing of about 1.2V, sufficient to hard-switch the transistors of the PA, mixer and RF divider. The oscillator also required approximately 1.2V of dc Voltage bias when oscillating.

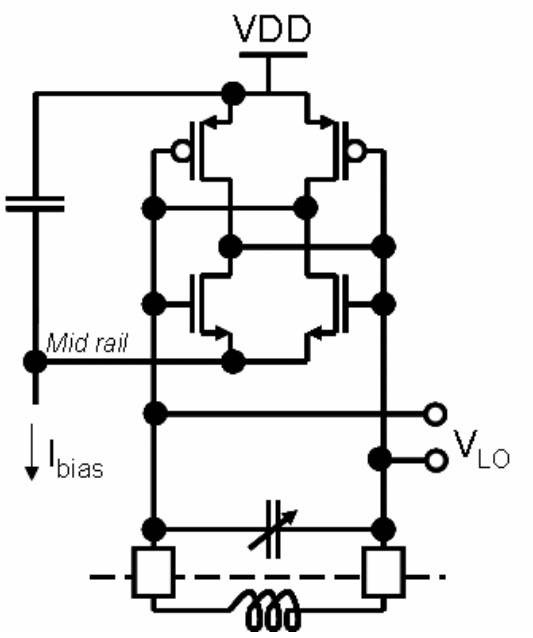


Figure 2.2. 900MHz oscillator. Resonance is provided by a variable capacitor and off-chip inductor. Energy dissipated by losses in the resonant tank are restored by a pair of cross-coupled CMOS inverters, whose lower rail forms the Mid rail for the system. This mid rail is coupled to VDD through a large capacitor which provides an effective AC ground.

In order to be as flexible as possible when integrated with a microprocessor and RAM, the oscillator's frequency tuning was made fully digital. A tuning range of 16% was needed to cover the US ISM band as well as process and component variations. A frequency accuracy of better than 5kHz was chosen to ensure accuracy much better than the symbol spacing for FSK signals carrying 100kbs. Tuning was implemented with 1 pF of switchable capacitance, broken into 17 bits consisting of a large "band select" capacitor, three 4-bit binary-weighted switchable arrays, and a varactor controlled by a 4-bit DAC. To prevent gaps in the tuning range due to device

mismatch, the LSB of each 4-bit array controlled slightly less capacitance than the full capacitance of the next smallest array. The resulting overlap can be removed by an appropriately calibrated recoding of the control bits, requiring 112 bits of memory.

The switchable capacitances needed to be very high Q, ideally much better than that of the off-chip inductor. By tying the gate of an NMOS transistor to the oscillator core (biased near 2.5V) and its bulk, drain and source to ground, the device is put into deep inversion, presenting a low, distributed channel resistance in series with the oxide capacitance (effective Q~100). When the drain and source are tied to VDD, the channel becomes a floating node, and gate capacitance is dominated by overlap and depeletion capacitances, which are much lower, giving a large ON to OFF capacitance ratio (1:2.5 here).

The oscillator was tested using a 15nH off-chip inductor; the loaded tank Q was 30 and the oscillator covered a frequency range of 820MHz-960MHz with an accuracy of 3kHz while consuming approximately 400μW. Phase noise was -108dBc/Hz at 1MHz offset, and was dominated by bias noise and so rolled off at 40dB/decade due to bias filtering (see figure 2.3)

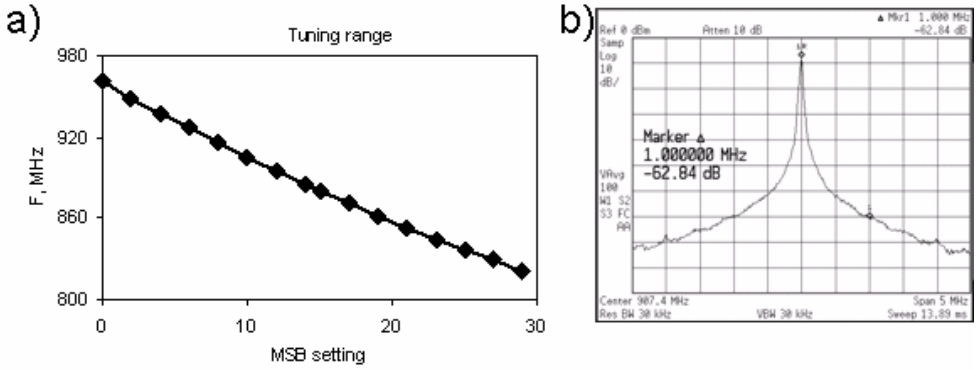


Figure 2.3. Oscillator results (measured in transmit mode). a) Tuning range for 5 MSBs of digital control. b) Spectrum of output: 5MHz span, 30kHz resolution BW.

Frequency Control

In order to be useful, the frequency of local oscillator must be controlled with high precision. This precision is necessary to guarantee that the receiver and transmitter and receiver are centered at the same carrier frequency with much higher precision than the band width of the transmitted data. This precision is achieved by comparing the oscillator frequency to a fixed frequency, much slower crystal oscillator, which also can be used to maintain a local clock for network synchrony. This frequency comparison requires dividing down the frequency of the local oscillator. Thus, a critical block for this process is a low power, high frequency divider.

RF divider.

To observe the frequency of the RF oscillator, a digital counter was used to count the number of RF cycles per cycle of a crystal oscillator. The counter comprised a series of T-flip-flop. Each consecutive stage switches half as often as the previous, so the overall power depends primarily on the first few stages.
The first factor of 8 frequency division was implemented using a 4-stage Johnson counter, shown in figure 3.4. Its constituent D-flip-flops consisted of two CMOS inverters interleaved with pass-gates, directly driven by the resonant tank of the oscillator. The circuit was self biased through a controllable current source and decoupling capacitor.

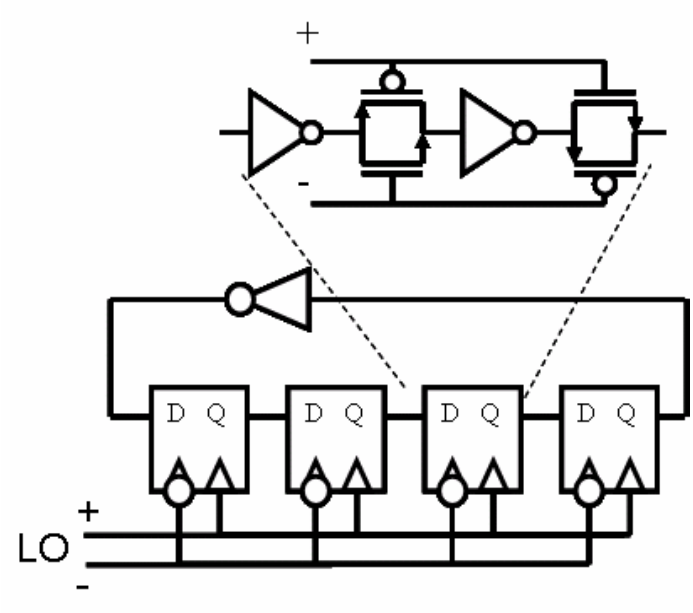


Figure 2.4. Schematic of 900MHz Johnson counter. Four dynamic flip-flops form a ring with one additional feedback inverter. Each flip-flop comprises two inverters interleaved with two complementary pass gates driven from the oscillator core with opposite polarity.

The total power consumed by the counter depends the number of stages, N, in the Johnson counter, and is set by the relationship:

$$P_{TOT} = F_{osc} \cdot V_{dd} \cdot C_{inv} \cdot \left(2 + 4/N + 2N/Q\right) \qquad \text{(eq. 2.2)}$$

Where *Cinv* is the capacitance associated with a given inverter or passgate (assumed equal). and the first term comes from one flip-flop switching each oscillator cycle one inverter. The second term comes from oscillator loading by gate capacitance of its pass-gates, and so is proportional to N, but, being resonated inversely proportional to the effective Q of the switches' gates. By using NMOS and PMOS transistors of

equal width and biasing them appropriately, an effective gate Q of about 10 was achieved.

The third term comes from the feedback inverter of the Johnson counter, and the ripple counter that follows it, each of which switch once per cycle of the Johnson counter.

Optimizing power for Q = 10, gives N = 4. The divider was measured to consume 60μW active power and functioned across the entire oscillator range (820MHz to 960MHz). This was followed by a chain of four custom dynamic T flip-flops and a reference re-timing circuit (discussed below), which together consumed an additional 18 μW. The resulting output frequency was ~7MHz, where standard cell logic can operate efficiently. This output was also used to generate the IF signal for the receiver and as the clock for various digital circuits.

Digital Feedback Loop

To demonstrate that the oscillator and divider could be used to generate a reliable RF frequency with low power, a digital feedback loop (see figure 2.5) was designed in standard cell logic to set the RF oscillator to any frequency desired within its tuning range and correct for disturbances introduced by supply and bias changes. A running count of oscillator cycles is maintained by a 4 stages of custom dynamic logic and 12 stages of standard cell logic. The standard cell part consisted of 3 sub-blocks, each with a 4 stage look ahead, such that settling time was set by the dynamic stages, two look-ahead circuits and the delay through a given standard cell subblock:

$4T_{dyn}+2T_{LA}+4T_{SS} < 4ns$ *(eq. 2.3)*

The state of this composite counter is latched once per cycle of the 32kHz reference crystal oscillator. The reference is retimed by the falling edge of the Johnson counter to ensure that the ripple counter has 4ns to settle before being sampled. Successive latched counts are subtracted, giving the ratio between oscillator and reference frequencies. This ratio is compared with a desired ratio, and errors are accumulated and fed back to the oscillator. Although the loop settles slowly (time constants on the order of 200μs), it is very low power, consuming 5 μW. This implementation could be coupled with faster start-up algorithms, or can be pre-loaded with previously saved settings.

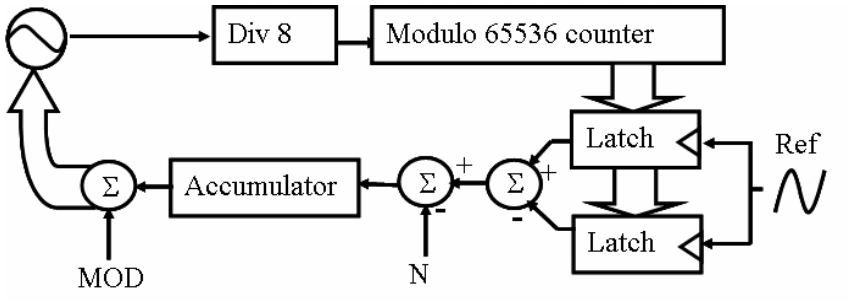


Figure 2.5. Digital feedback loop for frequency centering. The 900MHz oscillator output is divided by 8, and feeds into a 16-stage ripple counter. The state of this counter is latched in by the reference signal. Successive states are subtracted, and the difference is compared to N. Any error is accumulated and fed back.

Transmitter

In transmit mode, the oscillator drives the "power amplifier" (PA), which is designed to efficiently couple power into a relatively low impedance antenna. Frequency (or phase) modulation is introduced by directly modulating the local

oscillator. For most tests simple FSK modulation was used. Because the oscillator is at the same frequency as the PA, there is some risk of the PA interfering with the loop gain of the oscillator. However, since the output power of the PA is only slightly higher than that dissipated by the oscillator, there is never enough power coupled into the oscillator tank to significantly degrade performance.

The design challenge in the transmitter was to design a PA that would efficiently drive a signal from the high impedance oscillator tank onto a low impedance antenna (assumed to be in the range of 50-200Ω). For reasons of cost and process availablity, high-Q matching networks were not an option, beyond including some series inductance which could be built into the antenna. Using 1mW from a 3V supply, we have about 350μA bias current (once divider current is siphoned off). Stacking with the oscillator and reusing current leaves 1.8V available for the PA, which for a class-A amplifier implies an optimum load impedance of 2.5kΩ. A switching, push-pull PA, using a common drain amplifier to push 2Ibias onto the antenna for one half-cycle, and a common source to pull 2Ibias off for the other [10], gives an effective source impedance of 1.25k. Stacking two such amplifiers and ac-coupling their outputs, as shown in figure 2.6, pushes or pulls 4IBias from the antenna each half cycle, and divides the available headroom so that each stage has 0.9V available. This set-up results in an effective source impedance of 325Ω, which can then be matched to the antenna through a single inductor.

The transistor gates were biased through 60k resistors to appropriate levels for optimal switching, while the transistors were sized to optimize efficiency, trading off the load they present to the oscillator and the amount of series resistance they impose

when fully on. A tail current sink was necessary to fix bias current and so prevent unwanted feedback between oscillator amplitude and bias current. Because the voltage swing on the oscillator is somewhat larger than on the PA output, all of the transistors will be driven into triode when they are on. The PA adjusts to changes in bias current or load impedance by shifting the bias points of the intermediate nodes. Thus when output swing is reduced, the internal nodes will increase their voltage compared to ground. This degrades efficiency, since headroom is used less efficiently, but because increasing the biasing of the transistors' sources relative to their gates reduces the duty cycle of each switch, the PA is driven into class-C operation, improving efficiency. Thus the PA can maintain reasonably constant efficiency across a wide range of bias currents and load impedances.

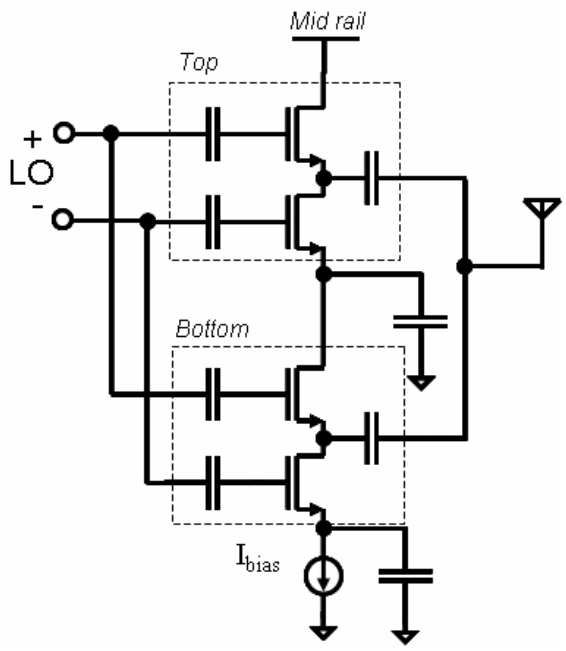


Figure 2.6. Stacked antenna driver. Push-pull amplifiers are AC coupled in parallel to the antenna. Bias current is set by a current mirror passes through the output stages in series. Large intermediate capacitors AC ground the intermediate supply rails. Gate bias resistors not shown.

The transmitter was tested driving a 50Ω load in series with 12nH. By varying the bias current setting, the transmitter output power could be swept from 50 μW to 350μW while maintaining a PA efficiency of 20-35% and a system efficiency (including oscillator, dividers, and biasing) between 10% and 19% (see figure 2.7). Maximum output power and efficiency were achieved for a load impedance of 100Ω, with output power about 0.3dB higher than for 50Ω and a peak efficiency of 20%. Higher output powers up to 1mW were possible by de-stacking the oscillator and PA to provide more headroom, at the cost of reducing system efficiency to 15%.

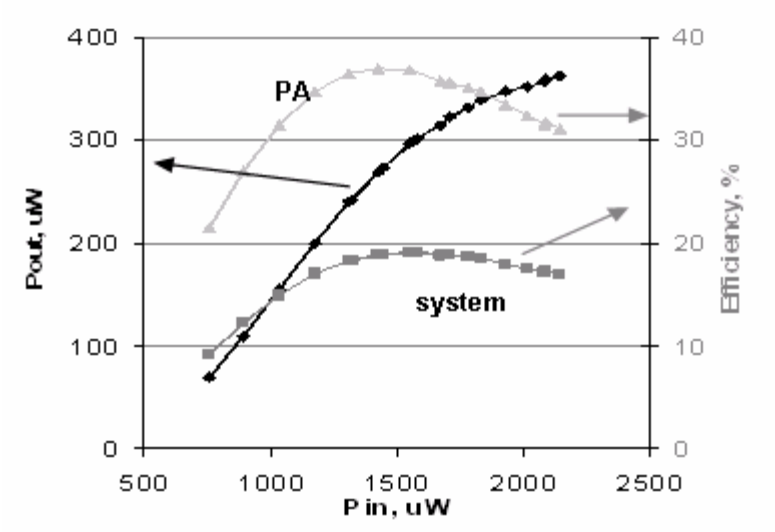


Figure 2.7. Output power (left y-axis) ranges from 80μW to 360 μW vs input power (x-axis) rnaging from 750 μW to 2200 μW. Power amplifier (PA) and overall system efficiency (right y-axis) vs input power. System efficiency includes PA, oscillator and dividers.

Receiver Architecture

In designing a low power receiver, the number of circuits operating at RF frequencies was minimized. In order to be both robust and low cost, the receiver needed to be insensitive to wide band interference, so that it could function without an off chip filter. Several architectures were investigated, as shown in figure 2.8, including: 1) A simple rectifier-based ASK receiver consisting of an RF amplifier and rectifier, 2)A super-regenerative receiver wherein an oscillator on the cusp of oscillation is fed the RF input and provides near-infinite gain, 3) a direct conversion receiver (DCR) and 4) A low-IF receiver.

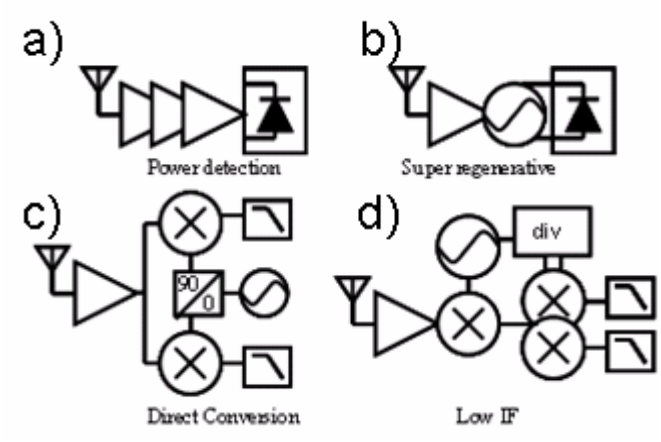


Figure 2.8. Architectures for low power radio receivers (from [9]). a) simple power detection, b) super-regenerative power detection c) Direct Conversion Receiver (DCR) d) Low IF receiver.

The simple rectifier will only efficiently demodulate voltage signals on the scale of 26mV or greater (see eq 1.3), and therefore will have a sensitivity of approximately

$(0.026)^2/(ZinA_V) \rightarrow -26.7-10\log(Av)$ dBm (eq 2.3)

for a 50Ω input impedance. This requires significant voltage gain to achieve reasonable sensitivity (about 50dB of gain for -80dBm sensitivity) which implies significant power consumption (thus, such an approach only makes sense where sensitivity is not important). This approach also has the problem of having essentially no frequency selectivity, and so requires a high-Q RF filter as part of its chain to reject interferers: such high-Q filtering implies an off-chip filter using a specialized component and little tuneability.

Super-regenerative receivers create a high gain, extremely narrow band amplifier using an RF oscillator held near the cusp of oscillation. Following this with

an envelope detector provides AM demodulation, resulting in a very simple receiver topology with the potential for low power operation. However, super-regenerative receivers are quite sensitive to pulling by interfering signals and generally suffer from slow settling and so must operate at correspondingly low data rates[6].

Direct-conversion and low-IF receivers amplify incoming RF signals and mix them with an RF oscillator to translate them to lower frequencies where voltage gain requires less power and channel selection may be done with on-chip filters. Though the RF oscillator increases power consumption compared to simpler topologies, such receivers enable multi-channel communication, handle a wide variety of modulation techniques, and resist interference.

Direct conversion only requires an LNA and mixer running at RF frequencies, and permits band-select filtering early in the receive chain. However, to demodulate most signals, a DCR must provide a quadrature phase split at RF frequencies (implying extra overhead power). In addition, DCRs are sensitive to interference through 2^{nd} order nonlinearity, DC offset and flicker noise[11, 12]. DC offset and flicker noise were avoided by using a low-IF receiver with an IF of 1.7MHz, and minimized power by using a single-phase RF mixer without image rejection. A block diagram of the receiver is shown in figure 2.1.

This approach saves power by providing gain at a low frequency IF while avoiding the DC offsets, IIP2 and flicker noise associated with a DCR. One issue with this approach is that the single RF mixer provides no image rejection. This lack degrades noise figure by 3dB by down-converting noise from both the wanted band and the image band and makes the radio sensitive to interfering signals at the image

frequency. However, providing image rejection carries a cost in terms of power consumption and/or signal degradation. At the very least, an image rejecting front end would require separate I and Q paths, implying two mixers and two IF chains. Based upon the final design used, an additional IF chain would consume approximately 75μW, power which would be diverted from the front-end, and would be expected to degrade noise figure by ~ 1.25dB. Image rejection also requires phase-splitting either the LO or input RF signal before mixing. Such phase splitting is commonly accomplished one of three ways (shown in Fig 2.9). These methods are: 1) Using an oscillator at twice the frequency and including a frequency divider between the oscillator and mixers[3]; 2) including an RC polyphase filter between the existing oscillator and mixers, or 3) including a polyphase filter between the LNA and mixers[12, 13]. Each of these approaches either consumes additional power or degrades noise, or both.

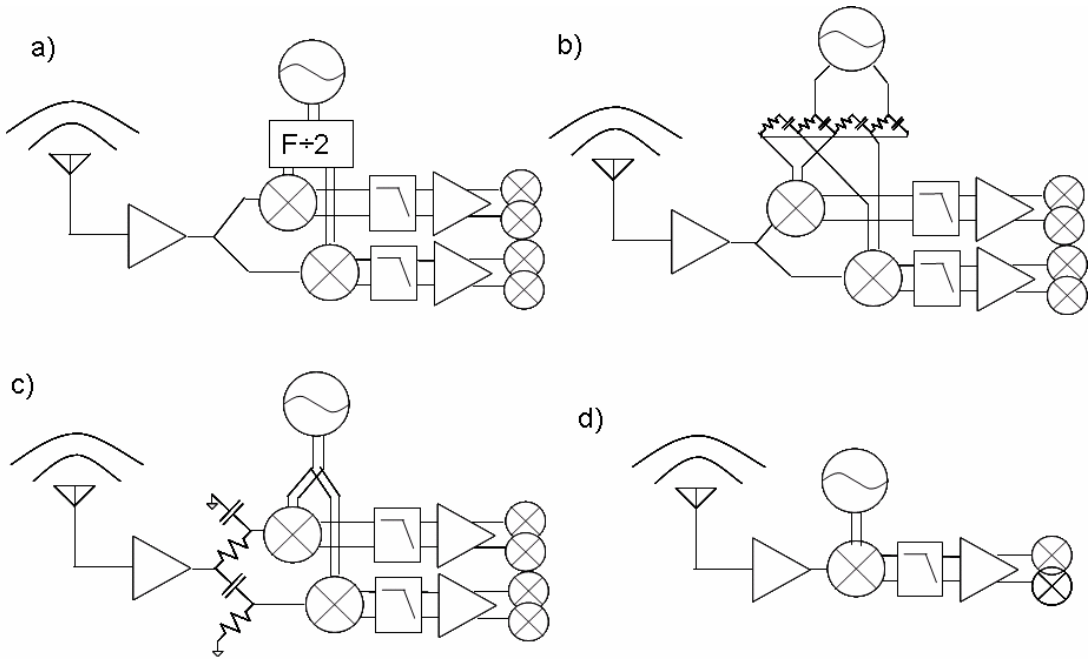


Figure 2.9 Common methods of generating image rejection: a) Run Oscillator at 2× received frequency, extract rising and falling edges for 0 and 90 degrees. b) Phase split Oscillator output with RC polyphase filter. c) Phase split RF signal after LNA with RC polyphase filter. d) No image rejection: much less circuitry, so lower power.

A divider running at 2× the present LO would consume at least twice the power of the existing divider, an additional 60 μW. Furthermore this divider would have to drive the mixers in a non-resonant mode costing about 30 μW. All of this power, when diverted from the receiver would further degrade noise figure by about 2dB. In addition, this would complicate the transmitter design (which presently is based on an on-frequency LO, not one at 2× the transmit frequency) and degrade its efficiency. Thus, this approach would not improve noise while complicating the design.

Inserting an RC polyphase filter between oscillator and mixer will automatically reduce LO swing on the mixers by 3dB, and in order to prevent more than 3dB addition loss due to voltage division between the polyphase and gate capacitance (based on the existing mixers), the resistors in the polyphase would need to have a value of less than 10kΩ. This impedance would load the oscillator, consuming at least an additional 90 μW, once again degrading noise by an additional 2dB. Furthermore, mixer noise will be effectively doubled by its reduced LO swing, further degrading noise. Once including image rejection degrades noise is more than a simple lack of image rejection would.

Finally, including a polyphase between LNA and mixers will automatically attenuate the wanted signal by 3dB. In addition the polyphase network will present a shunt path at the LNA output, causing an additional attenuation. This shunting effect can be reduced by increasing the impedance of the polyphase network, but this automatically introduces extra noise, such that overall noise performance (at least in this process using the amplifier used, etc) is always degraded by at least 3dB for any polyphase component values.

Thus, a design with no image rejection proved the lowest noise in the extreme low power regime. To address the risk of narrow-band interferers at the image frequency, the IF frequency was made programmable to allow dodging of such interference.

Almost all of the power in the receiver was spent on LO generation and RF gain. Bias current was set by the requirement of robust positive feedback in the LO (see chapter 3), and then this current was redistributed to provide optimal noise and

linearity. The majority of the current from the oscillator went to the front-end (about 200μA), followed by the IF amplifier, which used 50μA, and two baseband amplifiers consuming 13μA each. This general distribution was chosen to minimize total input-referred noise. The total input referred noise can be expected to follow the equation:

$$v_n^2{}_{tot} = v_n^2{}_{LNA} + v_n^2{}_{IF}/A_{FE}^2 + v_n^2{}_{BB}/(A_{LNA}^2 \cdot A_{IF}^2) \quad \text{(eq 2.4)}$$

That is each stage contributes noise divided by the preceding stages' gain. If one assumes (imperfectly) that input-referred squared voltage noise in an amplifier is proportional to bias current in that stage, then total noise will be:

$$v_n^2{}_{tot} \propto I_{LNA}^{-1} + I_{IF}^{-1}/A_{FE}^2 + I_{BB}^{-1}/(A_{IF}^2 A_{FE}^2) \quad \text{(eq 2.5)}$$

if we then define:

$$I_1 = I_{IF} + I_{BB}$$

$$I_{tot} = I_1 + I_{LNA} \quad \text{(eq 2.6)}$$

We can find the combined noise of the IF and baseband to be:

$$v_n^2{}_1 \propto (I_1 - I_{BB})^{-1} + I_{BB}^{-1}/(A_{IF}^2) \quad \text{(eq 2.7)}$$

which is minimum for a given I_1 when

$$I_{BB} = I_1(1 + A_{IF}) /(A_{IF}^2) \quad \text{(eq 2.8)}$$

Similarly overall noise is roughly minimum when

$$I_1 = I_2 (1 + A_{FE})/(A_{FE}^2) \quad \text{(eq 2.9)}$$

Thus, each stage should burn approximately 1/A the current of the previous stage. The final design tended to burn more current in the IF and baseband stages than this simple analysis would suggest. Partly this is because the first stage is single ended, while succeeding stages are differential and so require approximately twice the current to achieve the same noise. Furthermore extra current was needed in later

stages to suppress flicker noise [14], which tends to degrade noise for near-dc signals (ie baseband signals) but not RF signals.

The RF and IF mixers were both passive, providing low power consumption and very good linearity. Each mixer was followed by a programmable-value capacitance, setting its output bandwidth. These capacitors interacted with the mixers to low pass filter the output of each stage, suppressing unwanted interferers, preventing them from saturating subsequent stages. This receiver was designed to be general purpose in terms of modulation scheme and data rate. However, to demonstrate true data transmission, a binary frequency-shift-keyed (FSK) demodulator also included after the baseband amplifiers capable of demodulating up to 100kb/s.

Front-end Design

The receiver front end is shown in figure 2.10. It performs the functions of low noise amplification, RF down-conversion and IF low-pass filtering in a structure of 4 transistors and 3 capacitors.

A self-biased CMOS inverter acts as an amplifier and converts the input voltage to current, which is alternately driven to one of two output capacitors by a pair of passive NMOS switches. These switches' gates are driven directly by the oscillator, sampling the inut at the LO frequency. The circuit can be modeled as a discrete time system whose differential output voltage updates each LO cycle according to the equation:

$$V_{out}(kT+T) = V_{out}(kT)\frac{C_L/2}{C_L/2 + C_{PAR}} + \int_0^{T/2} V_{RF}(t+kT)\frac{gm}{C_L}dt - \int_{T/2}^{T} V_{RF}(t+kT)\frac{gm}{C_L}dt \quad \text{(eq 2.10)}$$

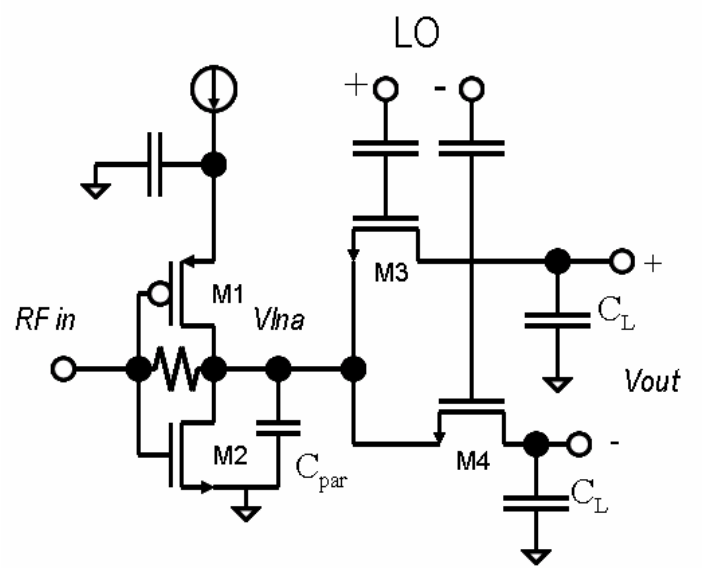


Figure 2.10. RF front-end: self-biased CMOS inverter (M1 and M2) provides gain. A large capacitor provides an AC ground from the PFET. C_{par} is not an explicit capacitor, while M3 and M4 sample the signal onto the load capacitors (C_L).

Where gm is the combined transconductance of the N- and PMOS input devices and C_{PAR}, the parasitic capacitance at Vmid. The two integration terms describe how charge from the initial amplifier accumulate on the output capacitors during each half cycle of the LO. The first term captures the effect of parasitic capacitance, C_{PAR} at the intermediate RF node, which interacts with the output capacitors in a way similar to a switched-capacitor resistor, dissipating charge from the output with each cycle of the LO. If the input, $V_{RF}(t)$ is sinusoidal with frequency f_{RF}, then $V_{IF}(t)$ will also be a sinusoid with frequency $f_{IF} = |f_{RF} - f_{LO}|$ where $f_{LO} = 1/T$. Assuming that this is true, and that $f_{LO} >> f_{IF}$ then the relationship between the input (at f_{RF}) and output (at f_{IF}) can be approximated by the continuous time transfer function:

$$\frac{V_{IF}}{V_{RF}} = \frac{gm}{\pi} \frac{1}{C_{PAR} f_{LO} + j\omega_{IF} C_L / 2} \qquad \text{(eq 2.11)}$$

In-band gain thus depends on the ratio of the gm of the active devices to the parasitic capacitance at their drains. Since both the capacitance and transconductance depend upon the width of the active transistors, maximizing this ratio sets the sizing of the active transistors, optimized when V_{DSATN} is ~250mV and V_{DSATP} is ~ 500mV. The gate capacitance of these devices is less important, as it is in parallel with and dominated by parasitic loading from the bond pad and transmitter. The RF switches were sized for optimal system noise figure: very narrow switches generate excessive thermal voltage noise; very wide switches contribute excessive capacitance to C_{PAR}, reducing gain. The load capacitors were made programmable, and were selected to set the IF bandwidth when interacting with C_{PAR}. Overall the structure was designed to have a nominal voltage gain of 26dB, NF of 9dB and BW of 2.5MHz when driven by a 100Ω antenna in series with 14nH.

Biasing of the mixer gate voltages was delivered through minimum-width, 50kΩ resistors. This voltage was set up by two stacked, diode-connected NFETs set to roughly match the LNA and switch transistors so that when no LO was present, the mixer switches would be off but close to the turn on threshold. Ideally the switch transistors turn on strongly at the peak of the LO, but are completely OFF at the transition point so that no leakage can occur between the two output capacitors. Such leakage would reduce the gain of the front-end. To ensure the optimal bias could be found in the face of unexpected process variability or model errors, this voltage was made programmable.

This structure provides very good wide-band linearity by suppressing wide-band interferers not just at its output, but at the intermediate RF node at the LNA drain

devices as well. Because this node is alternately shorted to the two output capacitors at the rate of the LO, the voltage on this node should look a lot like an up-converted version of the voltage across the IF port. This voltage will be superposed with the voltage from RF current from the LNA passing through the finite impedance of the switches and load, giving an overall voltage transfer function of:

$$V_{MID} = V_{RF}\, gm\left(\frac{1}{\pi^2}\frac{1}{C_{PAR} f_{LO} + j\omega_{IF}\, C_L/2} + \frac{jC_L\omega_{RF}R_{CH}+1}{jC_L\omega_{RF}}\right) \qquad \text{(eq 2.12)}$$

Where the first term reflects the output signal from eq 2.11, and the second the instantaneous impedance presented by the switch plus load. Figure 2.11 shows that equations 2.11 and 2.12 predict similar results to numerical circuit simulations.

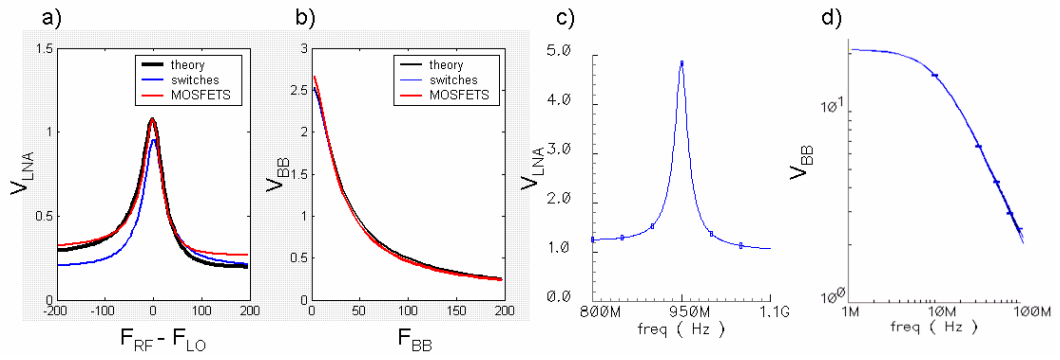


Figure 2.11. switched filtering effect. a) Comparison of equation 2.12, MATLAB simulations of ideal resistive switches and MATLAB simulations of level-1 MOS models of switches for V_{LNA}. b) Comparison of equation 2.11, MATLAB simulations of ideal resistive switches and MATLAB simulations of level-1 MOS models of switches for V_{BB}. c) SpectreRF (Cadance) simulation of V_{LNA} in figure 2.10, d) SpectreRF (Cadance) simulation of V_{BB} in figure 2.10 (note log scale)

This function appears to describe a high-Q band-pass filter, and, indeed, the first term can be approximated (for $f_{LO} >> f_{IF}$) by a 2nd-order transfer function as follows:

First, multiply top and bottom by $4/C_L$ and convert f_{LO} to ω_{LO}:

$$\frac{\frac{1}{\pi^2}}{C_{par} f_{LO} + j\omega_{IF} C_L/2} = \frac{\frac{4}{C_L \pi^2}}{\frac{2C_{par}\omega_{LO}}{\pi C_L} + j2\omega_{IF}} \qquad \text{(eq 2.13)}$$

Next add and subtract $j\omega_{LO}$ from the denominator:

$$- = \frac{\frac{4}{C_L \pi^2}}{\frac{2C_{par}\omega_{LO}}{\pi C_L} + j(\omega_{IF} + \omega_{LO}) - j(\omega_{LO} - \omega_{IF})} \qquad \text{(eq 2.14)}$$

Now multiply top and bottom by: $j(\omega_{LO} + \omega_{IF})$ *and substituting* $(\omega_{LO} + \omega_{IF}) = \omega_{RF}$:

$$= \frac{j\omega_{RF}\frac{4}{C_L \pi^2}}{\omega_{LO}^2 - \omega_{IF}^2 + j\omega_{RF}\omega_{LO}\frac{2C_{PAR}}{C_L \pi} - \omega_{RF}^2} \qquad \text{(eq 2.15)}$$

Finally since $\omega_{IF}^2 << \omega_{LO}^2$ we can neglect the ω_{IF}^2 term and rearrange get the simple 2nd-order bandpass function:

$$= \frac{j\omega_{RF}\frac{4}{C_L \pi^2}}{\omega_{LO}^2 + j\omega_{RF}\omega_{LO}\frac{2C_{PAR}}{C_L \pi} - \omega_{RF}^2} \qquad \text{(eq 2.16)}$$

That is, the filter has a center frequency equal to f_{LO}, and so automatically tracks this frequency, while it has a Q set by the ratio of parasitic and load capacitances, about 10pF/50fF in this design, giving a Q~200,. The maximum wide-band attenuation possible, assuming $R_{CH} >> 1/(\omega_{RF} C_L)$, is set roughly by the product $C_{PAR} \cdot \omega_{LO} \cdot R_{CH}$,

which improves with increased switch width, but asymptotically approaches a fixed limit for a given process and LO swing as the parasitic capacitance from the switches comes to dominate C_{PAR}.

By shunting the wideband signals at the LNA output, the mixer prevents large, wide-band interferers from driving the LNA transistors into triode, improving wide-band linearity. In this design, the out-of-band attenuation was simulated to be 9dB, giving a simulated out-band compression point of about -20dBm.

IF Chain Design

The IF chain comprised a differential amplifier, an 8-phase passive mixer and a digital DC-offset correction circuit. The amplifier design is shown in figure 2.12. It consists of a complementary set of N- and PMOS differential pairs, loaded by a large composite resistor. To maximize transconductance in the amplifier, the transistors were biased in weak inversion (V_{DSAT} = 50mV). Biased this way, the transistors are in an intermediate state between square-law behavior and (sub-threshold) exponential behavior. Bias current was sourced by a PMOS current source, and sunk by an NFET in weak triode, whose gate voltage was set through common mode feedback from the mid-point of the load resistor. Input bias voltage comes directly from the front end, and were equal to the self-biased point of the LNA's NFET. The output bias voltage was set by the feedback to the tail NFET, and ideally, it would be equal to the input bias voltage, maximizing available swing on the output. This was achieved by sizing the tail NFET such that for a drain current of Ibias, and drain voltage equaled V_{GS}

from the LNA minus V_{GS} from the IF amplifier active devices, the gate voltage would equal the V_{GS} of the LNA.

The amplifier was designed to have a maximum voltage gain of approximately 30dB set by the combined transconductance of the complementary differential pairs and the load presented by the output resistors. IF gain was programmable by switching the drains of the transistors to different intermediate points on the load. This setup kept output impedance constant, but permitted gain to change according to the resistance between the drains of the active devices. Constant output impedance is important since it interacts with the IF mixer load capacitors to set up the output low pass filtering bandwidth.

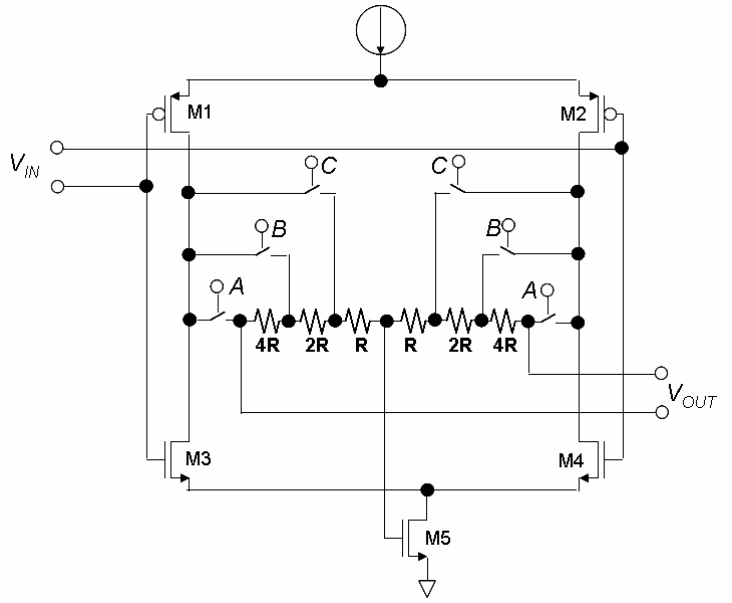


Figure 2.12. IF amplifier. Two PFETs (M1 and M2) and two NFETs (M3 and M4) act as complementary differential pairs, biased by a PFET current source and NFET (M5) in common mode feedback. Output load an common mode feedback are provided by a switchable resistor. Only one of switches A, B, or C are closed at a time.

The IF amplifier was followed and loaded by a quadrature, double-balanced passive mixer. The mixer was loaded by filtering capacitors and provided the same band-pass loading effect described for the RF mixer, suppressing out-of band interferers and DC offsets at the IF amplifier output. To prevent the load capacitors discharging each other, the LO signals switching the I and Q sampling switches did not overlap. This implies that each switch should only be on 25% of the time, with a total of 8 switches and 4 sampling capacitors required.

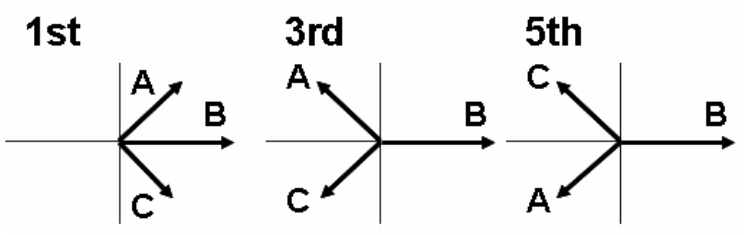


Figure 2.13. Principal of harmonic cancellation: Phasor diagram of three signals at 45 degrees split. If summed at he fundamental they add roughly in phase, but for the 3rd and 5th harmonics, the phase split increases and these harmonic signals actually cancel. Note that for this to work, the middle phase must be weighted by ~1.4 relative to the other two.

In fact twice as many switches and load capacitors were used. A standard mixer will mix down not just the IF, but also its odd harmonics with a gain of 1/m, where m is the number of the harmonic. This harmonic mixing makes the receiver sensitive to interferers at those harmonics. Although filtering immediately following the front-end should act to suppress these interferers before they reach the IF mixer, for the lowest harmonics (where the harmonic mixing is strongest), this filtering will

be relatively weak. To reduce such interference, the mixer was modified, using a method mathematically similar to that described by Weldon et. al.[15], so that it rejected the third and fifth harmonics. Each path (I and Q) comprised three mixers in parallel, driven by three 45°-split LO signals, whose outputs are recombined in a weighted sum to cancel the 3rd and 5th harmonic mixing terms, as shown in figure 2.12. Since the I and Q outputs can share some of these mixers (ie. the +45° mixer for I is the same as the -45° for Q) a total of 8 capacitors and 16 switches are required, driven by 8 non-overlapping 12.5% duty-cycle digital signals. These 8 sampled signals were then weighted and combined in two triple-input differential baseband amplifiers to yield the desired I and Q signals (see figure 2.13). Relative gain (1 for ± 45° vs 1.4 for 0°) was achieved by scaling transistor width and bias current in the differential pairs that made up these amplifiers.

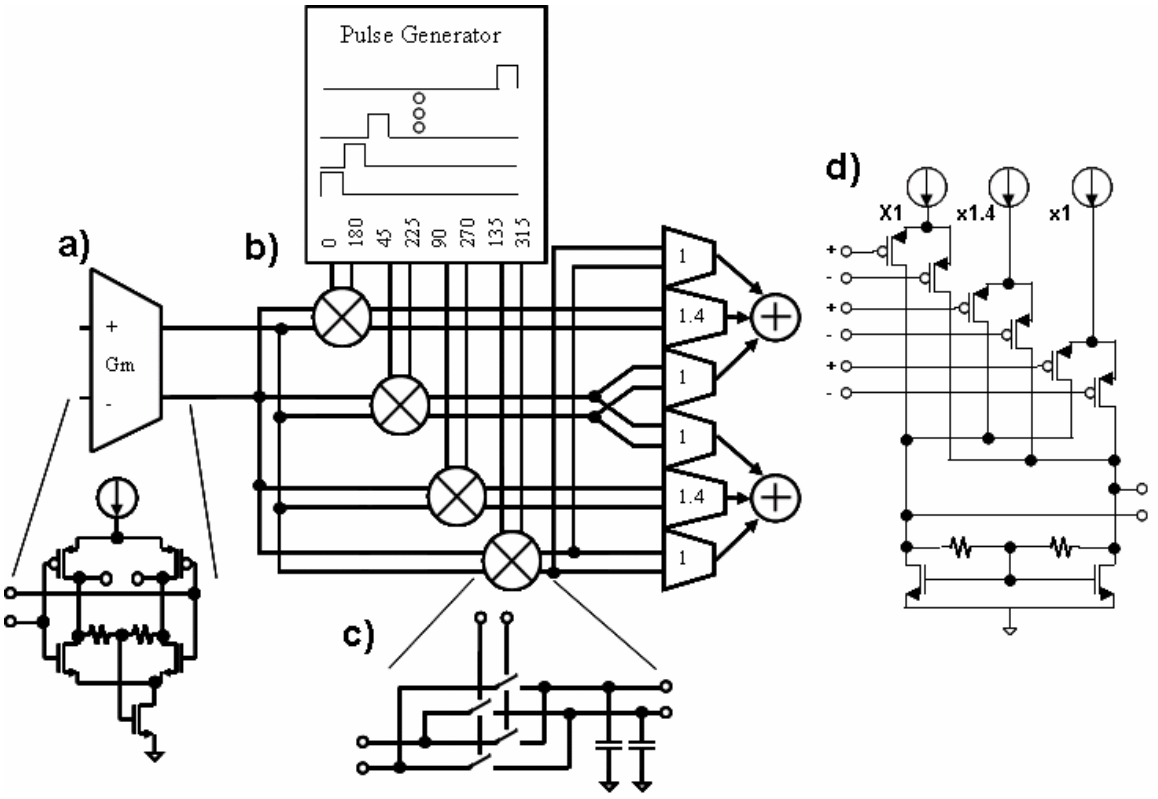


Figure 2.13. IF mixer. a) Input is from the IF amplifier which has a fixed output impedance. b) Mixer consists of 4 double-balanced switching mixers each driven by a pair of 12.5% duty cycle pulses 180º from each other. c) Each sub-mixer consists of four switches connected in double-balanced topology, loaded by two large capacitors. d) the resulting four differential outputs are recombined through a pair of 3-input differential amplifiers. Each input is an appropriately scaled differential pair. The resulting currents are summed and loaded by a pair of NFETS arranged in common mode feedback.

The signals that drove the mixer switches were derived from a divided-down LO signal at ~7MHz (see Chapter 3). This signal was then further divided, programmably, by 1,2,4, or 8, and then by 4 in such a way as to generate the desired 8 phases at 1.8MHz, 900kHz, 450kHz, or 225kHz. The non-overlapping signals were then generated using combinatorial CMOS logic designed in standard cells (see figure 2.14).

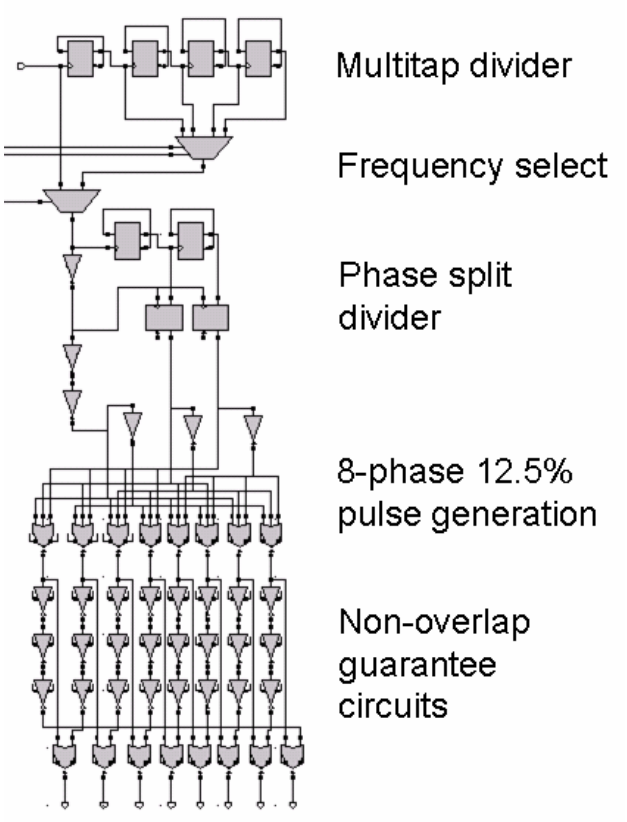


Figure 2.14. Non-overlapping 12.5% 8-phase pulse generator circuit.

As with the passive RF mixer described above, this passive mixer will reflect the filtering on its output to suppress input IF signals offset from the wanted signal. In this structure, output impedance of the amplifier contains both resistance and capacitance. Also, there are 4 distinct differential load capacitors, such that the update equation for each voltage depends on the other outputs. For the nth output, this equation is:

$$V_n(kT+T) = \left(C_L V_n(kT) - C_{PAR} V_{n-1}(kT)\right)\left(\frac{1}{C_L + 2C_{PAR}} \frac{4RC_L - T}{4RC_L}\right) + \int_{(n-1)T/8}^{nT/8} \frac{gm}{C_L}\left(V_{IF}(t+kT) - V_{IF}\left(t+kT+T/2\right)\right)dt \quad \text{(eq. 2.17)}$$

Where C_L and C_{PAR} are as defined as for the RF mixer, V_{IF} is the IF amplifier input, and gm and R are the transconductance and output resistance of the IF amplifier. This can be rewritten for the vector of outputs:

$$\vec{V}(kT+T)=\left(\frac{C_L(4RC_L-T)}{(C_L+2C_{PAR})4RC_L}\right)\begin{bmatrix}1 & 0 & 0 & -\delta\\ \delta & 1 & 0 & 0\\ 0 & \delta & 1 & 0\\ 0 & 0 & \delta & 1\end{bmatrix}\vec{V}(kT)+\vec{u}(kT) \qquad \text{(eq. 2.18)}$$

Where $\delta = C_{PAR}/CL$. The input u(kT) is a vector defined as follows:

Assume the input $V_{IF}(t)$ is roughly sinusoidal and close to the m^{th} harmonic of the IF such that

$$V_{IF}(t) = A\cos(m2\pi t/T+\Delta\omega t) \qquad \text{(eq. 2.19)}$$

Then the integral terms of eq 2.17 become, approximately:

$$u(m,n,k)=\frac{Agm}{C_L}\frac{T}{m2\pi}\left(\sin\left(\frac{mn\pi}{4}+\Delta\omega kT\right)-\sin\left(\frac{m(n-1)\pi}{4}+\Delta\omega kT\right)\right)$$

$$-\frac{Agm}{C_L}\frac{T}{m2\pi}\left(\sin\left(\frac{mn\pi}{4}+m\pi+\Delta\omega kT\right)-\sin\left(\frac{m(n-1)\pi}{4}+m\pi+\Delta\omega kT\right)\right) \qquad \text{(eq. 2.20)}$$

Which equals zero for even m and for odd m is

$$\frac{Agm}{C_L}\frac{T}{m\pi}\left(\sin\left(\frac{mn\pi}{4}+\Delta\omega kT\right)-\sin\left(\frac{m(n-1)\pi}{4}+\Delta\omega kT\right)\right) \qquad \text{(eq. 2.21)}$$

Which comes to

$$\frac{Agm}{C_L}4T\,\mathrm{sinc}\left(\frac{m\pi}{8}\right)\cos\left(\frac{m(2n-1)\pi}{8}+\Delta\omega kT\right) \qquad \text{(eq. 2.22)}$$

While the for 4 phases for the four outputs are:

mπ/8, 3mπ/8, 5mπ/8, 7mπ/8

so, the vector u(kT) is:

$$\vec{u}(kT) = \frac{A \cdot gm}{C_L} \frac{4}{m\omega_{IF}} \sin\left(\frac{m\pi}{8}\right) \begin{pmatrix} \cos\left(\frac{m\pi}{8} + \Delta\omega kT\right) \\ \cos\left(\frac{3m\pi}{8} + \Delta\omega kT\right) \\ \cos\left(\frac{5m\pi}{8} + \Delta\omega kT\right) \\ \cos\left(\frac{7m\pi}{8} + \Delta\omega kT\right) \end{pmatrix} \quad \text{(eq. 2.23)}$$

To find the transfer function of the IF mixer, we can then now take the z-transform of V:

$V(z)=(zI-aC)u(z)$ (eq. 2.24)

The poles of the system are just the eigenvalues of the matrix in equation 2.18,

$a(1-\delta+j\delta)$, $a(1-\delta-j\delta)$, $a(1+\delta+j\delta)$, $a(1+\delta-j\delta$

where the scalar term

$$a = \left(\frac{C_L(4RC_L - T)}{(C_L + 2C_{PAR})4RC_L} \right) \quad \text{(eq. 2.25)}$$

These values are associated with eigenvectors (as columns of a matrix):

$$E = \frac{\sqrt{2}}{4} \begin{bmatrix} 1-j & 1+j & -j\sqrt{2} & j\sqrt{2} \\ \sqrt{2} & -\sqrt{2} & -1-j & -1+j \\ 1+j & 1-j & -\sqrt{2} & -\sqrt{2} \\ -j\sqrt{2} & j\sqrt{2} & -1+j & -1-j \end{bmatrix} \quad \text{(eq. 2.26)}$$

If we write V(z) out in these terms, we get:

$V(z)=(E(zI-a\Lambda)^{-1}E^*)\mathbf{u(z)}$ (eq. 2.27)

Where Λ is the diagonal matrix of eigenvalues. We can rewrite **u(z)** as the product of a scalar term and a phase vector:

$$\vec{u}(z)=\begin{pmatrix} e^{jm\pi/8} \\ e^{jm\pi 3/8} \\ e^{jm\pi 5/8} \\ e^{jm\pi 7/8} \end{pmatrix}\left(\frac{gm}{C_L}4T\sin c\left(\frac{m\pi}{8}\right)\right)u(z)=\vec{\phi}(m)\left(\frac{gm}{C_L}4T\sin c\left(\frac{m\pi}{8}\right)\right)u(z) \quad \text{(eq. 2.28)}$$

But for each odd value of m, the phase vector is parallel to one of the eigenvectors and consequently orthogonal to the other three, introducing three zeros that cancel three poles, leaving only one pole that actually affects the behavior of the system. Thus for m = 1

$$\vec{V}(z)=\begin{pmatrix} j\sqrt{2} \\ -1+j \\ -\sqrt{2} \\ -1-j \end{pmatrix}\frac{1}{z-a(1+\delta-j\delta)}\frac{gm}{C_L}\sqrt{2}T\sin c\left(\frac{\pi}{8}\right)u(z) \quad \text{(eq. 2.29)}$$

Where u(z) is now just the scalar input to the IF amplifier. This is true for all odd values *of m, but each m extracts a different pole:*

m = 3: p = a(1 - δ – jδ)

m = 5: p = a(1 - δ + jδ) (eq. 2.30)

m = 7: p = a(1 + δ + jδ)

all of these poles are close to the fixed value*, a.* but each introduces a slight shift, leading to slightly asymmetric filtering for frequencies slightly higher or lower than the IF frequency, as can be in simulation (see figure 2.15b)

As with the RF mixer described earlier, the IF mixer output is reflected on its input. Thus the input of the IF selects for signals near the IF frequency and its odd harmonics with a bandwidth of *2a.* The effective input impedance of the mixer is shown in the simulation results in figure 2.15a.

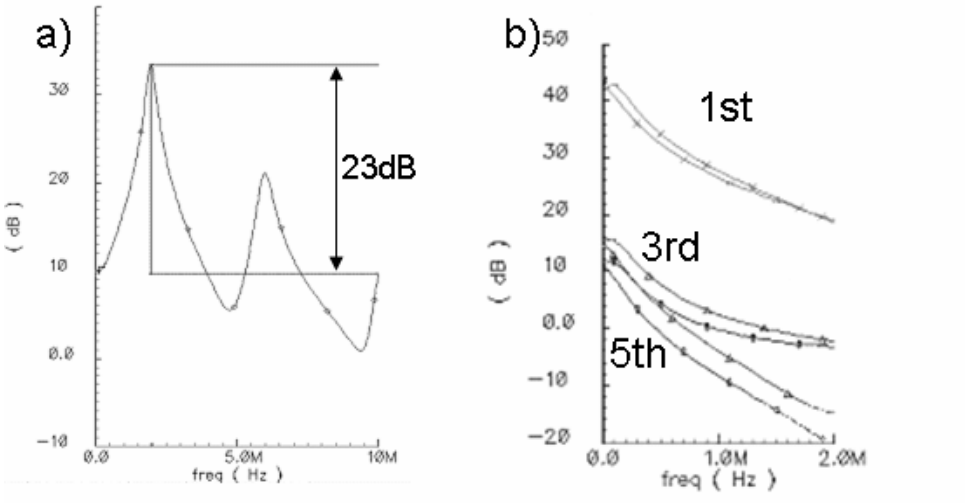


Figure 2.15. a) SpectreRF simulations of gain at IF amplifier output for a 0dB input signal. Note that signals close to the IF and 3rd harmonic are selected over other frequencies, including DC. b) Gain at baseband output. Note that 1st harmonic mixing is close to 30dB stronger than the mixing of the third harmonic, showing harmonic suppression. Also note that the peak of the filter is slightly offset from 0Hz, as predicted by equation 2.29.

This effect is especially important in that it will suppress DC offsets at the output of the IF amplifier. This suppression is ultimately limited by the output capacitance of the mixer. Ideally, each 1/8 cycle, the DC current will cause the output voltage to slew at a rate of I_{DC}/C_L such that

$$V(T/8) = V(0) + (I_{DC}/C_L)T/8 \quad \text{(eq. 2.31)}$$

Since each half cycle, the same capacitors are presented to the mixer output, but in the opposite configuration, we know that the system will settle to a state where V(0) = -V(T/8), such that the peak output due to the DC current will be

$$Vpeak = (I_{DC}/C_L)T/16 \quad \text{(eq. 2.32)}$$

This signal will then be presented to the succeeding differential pairs that follow the sampling capacitors. Thus, the amount of DC offset the IF amp can handle is really set by this peak swing, and not the actual output term at DC.

This DC offset can come from a variety of sources, including mismatch in the transistors themselves, from even order nonlinearity, and from self mixing in the RF mixer. This last mechanism is likely to dominate because the oscillator uses an off-chip inductor, whose bond wires are close to the antenna input, leading to strong coupling. This coupling and its consequent DC offset can be estimated as follows: First, the oscillator has a swing of about 1V. Each bondwire has an inductance of ~1.5nH (a very rough estimate) out of a total inductance of about 15nH, and so carries about 100mV of LO signal. Coupling between alternate bondwires can be estimated to be about 0.1. Thus the LNA input should have about 10mV of LO signal on it. This signal will be amplified and down-converted by the RF front-end, and because it will be exactly the LO frequency, it will down-convert to DC. Assuming front-end gain of 20dB, the IF amplifier can be expected to see a DC offset as large as 100mV, enough to compress the input, and to generate a large peak output on the order of 0.5V, compressing the output. This offset, however, should be relatively constant, and so can be cancelled. To perform this cancellation, a 6-bit current-mode DAC was included on the input of the IF amplifier. This DAC was not required to completely cancel the offset, but just to suppress it enough that succeeding stages would not be compressed. The DAC was implemented as a pair of binary-weighted multi-tap current mirrors connected to the two outputs of the front end. Each bit turned on a given tap on one side and turned it off on the other.

Now to find the final I and Q outputs, we just multiply V(z) by a matrix describing the weights used when recombining the outputs:

$$\begin{pmatrix} I(z) \\ Q(z) \end{pmatrix} = \begin{bmatrix} \sqrt{2} & 1 & 0 & -1 \\ 0 & 1 & \sqrt{2}j & 1 \end{bmatrix} \begin{pmatrix} 1 \\ \left(\frac{1+j}{\sqrt{2}}\right)^m \\ j^m \\ \left(\frac{1-j}{\sqrt{2}}\right)^m \end{pmatrix} V(z) \qquad \text{(eq. 2.33)}$$

Thus the two outputs, I, and Q are 90 degrees apart but show the same pole location, and so the same filtering.

FSK Demodulator

To demonstrate complete bits-to-bits communication, a simple FSK demodulator was also included. This demodulator included three poles of additional baseband filtering, two complex poles implanted in a Sallen-Key filter and one real RC pole, with a composite bandwidth of approximately 150kHz. This filtering was folloed by a switched-capacitor amplifier with 15 dB gain and DC-offset correction on both I and Q channels. The signals were then limited and fed to a bank of 4 D-flip-flops, each of which detected the relative phase of the I and Q signals when one of them passed through zero, providing an estimate of the FSK symbol[16]. These 4 estimates were then fed to a simple voting algorithm to detect either a "1" or a "0". Simulation of the demodulator showed a BER<10^{-3} provided the SNR on each of I and Q was >15dB.

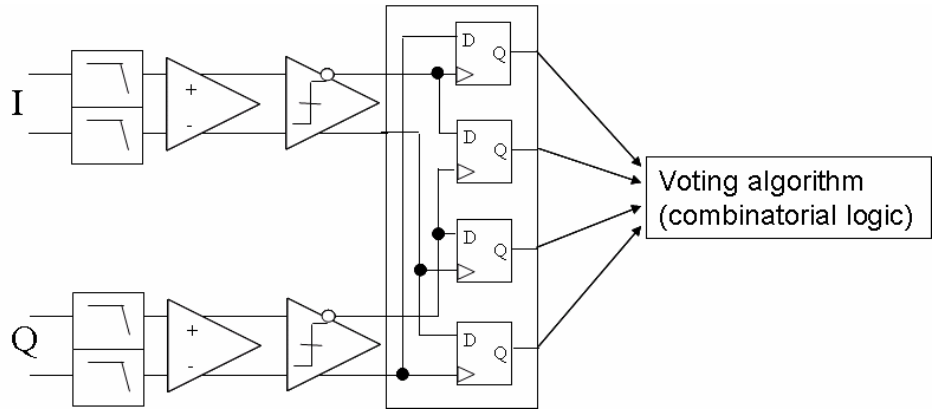


Figure 2.16. FSK demodulator

Measured Receiver Results:

All tests were performed using signals from a 50Ω source in series with 12nH and bondwire. For most of the tests, bias settings were at a nominal level, consuming 1.15mW from a 3V supply. NF was tested by observing the output noise floor of the I and Q channels, and calculating, based on in-band gain, the degradation of SNR relative to the input. Each (I and Q) channel showed a NF of 15dB, implying a double-side-band NF of 12dB. Sensitivity measurements showed a sensitivity (BER<10^{-3}) of -94dBm, for a 100kb/s data rate FSK signal with 300kHz tone spacing, consistent with the expected SNR of the demodulator. Reducing bias current in the LNA and IF amplifier tended to degrade this sensitivity, such that in the lowest power mode tested (Vdd = 2.5V, Pdc = 650μW) NF and sensitivity were degraded by 8dB. Strong interferers could act to desensitize the receiver in several distinct ways, including through distortion, reciprocal phase-noise mixing and spurious down-conversion by the IF mixer. Since these effects can work in concert to degrade performance, a generalized desensitization test was performed by supplying the

receiver with a -90dBm signal at 900MHz and the sweeping a single tone interfering signal in 200kHz steps from 865MHz to 935MHz and increasing the power of the interferer until BER was seen to degrade to approximately 10^{-3}. This permitted us to plot the maximum blocker strength vs frequency, and account for all of the different effects at once. The results of this test are shown in figure 2.17. Wide-band interferers desensitized the receiver at power levels of about -13dBm (tested from 800MHz to 1GHz in 10MHz steps), and this desensitization which seemed mostly due to front-end compression, The desensitization level is degraded closer in, reflecting the filtering characteristic of the mixer output and indicating that close-in desensitization is due to compression in the IF amplifier. Desensitization becomes worse at frequency offsets proportional to even multiples of the IF. The worst case is, of course, at the image frequency, where an interferer is equivalent to an interferer at the same frequency as the signal (the demodulator stays accurate for signal-to-interferer levels down to ~5dB, provided the interferer is constant envelope). Other frequencies where desensitization occurs correspond to odd harmonics of the IF. The comparitively mild desensitization at frequencies close to those corresponding to the 3^{rd} and 5^{th} harmonics indicates that the harmonic suppression in the IF mixer is working as expected.

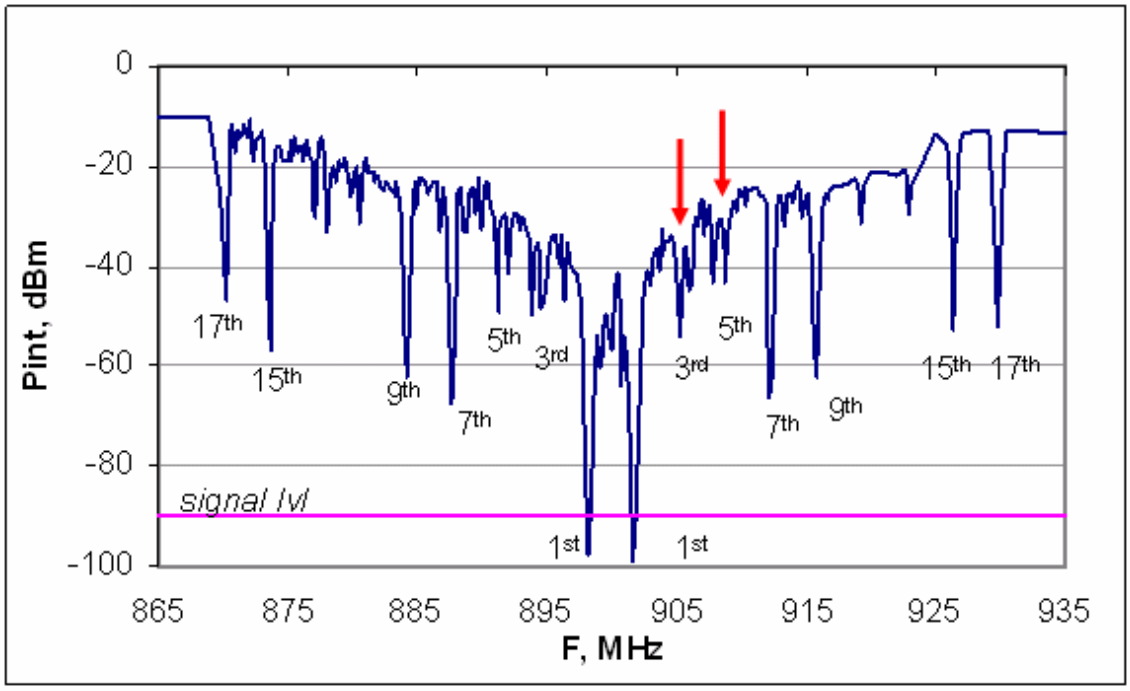


Figure 2.17. Blocker sweep. Single-tone interferer was swept in 200kHz steps and scaled to find the minimum power that generated errors in the demodulation of a -90dBm signal at 902MHz. At large offset, desensitization occurred for interferers of ~12dBm or stronger. This degraded closer in, and at odd multiples of the IF. Note that the 3rd and 5th (and 11th and 13th) harmonics have much less effect indicating that the harmonic canceling mixer is performing its function.

Thus the receiver meets all of our requirements with significant margin exceeding the target sensitivity of -80dBm by about 14dB, implying approximately twice the range indoors. Similarly the wideband desensitization requirement of -22dBm was beaten by almost 10dB, improving the probability that this receiver will withstand interference from any standard wireless device not immediately adjacent to the mote.

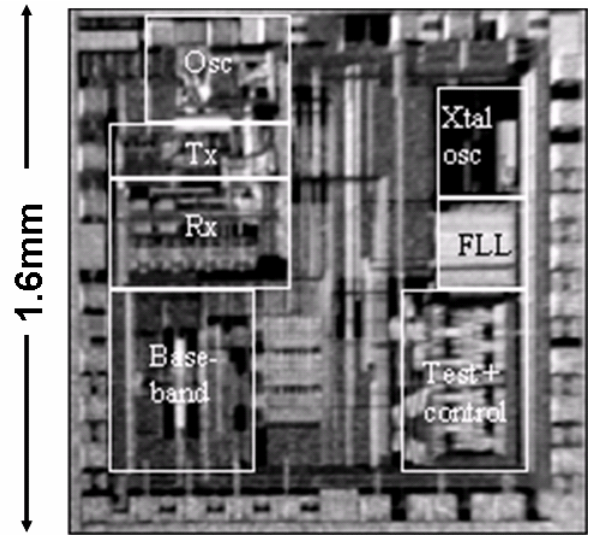


Figure 2.18. Die photo.

System Results

The Layout of the chip is shown in figure 2.18. As can be seen, the RF and IF components of the design required only 0.4mm^2, with the FSK demodulator and FLL requiring an additional 0.4mm2. The receiver and transmitter were each tested in conjunction with the FLL and demodulator, and two such chips were demonstrated passing bits at a rate of 100kbs using FSK with frequency spacing 300kHz. At nominal bias settings a link margin of 87dB was demonstrated while consuming 1.14mW in the receiver and 1.28mW in the transmitter. Each were also demonstrated operating in low-power modes, where a link margin of 74 dB was still possible, but now consuming 0.94mW in the transmitter and 0.65mW in the receiver. Table 2.1 summarizes the break-down of power by functional block. Finally, two radios in the nominal bias setting were demonstrated communicating at a range of 16 meters indoors through 2 concrete walls, confirming that the design was in fact capable of wireless communication at the range and data rate original sought while consuming

the amount of power specified (see figure 2.19). These tests were performed using a minimum of external components: battery, antenna, 32kHz crystal, oscillator inductor, and one series inductor with the antenna.

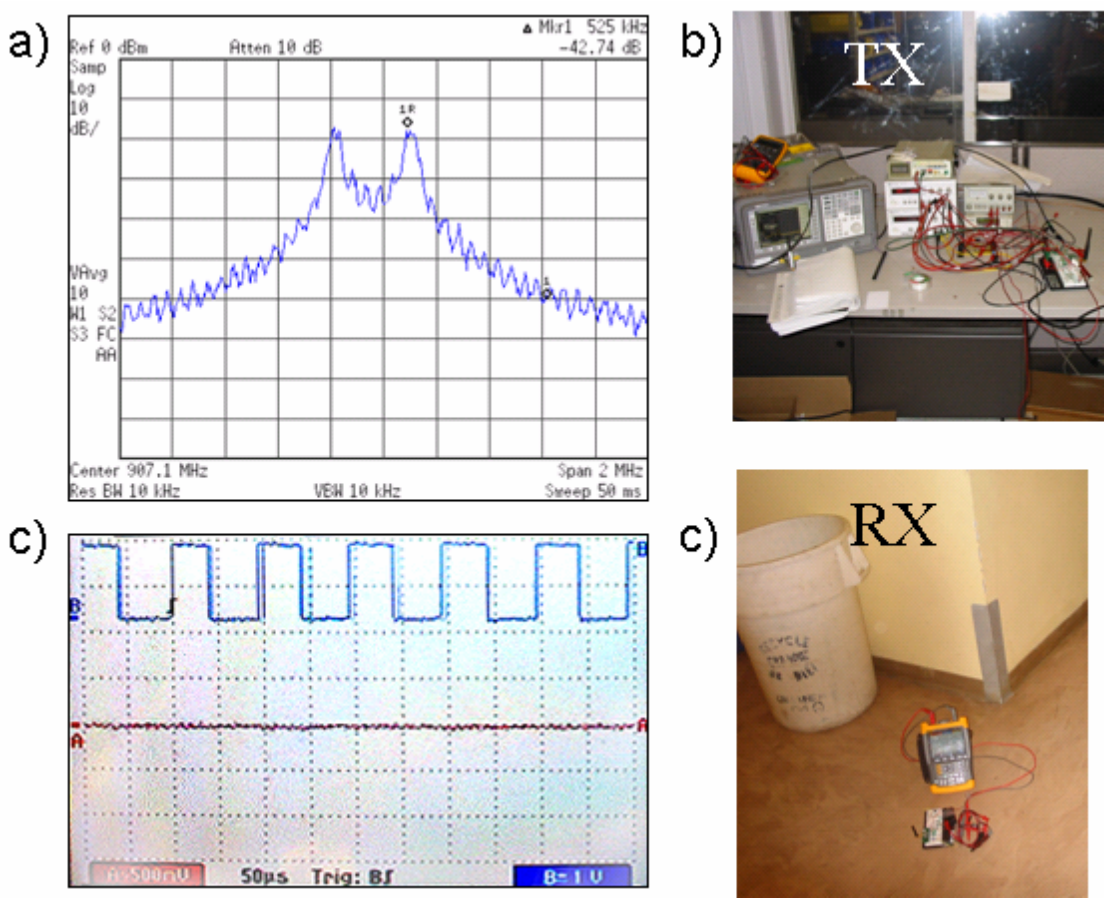


Figure 2.19. Wireless testing. a) Output spectrum of transmitter sending a 10kHz squarewave with 300kHz spacing between 1 and 0 states. b) Photograph of transmitter setup in northwest corner of Cory 476. c) Oscilloscope trace of receiver output demodulating 10kHz square wave at a range of 16 meters through concrete walls. d) Photograph of receiver in the hallway of Cory, 4th floor.

References

[1] M. D. Scott, K. S. J. Pister, and B. E. Boser, "An Ultra-Low Energy ADC for Smart Dust," *IEEE JSSC,* vol. 38, pp. 1123-1129, July 2003.

[2] B. Warneke and K. S. J. Pister, "An Ultra-Low Energy Microcontroller for Smart Dust Wireless Sensor Networks," in *IEEE ISSCC*, 2004, pp. 316 - 317.

[3] A.-S. Porret, T. Melly, D. Python, C. C. Enz, and E. A. Vittoz, " An ultralow-power UHF transceiver integrated in a standard digital CMOS process: Architecture and receiver," *IEEE JSSC,* vol. 36, pp. 452 - 466, March 2001.

[4] J. M. Rabaey, J. Ammer, T. Karalar, S. Li, B. Otis, M. Sheets, and T. Tuan, "Picoradios for wireless sensor networks: The next challenge in ultra-low power design," in *International Solid-State Circuits Conference*, San Francisco, 2002, pp. 200-201.

[5] T. Melly, A.-S. Porret, C. C. Enz, and E. A. Vittoz, "An Ultralow-Power UHF Transceiver Integrated in a Standard Digital CMOS Process: Transmitter," *IEEE Journal of Solid-State Circuits,* vol. 36, pp. 467-472, Mar 2001.

[6] B. Otis and J. M. Rabaey, "A 450μW-RX, 1.6mW-TX Super-Regenerative Transceiver for Wireless Sensor Networks," in *International Solid-State Circuits Conference*, San Francisco, 2005, pp. 396-397.

[7] A. Molnar, "An Ultra-Low-Power 900MHz Radio Transmitter For Wireless Sensor Networks," UC Berkeley, Berkeley Dec 2003.

[8] H. Hashemi, "The Indoor Propagation Channel," *Proceedings of the IEEE,* vol. 81, pp. 943-968, 2004.

[9] B. W. Cook, A. Molnar, and K. S. J. Pister, "Low Power RF Design for Sensor Networks," in *RFIC Symposium*, Long Beach CA, 2005.

[10] R. G. Meyer and W. D. Mack, " A wide-band class AB monolithic power amplifier " *IEEE Journal of Solid-State Circuits,* vol. 24, pp. 7 - 12, Feb 1989.

[11] R. Magoon and A. Molnar, "The Integration of Direct Conversion Transceivers for Modern Cellular Standards " in *RFIC Symposium* Seattle Washington: IEEE, 2002.

[12] R. Magoon, A. Molnar, G. Hatcher, J. Zachan, and W. Rhee, "A single-chip quad-band (850/900/1800/1900MHz) direct-conversion GSM/GPRS RF transceiver with integrated VCOs and Fractional-N synthesizer," *IEEE JSSC,* vol. 37, pp. 1710-1720, December 2002.

[13] A. Molnar, R. Magoon, G. Hatcher, J. Zachan, W. Rhee, M. Damgaard, W. Domino, and N. Vakilian, "A single-chip quad-band (850/900/1800/ 1900MHz) direct-conversion GSM/GPRS RF transceiver with integrated VCOs and Fractional-N synthesizer," in *ISSCC*, San Francisco, 2002, pp. 232, 233.

[14] P. Grey and R. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3 ed. New York: Wiley & Sons, 1993.

[15] J. A. Weldon, R. S. Narayanaswami, J. C. Rudell, L. Lin, M. Otsuka, S. Dedieu, L. Tee, K.-C. Tsai, C.-W. Lee, and P. R. Gray, "A 1.75-GHz highly integrated narrow-band CMOS transmitter with harmonic-rejection mixers," *IEEE Journal of Solid-State Circuits,* vol. 36, pp. 2003 - 2015, Dec 2001.

[16] G. Ahlbom, L. Egnell, and C. Wickman, "Digital demodulator for optical FSK signals," *Electronics Letters* vol. 26 pp. 290-292, Mar 1990.

Chapter 3

A High Compliance CMOS Current Mirror

Abstract

A simple modification to a standard CMOS Current mirror is presented, which significantly reduces the amount of drain-to-source voltage required across its output to maintain constant output current. The design uses two additional transistors to replicate the output voltage on the reference transistor, matching both VGS and VDS between the devices. The design was demonstrated to reduce the required headroom on the output by more than a factor of 2, and to increase output resistance by more than a factor of 5. This mirror is especially useful in low power applications where bias voltage is at a premium.

Introduction

One of the most common building blocks in analog integrated circuits is the current mirror. Current mirrors are commonly used anywhere that a DC current source is required. Current mirrors are constrained by requirements on noise, accuracy, output impedance, and required voltage headroom (compliance). Unfortunately these requirements typically must be traded off against each other: noise, accuracy and impedance are improved by increasing the V_{DSAT} of a CMOS mirror, but V_{DSAT} sets the minimum voltage that must be available across the mirror's output to maintain constant current, and so sets its minimum power dissipation. Various circuits have been proposed to improve basic current mirror performance[1]. In particular it has been suggested that by replicating the output voltage of the mirror on the drain of the mirror's reference device[2]. Here we present a simple circuit, wherein by the inclusion of 2 additional transistors and an additional reference current, a standard CMOS current mirror's voltage range can be extended well below V_{DSAT} while maintaining a constant current and without degrading noise or matching.

The basic circuit is shown in Figure 1. Transistors M1 and M2 form a standard 1:N ratioed current mirror, so that $L_{M2} = L_{M1}$, $W_{M2} = N \cdot W_{M1}$, sharing their gate and source nodes, and so have the same V_{GS}. Two additional transistors, M3 and M4 act to replicate the drain voltage on M2 and present it to M1's drain. The result is that M1 and M2 are in exactly the same bias state, and so maintain drain currents in a 1:N ratio over a wide range of bias states, including both saturation and deep triode.

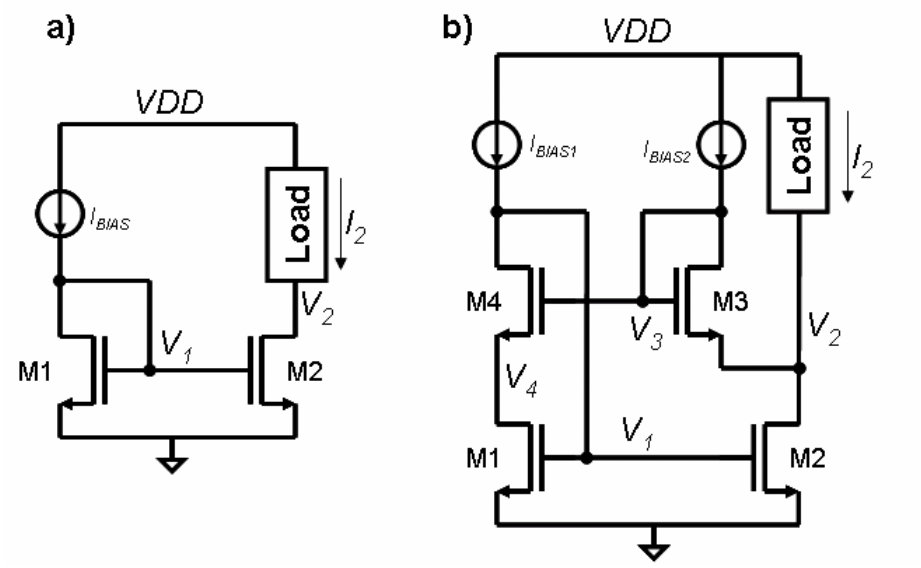


Figure 3.1 basic circuit topology. a) Typical CMOS current mirror. b) Enhanced mirror described here.

First-Order Large Signal DC Analysis

Transistors M1 and M2 share gate and source nodes. The drain node of M2 is connected to the source of M3, and M3 is diode connected, biased by I_{BIAS2}. If M3 and M4 are sized such that $V_{GS4} = V_{GS3}$, and M4 is in saturation, M4 acts as a source-follower and $V_4 \approx V_2$. This is best accomplished by scaling M3 and M4 to the same length, and scaling their widths such that $W_{M3}/W_{M4} = I_{BIAS2}/I_{BIAS1}$. The value of V_1 is set by negative feedback: $I_{D4} = I_{D1}$, and any difference between I_{D4} and I_{BIAS1} will tend to charge or discharge the gate capacitances of M1 and M2, shifting V_1 and so adjusting I_{D1} until it balances I_{BIAS1}.

When $V_2 < V_{DSAT*}$, defining $V_{DSAT*} = (I_{BIAS1}/k_1)^{1/2}$, the overdrive voltage on M1 and M2 when in saturation, M2 and so M1 are in triode, and their behavior can be approximated using a square-law model:

$$I_{D2} = Nk_1\left(V_1 - V_{TH} - V_2/2\right)V_2 \qquad \text{(eq. 3.1)}$$

$$I_{D1} = k_1\left(V_1 - V_{TH} - V_4/2\right)V_4 \qquad \text{(eq. 3.2)}$$

Since $V_4 = V_2$, when V_1 settles to a state such that $I_{D1} = I_{BIAS1}$ and $I_{D2} = N \cdot I_{BIAS1}$, then V_1 is set:

$$V_1 = V_{TH} + \frac{V_2}{2} + \frac{V_{DSAT*}^{\ 2}}{2V_2} \qquad \text{(eq. 3.3)}$$

Thus, feedback causes V_1 to increase to compensate for decreased V_2. The circuit must ultimately fail when V_1 approaches Vdd and so saturates the current source I_{BIAS1}. As will be discussed below, however, second-order effects become dominant well before this failure mode.

Under conditions where $V_3 - V_1 < V_{TH}$, M4 enters triode, and V_4 ceases to track V_2. Provided $\sqrt{I_{BIAS2}/k_3} < V_{TH}$, however, M1 and M2 will already be saturated and so relatively insensitive to changes inV_2. In intermediate cases, where M1, M2 and M4 are all in saturation, the circuit acts to increase the output resistance of the mirror.

Although these equations assume ideal square-law behavior, qualitatively similar behavior can be expected under other operating conditions such as velocity saturated conditions. Furthermore, optimization of noise and matching in current mirror usually requires the use of long-channel transistors in deep inversion, where square-law behavior is typically observed.

Small signal analysis

The small signal model of this circuit is shown in Fig 2. The overall circuit behavior can be translated to an effective output impedance relating V_2 to I_2. If we assume M3 and M4 act as an ideal voltage replica between V_2 and V_4 we can derive a comparatively simple relationship between V_2 and I_2:

$$\frac{I_2}{V_2} = \frac{j\omega N(N+1)(j\omega C_{GD1}C_{GS1} + gm_1 C_{GD1} + g_{DS1}C_{GD1} + g_{DS1}C_{GS1})}{gm_1 + j\omega((N+1)C_{GS1} + NC_{GD1})}$$

This is equivalent to the RC structure shown in figure 2a where

$$C_1 = (N^2 + N)\left(C_{GD1} + \frac{g_{DS1}}{gm_1}(C_{GD1} + C_{GS1})\right) \quad \text{(eq. 3.4)}$$

$$R_s = \frac{NC_{GD1} + (N+1)C_{GS1}}{N(N+1)(gm_1 C_{GD1} + g_{DS1}C_{GD1} + g_{DS1}C_{GS1})} \quad \text{(eq. 3.5)}$$

$$C_2 = \frac{NC_{DS1}C_{GS1}}{NC_{DS1} + (N+1)C_{GS1}} \quad \text{(eq. 3.6)}$$

Although this simplification is imperfect, it can be seen that the main positive feedback loop only operates for frequencies less than $1/(R_sC_1) \sim \omega_T/N$, where ω_T is the unity current gain frequency for M1 and M2, implying that delays from M3 and M4 are only relevant if they have strong effects below this frequency. Since the biggest capacitances associated with those transistors (C_{GS3} and C_{GS4}) operate to increase the coupling between V2 and V4, these capacitances should, if anything, act to enhance the operation of the circuit at high frequencies, rather then introducing delay. In the ideal case, R_p, the low frequency output impedance, is infinite, but as will be shown below, this value actually depends upon a variety of second order effects.

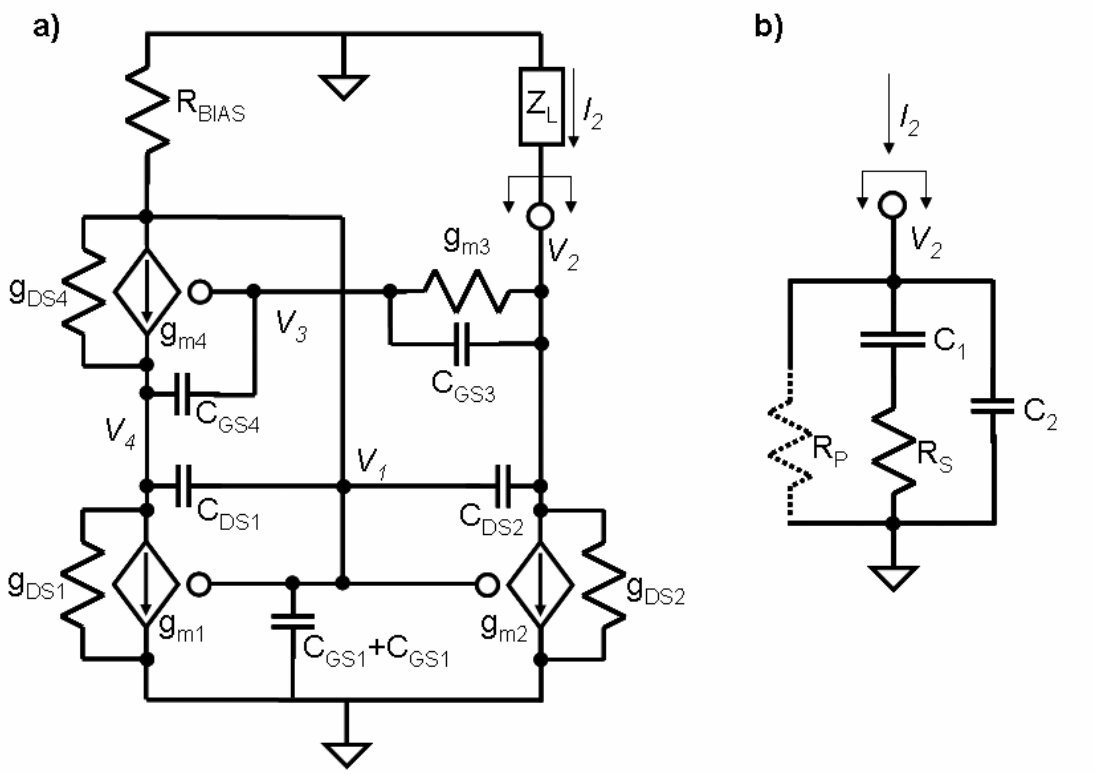


Figure 3.2. Small signal model of circuit enhanced current mirror. a) Small signal model of full transistor circuit. b) Model of equivalent output impedance, assuming $V_4 \approx V_2$.

Noise

Noise in the enhanced mirror is not substantially worse than it would be if it was unenhanced, for V_2>Vdsat*, with only a slight degradation due to noise from I_{bias2}. As V_2 decreases below Vdsat*, V1 increases, increasing channel conductance and thermal current noise at the drains of M1 and M2. Since thermal current noise is roughly proportional to channel conductance, which is roughly proportional to V_1-V_{TH}, i_d^2 increases roughly proportional to V_1-V_{TH} [3]. Flicker noise, in contrast, will tend to decrease as V_{DS} decreases[4] even if I_D is held constant. Thus, noise changes as V_2 is decreased and V_1 correspondingly increased, but this change depends upon

the noise mechanism, and this change is gradual, depending upon the degree to which the mirror is driven into triode.

Second Order Effects: mismatch and ro

Mismatch between transistors can be well approximated as being completely due to V_{TH} mismatch. Since the two important matching parameters are between M1and M2, and between M3 and M4, we can lump these mismatches into two voltage deviations, $\Delta V_2 = V_{TH2} - V_{TH1}$ and $\Delta V_4 = V_{TH4} - V_{TH3}$. The effect of ΔV_2 is to change Io by $\Delta Io = \Delta V2 \cdot N \cdot gm_1$, exactly as in an unenhanced mirror. ΔV_4 has the effect of slightly mismatching V_{DS1} relative to V_{DS2}. This effect becomes proportionally stronger as V_2 decreases. Combining these effects we find that ΔIo depends on V2, ΔV_4, and ΔV_2:

$$\frac{\Delta Io}{Io} = \Delta V_2 \frac{2V_2}{V_{DSAT*}^{\ 2}} + \Delta V_4 \left(\frac{1}{V_2} - \frac{V_2}{V_{DSAT*}^{\ 2}} \right) \qquad \text{(eq. 3.7)}$$

A DC dependence of Io on V2 implies finite real impedance at the output, shown as R_p in Figure 2b. By taking the derivative of ΔIo with respect to V_2 we can find the equivalent conductance due to mismatch:

$$g_{out} = Io \left(\frac{2\Delta V_2}{V_{DSAT*}^{\ 2}} - \frac{\Delta V_4}{V_{DSAT*}^{\ 2}} - \frac{\Delta V_4}{V_2^{\ 2}} \right) \qquad \text{(eq. 3.8)}$$

Since ΔV_4, and ΔV_2 can be either positive or negative, the conductance of the output can also be either sign. Other forms of mismatch, which can be described as mismatch in k, can be incorporated into this same framework: mismatch between k_3 and k_4 will shift V_{GS3} relative to V_{GS4}, and as such can be seen as simply contributing

to the value of ΔV_4. Mismatch between k_1 and k_4 acts as a change in the mirror ratio, and has no additional effects on g_{out}.

The output resistance ($1/g_{DS4}$) of M4 can also influence the effective output resistance of the mirror. When $V_{DS4} = V_1 - V_2$ is not equal to V_{DS3}, the output resistance of M4 generates an effective voltage mismatch of

$$\Delta V_{4R} = \frac{V1 - V2 - V_{GS3}}{r_{O4} gm_4} = \frac{V_{TH} - V_{GS3} - V_2/2 + {V_{DSAT*}}^2/2V_2}{r_{o4} gm_4} \qquad \text{(eq. 3.9)}$$

Since $V_{GS3} > V_{TH}$, this term will tend to take on negative values when M2 first enters triode, and then become increasingly positive is V_2 approaches zero. This results in an output conductance of

$$g_{out} = \frac{-Io}{r_{o4} gm_4} \left(\frac{V_{GS3} - V_{TH}}{{V_2}^2} + \frac{V_{GS3} - V_{TH}}{{V_{DSAT}}^2} + \frac{V_2}{{V_{DSAT}}^2} - \frac{{V_{DSAT}}^2}{{V_2}^3} \right) \qquad \text{(eq. 3.10)}$$

Which also takes on negative values as V_2 first decreases below V_{DSAT*}, but then becomes positive as V_2 continues to decrease. This conductance can be modeled as being in parallel with the conductance due to mismatch.

Finally the output resistance of the current source I_{BIAS1} is reflected to the output as a function of V_1, and results in an increasing output conductance as V_2 decreases:

$$g_{out} = -\frac{N}{R_{BIAS}} \frac{dV_1}{dV_2} = \frac{N}{2R_{BIAS1}} \left(\frac{{V_{DSAT*}}^2}{{V_2}^2} - 1 \right) \qquad \text{(eq. 3.11)}$$

Each of these effects on output impedance (ΔV_2, ΔV_4, r_{o4}, R_{BIAS1}) appears in parallel. Since both r_{o4} and R_{BIAS1} lead to decreased output impedance (and so decreased DC current) as V2 decreases, it is these terms that ultimately limit the

operating range of the enhanced current mirror. For reliable operation, furthermore, this operating range will be further limited by the worst case values of ΔV_2 and ΔV_4. Fig 3 shows simulation results for Io and g_{out} over various values of ΔV_2 and ΔV_4.

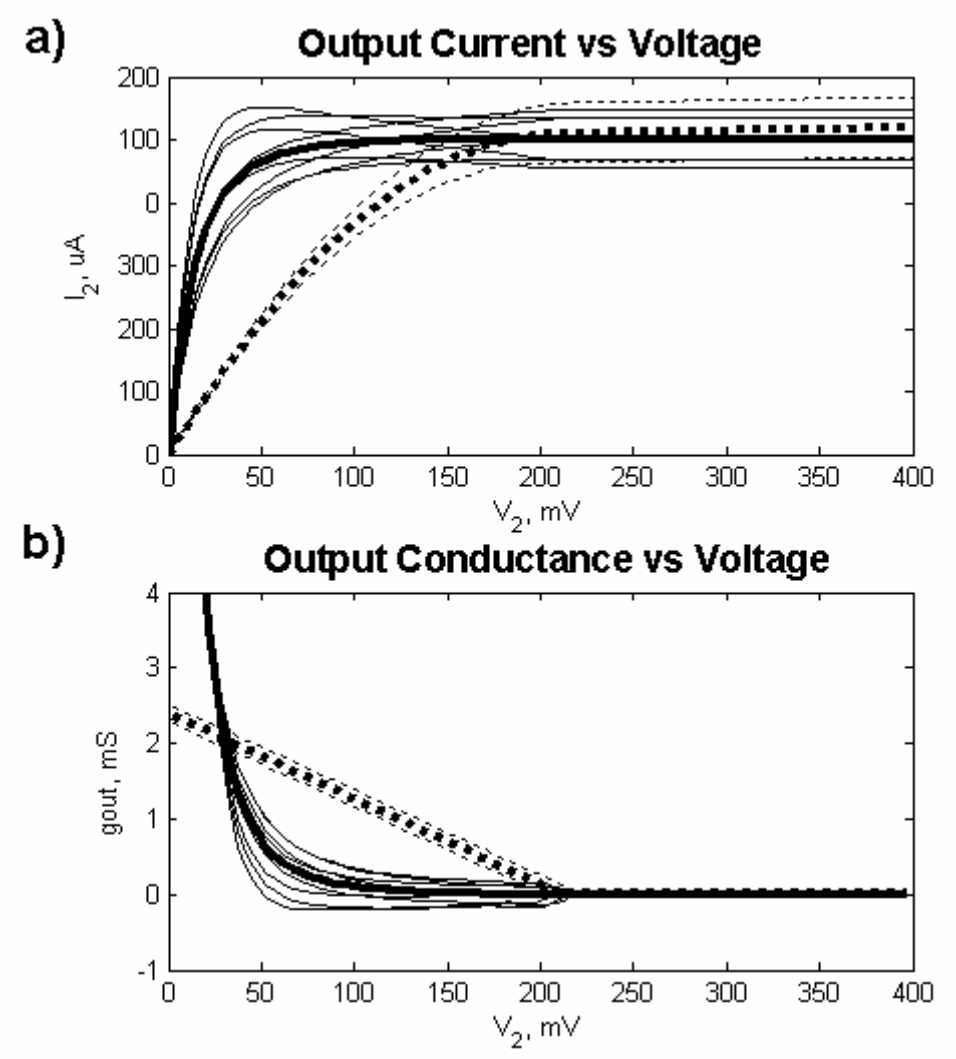


Figure 3.3 Simulated a) I-V and b) g-V curved of enhanced (solid lines) and unenhanced (dotted lines) current mirrors over various values of ΔV_2 and ΔV_4 (ΔV_{2RMS} = ΔV_{2RMS} = 10mV).

Because ΔV_2 and ΔV_4 can each take positive or negative values, and ΔV_{4R} takes on negative values for some values of V_2, it is possible for the DC output resistance of the enhanced mirror to take on negative values. This can potentially lead

to instability, but only if the magnitude of the real part of the load impedance presented to the mirror is larger than the magnitude of this effective negative output resistance. Specifically, if the load resistance meets the requirement:

$$R_L < \frac{1}{gm_2} \frac{V_{DSAT*}}{|\Delta V_2| + |\Delta V_4| - \Delta V_{4R}} \qquad \text{(eq. 3.12)}$$

then stability will be guaranteed. For most biasing situations, such as generating tail currents for differential pairs, this requirement is easily met, since $V_{DSAT*} >> |\Delta V_2| + |\Delta V_4|$, and R_L is often on the order of $1/gm_2$. In other scenarios, however, this requirement must be checked.

Measured results

Figure 4 shows measured results from a circuit of the type shown in Figure 1b. In the circuit implemented, N = 100, V_{DSAT} = 210mV, Vdd = 1.5V. The circuit also included a switch that could be closed across the source and drain of M4, causing the mirror to revert to its un-enhanced state (equivalent to Figure 1a). Output current vs V_2 are shown for both enhanced and basic (unenhanced) conditions. As can be seen, the mirror maintains a roughly constant DC current to much lower voltages in the enhanced state than in the un-enhanced state. For example DC current drops by 5% at 210mV in the basic state, but does not reach the same level in the enhanced state until Vout reaches 90mV, more than a factor of two improvement. Across this range, the output resistance averages about 15kΩ when the mirror is enhanced, versus 2.5kΩ when basic. At voltages greater than V_{DSAT}, the output resistance is also higher for the enhanced state, averaging 200kΩ, as compared to 28 kΩ. As expected, when V_2 increases beyond V_{GS2} the difference in output impedance disappears.

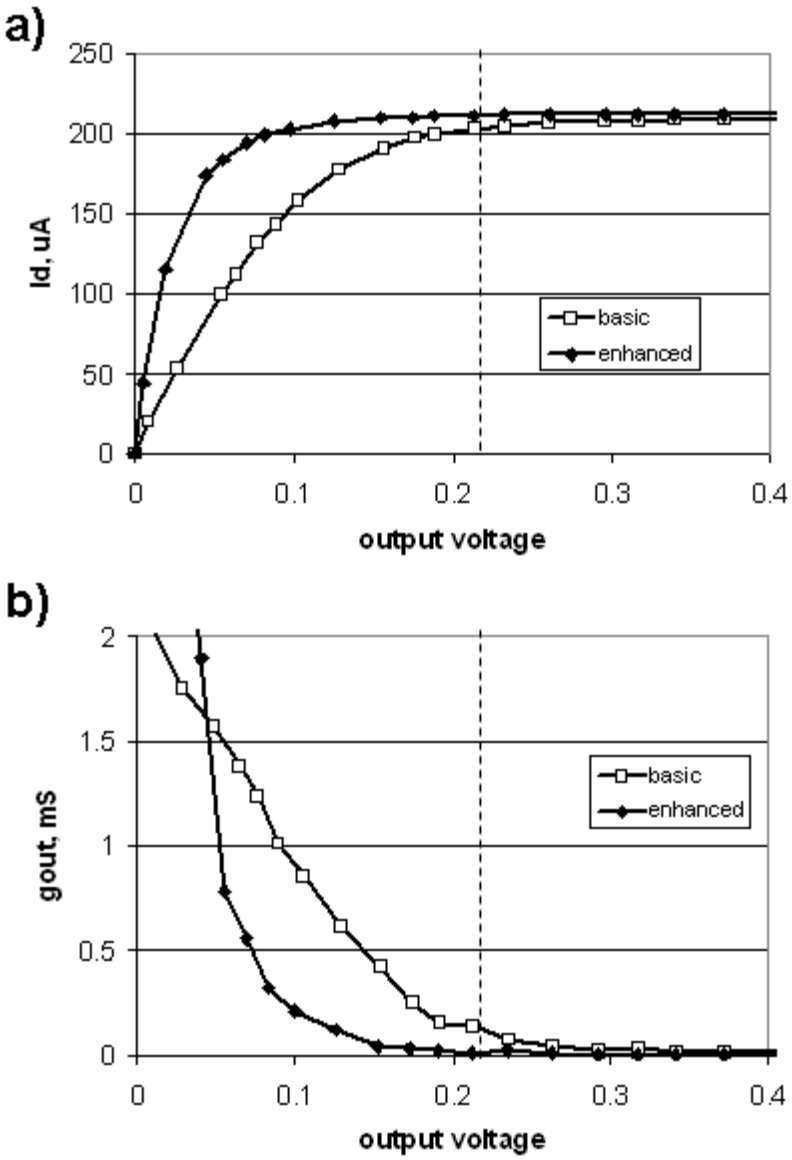


Figure 3.4 Measured results for a current mirror in enhanced and basic (unenhanced) modes of operation. a) DC current output versus output voltage (vertical line indicates approximate V_{DSAT}). b) Output conductance versus output voltage. Note that the enhanced state shows dramatically lower conductance from voltages well below V_{DSAT} to well above.

Conclusion

We present here a very simple enhancement on the standard CMOS current mirror that extends its operational range well below the Vdsat of the devices and at very little cost in terms of power, area and complexity. This circuit has been implemented as part of several low power circuits, including differential pairs, and self biased CMOS circuits (similar to current-starved ring oscillators)[5]. Use of this mirror topology has reliably improved operating ranges for battery voltages by more than 100mV.

References

[1] Charlon and W. Redman-White, "Ultra High-Compliance CMOS Current Mirrors for Low Voltage Charge Pumps and References," 2004.

[2] M. QuaranteNi, M. Poles, M. Pasotti, and P. Rolandi, "A HIGH COMPLIANCE CMOS CURRENT SOURCE FOR LOW VOLTAGE APPLICATIONS," in *ESSCIRC*, 2003, pp. 425-428.

[3] B. Wang, J. R. Hellums, and C. G. Sodini, "MOSFET Thermal Noise Modeling for Analog Integrated Circuits," *IEEE JOURNAL OF SOLID-STATE CIRCUITS,* vol. 29, pp. 833-835, July 1994.

[4] K. K. Hung, P. K. Ko, C. M. Hu, and Y. C. Cheng, "A Physics-Based MOSFET Noise Model for Circuit Simulators," *IEEE Transactions on Electron Devices,* vol. 37, pp. 1323-1333, May 1990.

[5] A. Molnar, B. Lu, S. Lanzisera, B. Cook, and K. Pister, "An Ultra-Low Power 900MHz RF Transceiver for Wireless Sensor Networks," in *CICC* Orlando FL: IEEE, 2004.

Chapter 4

Inhibitory Feedback to ON and OFF Bipolar Cells is Asymmetric in the Rabbit Retina

Abstract

The diverse morphological types of bipolar cell are divided into two general types, ON and OFF cells, defined by the polarity of their light elicited excitation, but the types of inhibitory input these two classes of cells recieve are very different. In OFF bipolar cells, inhibition is maximum at ight ON, so that the excitatory and inhibitory currents respond to light by adding in-phase with each other, enhancing the OFF voltage response. In rod bipolar cells, a known, distinct class of ON bipolar cell, excitatory and inhibitory currents both increased at light ON, and so added out-of-phase, such that inhibition suppresses the ON voltage response at all time scales. In half of ON cone bipolar cells the inhibitory currents are in-phase with excitation. In most of the remaining ON cone bipolar cells inhibitory currents are out-of-phase with excitation,

but with a slight delay, causing these cells to preferentially respond to transient stimuli. In-phase inhibition is carried by glycinergic amacrine cells spanning the ON and OFF sublaminae, while out-of-phase inhibition was carried by GABAergic amacrine cells within the ON sublamina. Thus while the ON and OFF bipolar cells each receive inhibition from the other sublamina, only the ON bipolar cells receive inhibition from within the same sublamina, a clear asymmetry in the flow of feedback inhibition in the retina. The excitatory-inhibitory interactions we have found in bipolar cells reflect similar interactions reported elsewhere in the retina and in higher visual centers.

Introduction

Retinal bipolar cells receive excitation from photoreceptors and provide excitation to a wide variety of inhibitory amacrine cells and to retinal ganglion cells, which in turn carry visual information to the brain. Mammalian retinal bipolar cells show considerable morphological and physiological diversity; MacNeil et al.[1] have distinguished 12 morphological types of bipolar cell by the stratification and morphology of their synaptic terminals. Bipolar cells also display a variety of excitatory receptors and synaptic morphologies that lead to diverse physiological responses. At the most basic level, ON and OFF bipolar cells (which depolarize to increments and decrements in light intensity, respectively) are distinguished by their excitatory receptors: OFF bipolar cells contain ionotropic, sign-preserving AMPA and kainite receptors, while ON bipolar cells contain sign-inverting metabotropic glutamate receptors. At a finer level, DeVries [2] has shown that distinct ionotropic glutamate receptors in OFF bipolar cells generate temporally distinct excitatory inputs

from photoreceptors. Bipolar cells also receive inhibitory inputs from, and deliver excitation to, a variety of amacrine cell types [3-5]. These feedback connections create an additional level of physiological diversity in bipolar cells [6, 7]. Because retinal bipolar cells are the first stage of the visual pathway with significant diversity, understanding bipolar cell physiology is a prerequisite to gaining insight into the origin of the broad range of signals that the retina sends to higher visual centers [8]. In this study, we attempted to characterize the interactions between excitation and inhibitory feedback that shape bipolar cell light responses. Our measurements show that the responses of all 12 morphological bipolar cell types are shaped by only four different forms of excitatory-inhibitory interaction, and that these interactions are distinct between the ON and OFF pathways. Application of a variety of specific inhibitory blockers revealed that these interactions are pharmacologically distinct. The inhibition that flows *between* the ON and OFF sublaminae is glycinergic and, surprisingly, enhances rather than suppresses the bipolar voltage response. The inhibition that flows *within* the ON sublamina, is GABAergic, and suppresses voltage response. We found very little inhibition that flowed within the OFF sublamina, revealing a striking asymmetry between the ON and OFF pathways of the retina.

Results: Excitation and inhibition interact in 4 distinct ways

Physiological and morphological identification of basic bipolar cell types.

Whole cell patch clamp recordings were made from 169 bipolar cells and each cell was classified as ON or OFF based on excitatory responses to high-contrast light and dark flashes. Of these, 60 were identified as OFF cells and 105 as ON cells based upon morphology and the sign of their responses (see methods). Among the ON cells, 55 were identified as rod bipolar cells based on of the location and shape of their axon terminals, which characteristically lie along the proximal border of the IPL [1]; the remaining 50 were identified as ON cone bipolar cells. An additional 93 bipolar cells were imperfectly patched, yielding incomplete data, often consisting of only the voltage response of the cell. Data from these imperfect cells were consistent with those found in more complete whole-cell recordings. We performed the majority of our experiments using a potassium-based intracellular solution (see methods), but we performed 60 of 169 experiments using a cesium-based solution, in order to confirm that our results were not simply a consequence of imperfect voltage clamping due to potassium leakage currents. The same forms of excitatory and inhibitory currents were found while using either solution, leading to similar voltage responses. These two data sets have been pooled in all subsequent analyses.

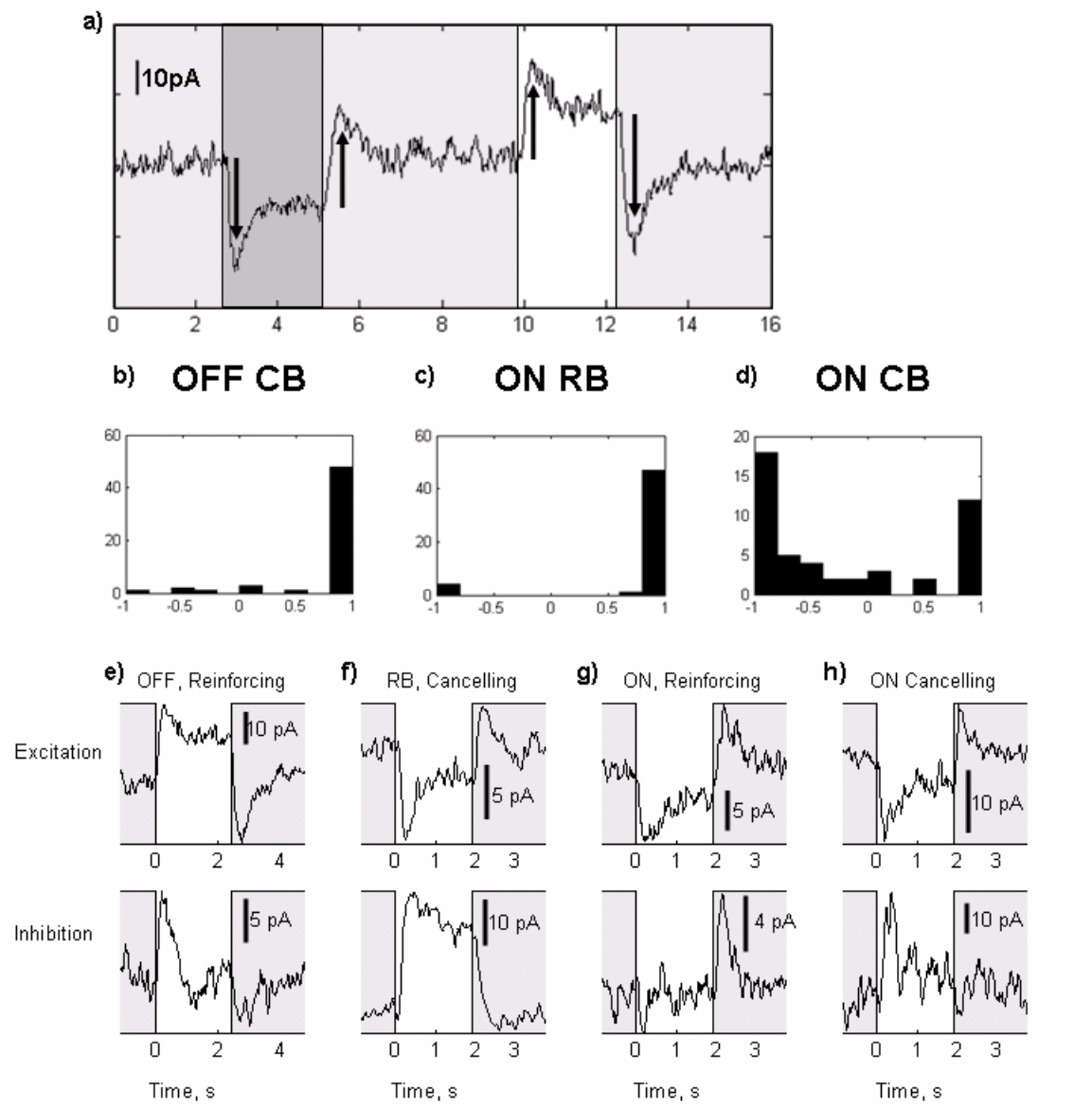


Figure 1. Examples of basic interactions between excitation and inhibition, as revealed by ±100% contrast flashes. a) Basic excitatory response of an OFF cell to dark and light flashes, arrows indicate transitions used to identify polarity. b-d) Histograms of inhibitory polarity for the three general classes of bipolar cell. Note that while both OFF and rod bipolar cells show a strongly dominant ON inhibition with only a few outliers, ON cone bipolar cells, show large numbers of cells with ON and OFF inhibition. Columns e-h): example excitatory and inhibitory traces from each class of cell: e) OFF cone bipolar cell: Excitation decreased (became less

negative) at light onset and rebounded at light offset; inhibition increased (became more positive) at light onset and decreased at light offset, these inputs were in-phase and so act to reinforce each other . Column f) Rod Bipolar: Both excitation and inhibition increased at light onset and decreased at offset, acting to suppress one another. Column g) ON cone bipolar: Both excitation and inhibition were inward, a reinforcing interaction. Column g) ON cone bipolar: Excitation was inward; inhibition was outward, and so canceled excitation.

All OFF cone bipolar cells receive in- phase, reinforcing inhibition

OFF bipolar cells showed a common response form: at the onset of a dark flash, excitation (an inward current) increased, and inhibition (an outward current) decreased; at the onset of a bright flash, excitation decreased and inhibition increased. Thus, the time courses of excitation and inhibition were in-phase: both currents become more inward at light OFF and more outward at light ON, as shown in Fig. 1a. This in-phase, reinforcing interaction appeared in 53/60 OFF cells. Of the remaining cells, 2 received no inhibition and 5 received ON-OFF inhibition, with ON inhibition dominating in 3/5 cases.

Rod bipolar cells receive canceling inhibition

All but 3 of the 55 rod bipolar cells recorded in whole-cell patch showed a common interaction: both excitation and inhibition *increased* in response to bright flashes, and both *decreased* in response to dark flashes. In these cells inhibitory currents were out-of-phase with excitatory currents and therefore cancelled excitation, acting to suppress the voltage response as shown in Fig 1b.

In cases where, resting potentials close to -35mV were recorded excitation dominated voltage response. In 17/55 cases, the initial magnitude of excitation to rod bipolar cells was much larger than the magnitude of inhibition but ran down over a period of approximately 5 minutes. Such rapid run-down was not seen in cone bipolar cells except immediately preceding cell death.

ON cone bipolar cells receive either reinforcing or canceling inhibition

ON cone bipolar cells showed two distinct interactions. In 27/50 ON cells, we measured an in-phase interaction between excitation and inhibition similar to that described above for the OFF cells: bright flashes increased excitation and decreased inhibition while dark flashes decreased excitation and increased inhibition.

In a separate set of ON cone bipolar cells (14/50), both excitation and inhibition increased in response to bright flashes and decreased in response to dark flashes (see Fig. 1d). Here, the inhibitory currents cancelled excitatory currents similar to the interaction measured in rod bipolar cells.

Nine ON cone bipolar cells showed little or no inhibition in response to light stimuli, but showed a significant amount of inhibitory noise.

Thus, while virtually every OFF bipolar cell receives the same kind of (ON) inhibition, ON bipolar cells receive a wider array of inhibitory inputs, with different types of inhibition seen both between ON rod and cone bipolar cells, but also among ON cone bipolar cells.

Sinusoid responses reveal the same 4 distinct interactions

Flash responses like those shown in Fig 1 fail to resolve fine temporal detail. To better characterize the temporal aspects of bipolar cell responses, we stimulated 153/169 preparations with stripes whose intensity varied sinusoidally in time as shown in Fig 2a. Looking at the fundamental frequency response, we characterized the amplitude of each cell's excitatory and inhibitory currents and voltage as a function of temporal frequency, as shown in Fig 2b. In order to analyze the interaction between excitation and inhibition at different temporal scales, we calculated the relative phase of their responses, with 0 degrees corresponding to reinforcement, and 180 degrees corresponding to cancellation (see Fig 2c-e).

All measured responses showed a significant reduction in amplitude at frequencies above 10 Hz, probably indicating a roll-off in the underlying signals generated in the outer retina. Subsequent analyses were only performed on frequencies below 10Hz. We primarily analyzed the fundamental frequency response, since this was sufficient to explain 90% of the periodic response in more than 90% of the cells. Higher harmonics were also investigated, but did not yield additional insight into the basic interactions between excitation and inhibition.

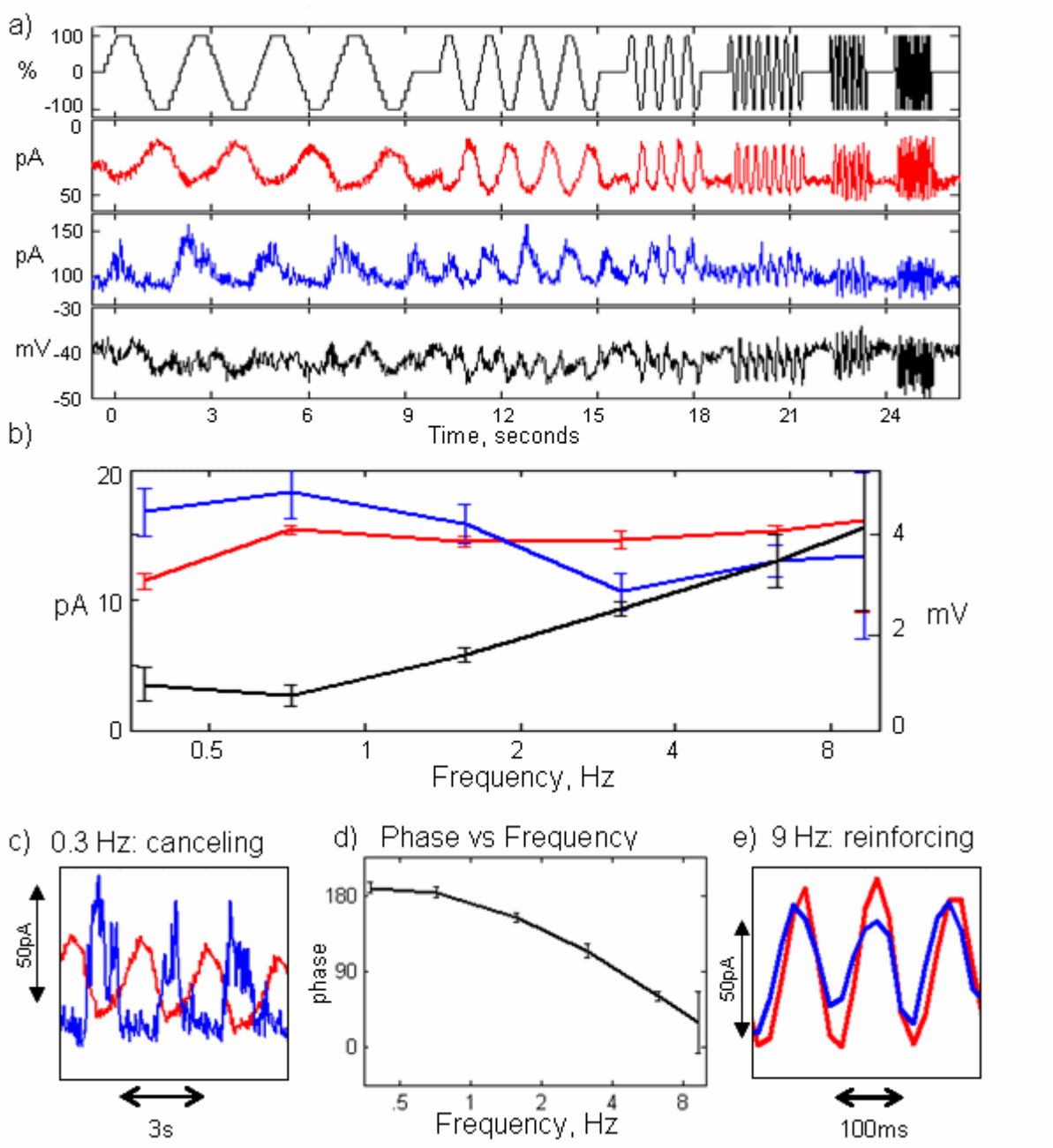


Figure 2. Sinusoidal responses of an ON cone bipolar cell. a) Basic recordings: stimulus intensity (top row, black) as a percent deviation from background, excitatory (red) and inhibitory (blue) currents and voltage response (bottom, black). b) Amplitude of response for excitation, inhibition and voltage (same color scheme as in a) plotted versus stimulus frequency, note that while excitation and inhibition respond

at all frequencies, voltage response is strongest at high frequencies. Error bars indicate estimated noise in response (see methods) c, e) Expanded scale overlay of current inputs at highest and lowest frequencies; note that the currents cancel at low frequencies (c), but reinforce at high frequencies (e). d) Calculating the phase difference between excitation and inhibition at each frequency yields a curve showing a gradual phase shift from 180 to 0 degrees.

OFF cone bipolar cells showed excitatory and inhibitory activity significantly above noise levels at all frequencies below 10 Hz. The phase difference between excitatory and inhibitory input currents remained close to 0 degrees at all frequencies for all OFF cells (see Fig 3a, 5b). Thus, inhibition acted to reinforce the excitatory response over a wide range of temporal scales.

Rod bipolar cells showed a strong low-pass response peaking at 1 Hz and rolling off sharply at higher frequencies (see Fig 3b). This form of frequency response was present in both excitation and inhibition, although inhibition tended to be more strongly biased toward lower frequencies than excitation. The phase relationship between excitation and inhibition was consistent across all frequencies, maintaining a constant 180-degree phase shift as shown in Fig 3b and Fig 5b. Hence, inhibition acts to oppose the excitatory response of rod bipolar cells at all temporal scales.

For those ON cells identified as having reinforcing inhibition to flashed stripes, the phase relationship between excitation and inhibition was consistently close to 0 degrees at all frequencies (Fig 3c, 5d). In those ON cone bipolar cells where inhibition cancelled excitation, the sinusoidal responses were more complex. Excitation and inhibition typically responded in a broad-band fashion, however, their phase

relationship changed dramatically across frequencies. At low frequencies, excitation and inhibition were roughly 180 degrees out-of-phase as shown in Fig 2c, and so the signals cancelled, resulting in a weak voltage response. At high frequencies, the phase relationship shifted to nearly 0 degrees as shown in Fig 2e, so that excitation and inhibition reinforced one another, resulting in a stronger voltage response. This phase changed smoothly with frequency, and can be well approximated by a simple time delay of 50 ms. This delayed cancellation resulted in a characteristically strong high-pass voltage response in 9 of 14 ON cone cells that showed canceling inhibition to flashes. As can be seen in Figs 2b, 3d and 5b, the voltage response was much more high-pass than either excitation or inhibition.

In summary, we measured 4 general types of excitatory-inhibitory interaction in bipolar cells: 1) reinforcement in all OFF cells 2) reinforcement in approximately half of ON cone bipolar cells; 3) cancellation, which appears in all rod bipolar cells, and 4) delayed cancellation, which appears exclusively in a subset of ON cone bipolar cells.

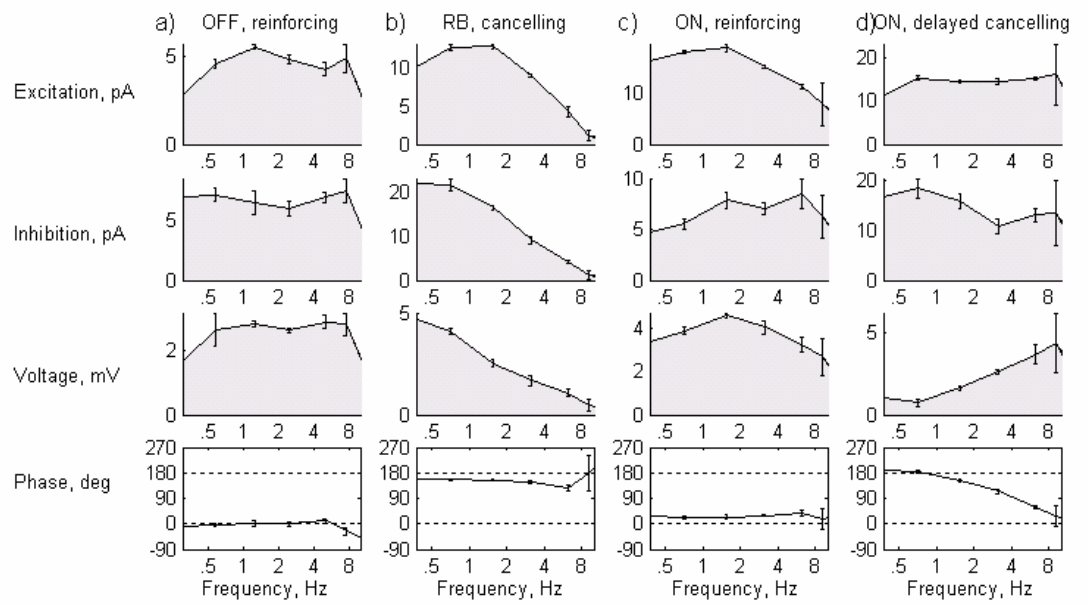


Figure 3. Examples of typical frequency responses for the 4 interaction types: Column a) OFF cell: excitation, inhibition and voltage show response at all frequencies below ~10Hz, the phase difference between excitation and inhibition stays close to 0 degrees at all frequencies, indicating reinforcement at all time scales where the cell responds. Column b) rod bipolar cell: excitation, inhibition and voltage all show much stronger responses at frequencies below 3Hz than above; the phase relationship between excitation and inhibition stays close to 180 degrees, indicating cancellation at all time scales. Column c) ON cone bipolar showing reinforcement, excitation inhibition and voltage respond to all frequencies below ~10Hz, and phase stays close to 0 degrees. Column d) ON cone bipolar showing cancellation; excitation and inhibition show responses to all frequencies, but their relative phase is variable, canceling at low frequencies and reinforcing at high frequencies. This results in a voltage response that is much stronger at high frequencies than at low frequencies.

Bipolar cells of various morphologies show the same basic interactions

We imaged each bipolar cell after recording and compared morphologies for each of the different types of interaction. Although we could not find a one-to-one correspondence between every cell we imaged and morphological classes described elsewhere, we did find individual examples that corresponded closely to each class described by MacNeil et al [1]. These examples appear in Fig 4a, along with tentative classification according to the system of MacNeil et al [1]. We also characterized each cell according to its axonal morphology, measuring three parameters: the depth of the axon terminal, its lateral width and its vertical spread within the IPL, as shown in Fig 4b. The distribution of morphological parameters is similar to that reported previously by MacNeil et al [1] leading us to believe that our dataset encompasses most, if not all of the major morphological types. As reported previously [5] the axon terminals of ON and OFF cells were segregated to different sublamina of the IPL, with OFF cells stratifying in the outer 40% of the IPL, and ON cells stratifying in the inner 60%. A scatter plot of cells, showing these morphological parameters, and color coded by interaction types, is shown in Fig 4c.

The OFF cells displayed a wide variety of morphologies including: 1) mono-stratified cells with axonal processes 20μm to 45μm wide, stratifying close to the INL; 2) narrower, diffusely stratifying cells spanning various parts of the OFF sublamina, and 3) a set of more monostratified cells close to the ON-OFF boundary. Of the OFF cells that showed an ON-OFF inhibition, 3 showed a similar morphology: a narrow, vertically diffuse axonal tree stratifying close to the ON-OFF boundary in the IPL. These cells may represent a distinct cell type, corresponding to CBa1-2n [1].

Rod bipolar cells showed a characteristic morphology: they had a few large, closely spaced axon terminals close to the GCL, corresponding to descriptions elsewhere [1, 9]. ON cone bipolar cells showed a diversity of morphologies similar to that of OFF cone bipolar cells, though typically with slightly wider, more monostratified axonal arbors. Delayed cancellation cells appeared in two distinct morphological types: 1) a monostratified, narrow (width= 25μm) type stratifying close to the IPL midline, and 2) a wider-field type (width = 35-50μm), stratifying closer to the ganglion cell layer, corresponding, respectively, to types CBb3 and CBb4 [1]. ON reinforcing cells appeared to include several distinct morphologies, with both monostratified and diffuse forms, at depths from the ON/OFF boundary to approximately 85% depth, likely corresponding to the remaining ON morphological types described by MacNeil et al [1].

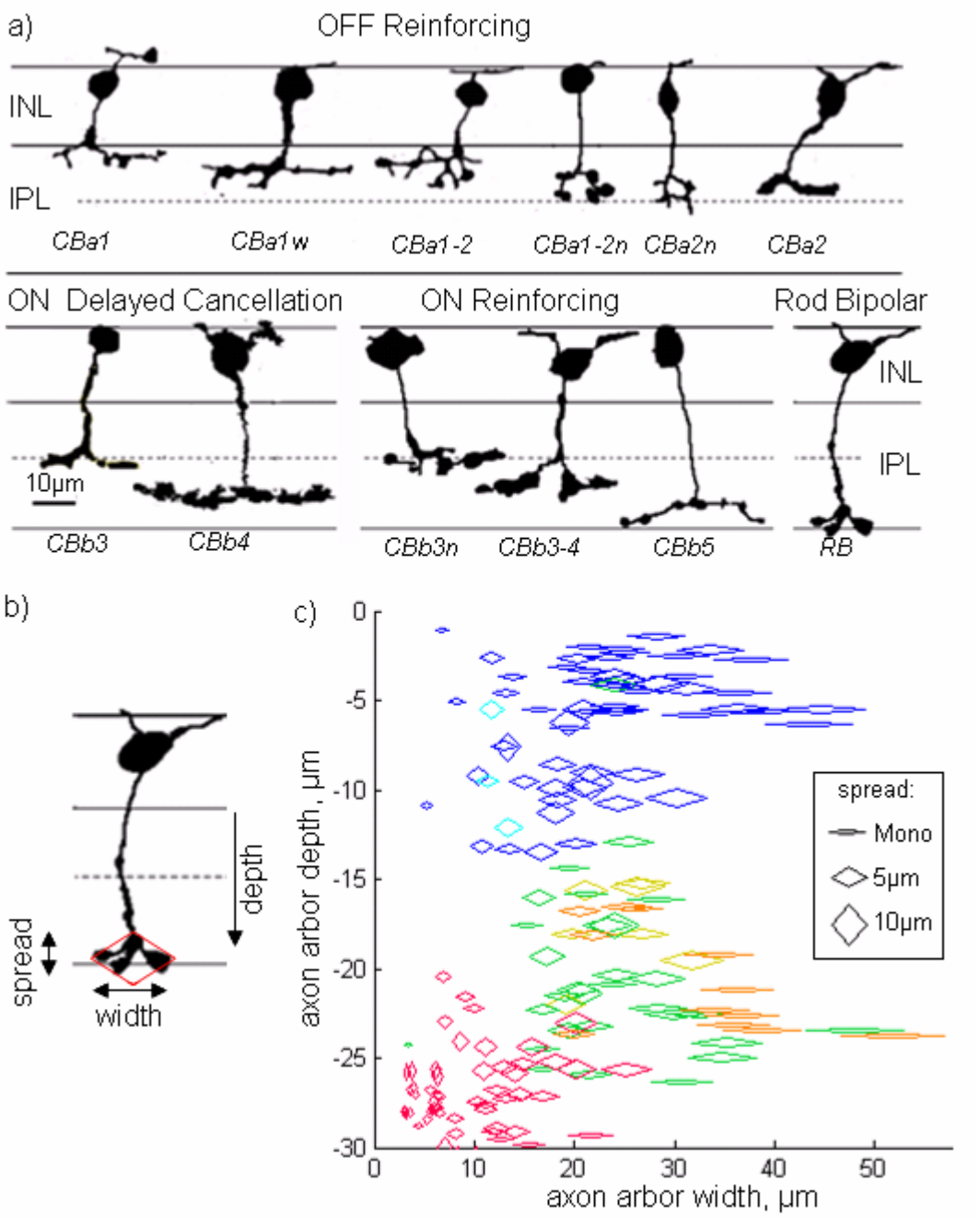


Figure 4 Similar axonal morphologies show similar interactions. a) Example morphologies for each of the different interaction types and tentative identification based upon [1].. *b) Parameters used to analyze morphology; depth is from the edge of the INL to the middle of the axonal arbor, width is the lateral extent of the arbor, spread is the vertical extent axonal branching. c) Scatter plot of morphological*

parameters of different cell types; diamond dimension are proportional to vertical spread and lateral width of processes. Red: rod bipolar cells, blue: reinforcing OFF cone bipolar cells, cyan: OFF cone bipolar cells with ON-OFF inhibition, green: ON reinforcing cone bipolar cells, orange: ON delayed cancellation cone bipolar cells, yellow: ON cone bipolar cells showing no inhibition.

Objective clustering verifies the 4 general interaction types

We sought to confirm the 4 distinct interaction types through the implementation of an objective clustering algorithm. Starting with the 153 cells from which we had obtained sinusoidal light responses, we calculated the "distance" between the sinusoid responses and morphological parameters of each possible pair of cells, and then iteratively combined pairs of cells or clusters separated by the smallest distance (see Methods). Monitoring the distance between clusters as a function of the number of clusters (see Fig 5a) we observed that cluster size increased slowly, as cells of like type combined into ever larger clusters until a relatively small number of clusters existed, at which point the distance between clusters began to increase more quickly with each iteration. The number of clusters with more than two members first increased as individual cells collected into an increasing number of separate clusters, and then decreased as these clusters coalesced toward a single super cluster. When the algorithm had reduced the data to 14 units, (5 multi-cell clusters and 9 individual cells) more than 95% of the morphologically identified rod bipolar cells were gathered into a single cluster. Thus it seemed reasonable to assume that terminating the algorithm at any point after achieving these 14 units would yield meaningful clusters. At the point

where there were 11 units, all but 7 cells had collapsed into 4 clusters, each with more than 5 members. Fig 5b shows that overlaying the amplitude and phase responses of the cells in these clusters revealed that the clusters roughly corresponded to the 4 different interactions described above. OFF reinforcing cells, rod bipolar, and ON reinforcing each dominate a single cluster, and the remaining cluster is made up of 8 ON delayed cancellation cells and 3 (high-pass) ON reinforcing cells. Two ON delayed cancellation cells were clustered with the rod bipolar cells and three others with the ON reinforcing cells.

To assess the significance of this clustering, we tested each clustering event against the null hypothesis that variation within a "real" cluster should be well described by a jointly Gaussian distribution. To find the probability that a given pair of clusters could be the product of a single such distribution, we generated more than 2000 artificial data sets and performed the clustering algorithm on them (see Methods). Using this approach we found that only 4 clusters in the entire cluster tree could be considered distinct with greater than 98% confidence. These clusters corresponded to the four clusters described above. Thus the four interaction types described correspond to the four most statistically significant clusters of cells.

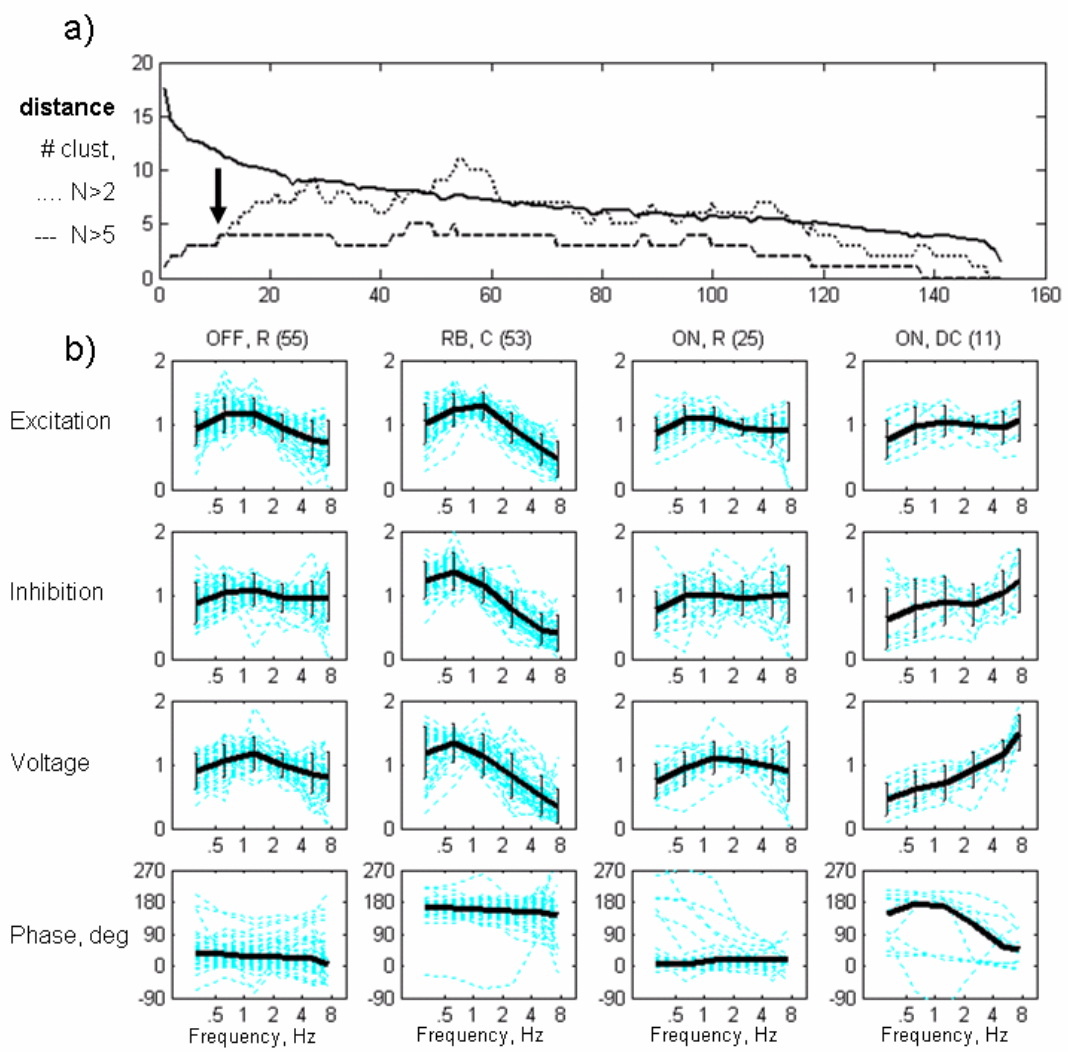


Figure 5. Clustering results: a) Graph of distance between closest clusters (solid), number of clusters with 3 or more members (dotted) and 5 or more members (dashed). Stopping at 11 clusters (arrow) yields the 4 multi-cell clusters shown in b). b) Composite responses and average response for each cluster vs. stimulus frequency. Dashed blue lines correspond to individual cell responses; solid black lines reflect average responses for each cluster. Amplitude of response in each case was normalized such that the average response across frequency for each cell was unity. Each column title refers to the dominant class of interaction (R=reinforcing, C=cancellation, DC= delayed cancellation); number of cells in cluster in parentheses.

APB confirms ON inhibitory pathways.

To confirm that the ON components of the responses reported above originated from the ON bipolar cells, we bathed the tissue in APB, a known agonist of metabotropic glutamate receptors [10]. APB eliminated excitation in ON cone (5/5) and rod bipolar cells (4/4), but not OFF bipolar cells (6/6). Furthermore, APB eliminated the reinforcing inhibition to OFF cells (7/7), indicating that this inhibitory signal originates in the ON bipolar cells and represents a crossover interaction between the ON and OFF pathways (see Fig. 6a). The canceling inhibition seen in (ON) rod bipolar cells was also eliminated by APB (4/4), indicating that this signal originates in the ON sublamina (see Fig 6b).

Reinforcement to ON cone bipolar cells was preserved in 1/3 cells, which stratified close to the midline of the IPL, but was eliminated in the other 2/3 cells, which stratified closer to the edge of the GCL. This partial blockade seems to indicate that some reinforcing signals originate in the OFF system while others actually originate in the ON system. Delayed cancellation, seen in many ON cone bipolar cells, was suppressed by APB in 2/2 cases. The effects of APB were consistently reversed by washing in control Ames' solution.

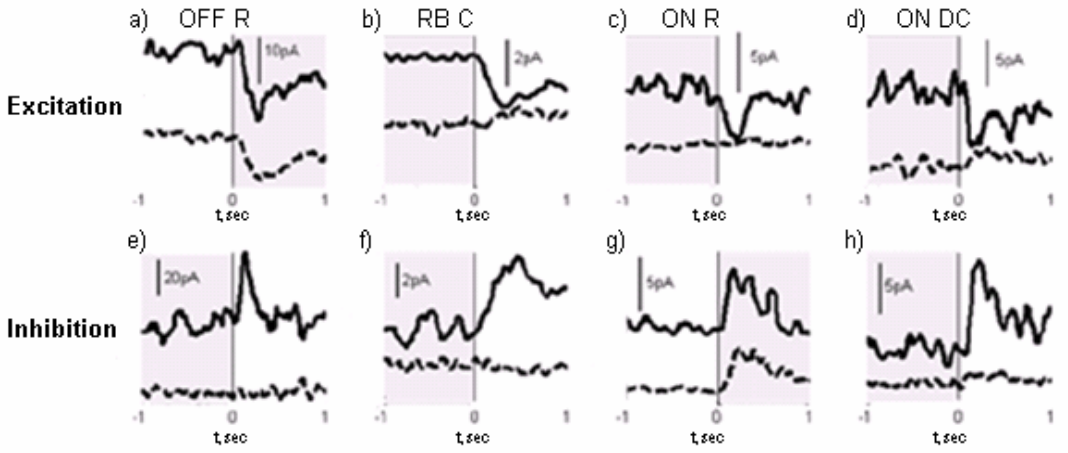


Figure 6. Effect of APB on excitation and inhibition in response to flashed stimuli. Columns correspond to different cell types: OFF reinforcing, rod bipolar (RB) canceling, ON reinforcing and ON delayed cancellation. Solid line is the control condition, dashed line is with APB. APB eliminated excitation to all ON cells (b-d) but not OFF cells (a). APB eliminated light responsive inhibition to OFF cone bipolar cells (e), rod bipolar cells (f) and delayed cancellation ON cone bipolar cells (h). Inhibition to ON reinforcing cells was not eliminated (g) in roughly ½ of cases. The y axis represents picoamps scaled as shown by scale bar, x axis represents time, in seconds. Each column title refers to a class of interaction (R=reinforcing, C=cancellation, DC= delayed cancellation).

Pharmacological identification of GABAergic and glycinergic inhibition

The inhibitory interactions described above are likely mediated by amacrine cell interneurons that release either glycine or GABA [11-23] [24-29]. We explored which neurotransmitters mediated different interactions using specific antagonists for the receptors of each of these transmitters, as illustrated in Fig 7.

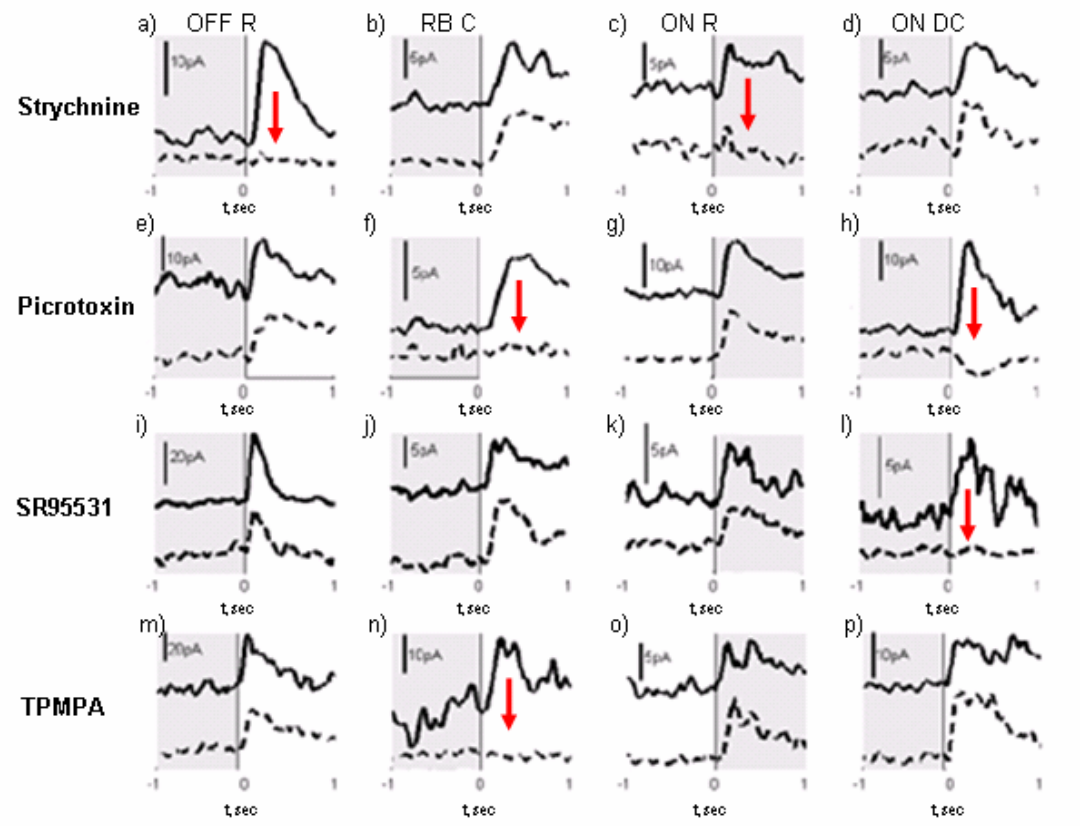


Figure 7. Effect of blockers on inhibitory currents measured in response to flashed stimuli. Columns indicate cell types: OFF reinforcing, rod bipolar canceling, ON reinforcing and ON delayed cancellation. Solid lines indicate the control condition, dashed lines indicate the blocked condition. Red arrows indicate cases where inhibition was suppressed. Strychnine (a-d) eliminated inhibitory light responses in the OFF reinforcing cells (a), and reduced inhibitory light responses in the ON reinforcing cells (c), but had little effect on cancellation inhibitions. Picrotoxin (e-h) suppressed inhibition to rod bipolar cells (f) and ON delayed-cancellation cone bipolar cells (h), but not reinforcement. SR95531 (i-l) selectively blocked delayed cancellation (l). TPMPA (m-p) selectively suppressed cancellation to rod bipolar cells (n).

Inhibition to OFF cells is blocked by strychnine but not by GABA receptor antagonists

In all 8/8 OFF bipolar cells tested, reinforcing inhibition was blocked by strychnine, a glycine receptor antagonist (see Fig 7a). This effect was seen in multiple morphological types, stratifying at various depths and with both wide and narrow ramifications. Picrotoxin, a GABA receptor antagonist, acted to reduce the high-frequency component of reinforcing inhibition to some OFF cells, but did not block the basic inhibitory flash response (Fig. 7e). This effect was found in 12/14 trials; in the 2 remaining cases, inhibition was partially suppressed, and in 1 of those 2 cases, when picrotoxin was washed out and replaced by strychnine, inhibition was completely eliminated. Thus, any picrotoxin effect in these two cases is likely presynaptic. Neither SR95531, a $GABA_A$ receptor antagonist (4/4 trials) nor TPMPA, a $GABA_C$ receptor antagonist (3/3 trials) suppressed inhibition to OFF bipolar cells. These results strongly suggest that the reinforcing inhibition to OFF cells is carried by glycine and not by GABA.

Inhibition to rod bipolar cells is blocked best by TPMPA

The canceling inhibition seen in rod bipolar cells was unaffected by strychnine in 7/13 cases, and partially suppressed in 6/13 cases. In rod bipolar cells, picrotoxin completely eliminated inhibition in 9/23 cases as shown in Fig 7f and partially suppressed inhibition in 12/23 cases. These effects were reversed by washing in control Ames' solution. SR95531 had no effect on inhibition to rod bipolar cells in 5/5 trials. TPMPA, however, strongly suppressed canceling inhibition to rod bipolar cells in 7/7 trials. Thus, it seems that canceling inhibition to rod bipolar cells is mediated primarily by $GABA_C$ receptors, but that glycine may play a secondary role.

Reinforcing inhibition to ON cone bipolar cells is reduced by strychnine

In ON cells, reinforcing inhibition was reduced by strychnine in 5/6 cells, eliminating most of the inhibitory flash response. With sinusoidal stimulation, strychnine acted to convert a broad-band inhibitory response into a high-pass response. As shown in Fig 7c, this effect corresponds to strychnine eliminating the sustained part of the inhibitory flash response leaving only a very transient inhibitory current. In one case where strychnine had no effect, APB completely eliminated the apparent reinforcing inhibition. Picrotoxin had little effect in ON reinforcing cells in 5/6 cases. Neither SR95531 nor TPMPA blocked reinforcing inhibition to ON cells in 4/4 cases each. Thus it seems that glycine is involved in reinforcing inhibition to ON cells as well as to OFF cells.

Delayed cancellation is suppressed selectively by GABA receptor antagonists

Delayed cancellation inhibition in ON cells was largely unaffected by strychnine, maintaining a wide-band inhibition with the appropriate delay in the 3/3 cases measured. However, delayed cancellation was strongly suppressed by picrotoxin in 3/3 cases. A weaker reinforcing inhibition was unmasked by picrotoxin but was suppressed by the addition of strychnine in the one case where both drugs were tried. Hence, the delayed canceling inhibition seen in some ON cone bipolar cells is a predominantly GABA-mediated signal, but some glycinergic reinforcing inhibition is also present. The effect of picrotoxin on delayed cancellation could be completely reversed by washing with control solution. Three ON bipolar cells showing delayed cancellation were treated first with TPMPA and then after a wash, with SR95531. In two cases, delayed cancellation was unaffected by TPMPA but blocked by SR95531;

in the third case, inhibition was blocked by TPMPA, while SR95531 actually increased inhibitory light response. Thus delayed cancellation to ON cone bipolar cells appears to be a GABAergic signal, but may act through either $GABA_A$ or $GABA_C$ receptors.

In summary, glycinergic inhibition is predominant in reinforcing interactions between bipolar cells, where ON inhibition reinforces OFF excitation and OFF inhibition reinforces ON excitation. GABAergic inhibition is predominant in interactions where ON inhibition cancels ON excitation.

Discussion

There are more than 27 morphologically distinct classes of amacrine cell in the rabbit retina [30], yet the inhibition they generate in bipolar cells interacts with excitation in only 4 different ways. The circuitry that underlies these interactions is discussed below.

OFF bipolar cells receive reinforcing glycinergic ON inhibition

In OFF bipolar cells, inhibition is eliminated by APB (Fig 6) suggesting that it is derived from the ON pathway. Bipolar cell axon terminals are confined to either the ON or OFF sublamina, so this inhibition must "cross over" carried by diffusely stratified ON amacrine cells that span the ON and OFF sublaminae. The inhibition is eliminated by strychnine as shown in Fig 7, suggesting that these diffuse amacrine cells are glycinergic, as shown in Fig 8a. This is consistent with previous studies demonstrating that: 1) diffusely stratifying cells have narrow field ramifications [30,

31], 2) narrow field amacrine cells are glycinergic [32] and 3) OFF cone bipolar cells show significant glycine sensitivity [23].

The majority of OFF bipolar cells receive inhibition over a wide band of frequencies. This wide-band response is unlikely to originate in the delayed-cancellation ON bipolar cells, which carry a high-pass response, or from rod bipolar cells, which carry a low-pass response. The inhibitory signals to OFF bipolar cells most likely originate in ON cone bipolar cells receiving reinforcing inhibition, since these cells show a broad band response.

The AII amacrine cell is known to inhibit OFF bipolar cells. Under dark adapted conditions, AII cells are glycinergic, receive ON excitation from rod bipolar cells, and inhibit OFF bipolar cells [33, 34]. It is unlikely that the (wide-band) inhibition we have recorded in OFF cells originates in (low-pass) rod bipolar cells. However, AII cells are also driven by electrical coupling from ON cone bipolar cells and could supply wide-band ON inhibition to the OFF bipolar cells via this pathway [35].

Rod bipolar cells receive GABAergic ON inhibition

Rod bipolar cells receive ON excitation and ON inhibition which are both eliminated by APB. This ON inhibition may be fed back from other rod bipolar cells as shown in Fig 8d, perhaps carried by A17 cells [36, 37]. If so, this inhibition does *not* reflect the action of a purely reciprocal synapse such as described by Chavez et al. [38] since the bipolar cell being recorded from is voltage clamped, preventing any modulation of its own glutamate release. Furthermore, in most of the rod bipolar cells we recorded, inhibition outweighed excitation, generating a weak hyperpolarization at light ON. If this hyperpolarization reflects the physiological behavior of rod bipolar cells, they can

not be the source of ON inhibition. More likely, the inhibitory input to rod bipolar cells originates in ON cone bipolar cells as shown in Fig 8e, and acts to suppress rod signals under photopic conditions.

The generally low-pass nature of excitation to rod bipolar cells probably reflects the slow release dynamics in rods [39-41]. The extreme low-pass nature of inhibition to rod bipolar cells must be a consequence of slow inhibitory receptors and/or of a very slow inhibitory interneuron. Rod inhibition was suppressed or completely eliminated by picrotoxin and TPMPA as shown in Fig 7, and is therefore likely carried by GABAergic amacrine cells acting at $GABA_C$ receptors. This is consistent with recent studies showing that rod bipolar cells have a significant number of $GABA_C$ receptors [42, 43] with slow dynamics [44].

Some ON bipolar cells receive reinforcing glycinergic OFF inhibition

ON reinforcing cone bipolar cells receive an inhibitory current at light OFF that is not reliably eliminated by APB, indicating that it originates in the OFF pathway. In most cases, strychnine suppressed OFF inhibition making it much more transient as shown in Fig 7c. Thus, the inhibitory signal in these ON bipolar cells is most likely carried from the OFF sublamina by glycinergic amacrine cells, as shown in Fig 8b.

But OFF glycinergic inhibition to ON bipolar cells cannot completely explain these results because: 1) In some cases reinforcement *is* eliminated by APB, and 2) recent studies [23, 45] have shown that some classes of ON cone bipolar cells, particularly those stratifying close to the GCL, contain few glycine receptors.

Electrical coupling via AII amacrine cells may provide an explanation: ON cone bipolar cells make electrical synapses with AII amacrine cells [33, 46-48] and it is

unlikely that these AII amacrine cells are voltage clamped across the gap-junctions. So, the apparent OFF inhibition measured in deeply stratifying ON cone bipolar cells may be transmitted via unclamped AII amacrine cells from other ON bipolar cells. These signals would contain an APB sensitive component, but could also carry strychnine-sensitive components from the shallower ON bipolar cells. Taken together, it is likely that the measured glycinergic reinforcing inhibition from the OFF system to ON cone bipolar cells represents a combination of direct inhibition to cells close to the ON/OFF boundary combined with input from unclamped gap-junctions to AII cells.

Some ON cone bipolar cells receive delayed GABAergic ON inhibition

ON delayed cancellation cone bipolar cells receive ON inhibition that is eliminated by APB, as shown in Fig 6d. The wide-band nature of the inhibitory response shown in Fig 3d suggests that inhibition originates, not from (high-pass) delayed cancellation bipolar cells, but from wide-band reinforcing ON bipolar cells, as shown in Fig 8c. This inhibition is eliminated by picrotoxin and SR95531 or TPMPA, and so is likely mediated by GABAergic amacrine cells, consistent with recent work indicating that ON cone bipolar cells have significant GABA sensitivity [23].

GABAergic inhibition to cone bipolar cells acts to suppress low frequency or sustained responses while enhancing the response to high frequency or transient stimuli. This is distinct from the signals seen in rod bipolar cells which suppress response across the whole band. These high-pass bipolar cells stratify at depths corresponding to the ON beta, ON alpha, and ON parasol cells [8], and may provide the transient excitatory input to these classes of ganglion cell.

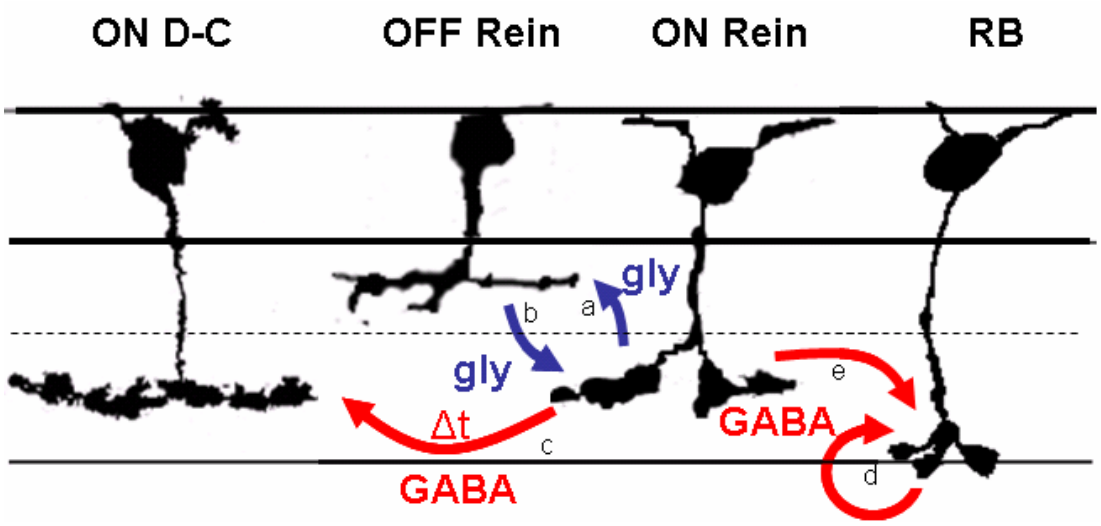


Figure 8. Circuitry providing inhibitory feedback to bipolar cells, blue arrows indicate glycinergic amacrine cells, red arrows indicate GABAergic amacrine cells. OFF cells receive sustained glycinergic inhibition from ON cells (a). OFF cells also provide a predominantly glycinergic inhibition to ON reinforcing cells (b). Delayed cancellation ON cells receive a sustained, delayed, GABAergic inhibition from reinforcing ON cells (c) that acts to enhance their response to very transient stimuli. Finally, rod bipolar cells receive a predominantly GABAergic inhibition originating in either ON cone bipolar cells (e) or other rod bipolar cells (d). Rein = reinforcing. D-C = delayed cancellation.

Asymmetries in the feedback circuits to bipolar cells

Fig 8 summarizes our understanding of the amacrine cell pathways that underlie inhibition to bipolar cells. One primary feature of these pathways is cross-lamina inhibition carried between the ON and OFF cone bipolar cells, which is predominantly glycinergic (blue arrows). There is significantly less OFF-to-ON inhibition than ON-to-OFF inhibition, revealing a significant asymmetry in signal flow between ON and OFF sublaminae. The other main feature of these pathways is inhibition within the ON

sublamina, which is predominantly GABAergic (bold red arrows). Yet there is little or no OFF to OFF inhibition, suggesting an additional asymmetry between the ON and OFF pathways.

These asymmetries are reflected in the inputs to ganglion cells as well [49]. Delayed cancellation interactions are also seen in a subpopulation of ON ganglion cells [8], where the delay serves to truncate the sustained components of excitatory inputs, leading to a more transient output. This interaction appears in approximately 70% of ON ganglion cells but is rarely seen in OFF ganglion cells. Similarly, the prevalence of reinforcing inhibition in cone bipolar cells (all OFF cells and ~ 50% of ON cells) closely matches the presence of a similar interaction in ganglion cells, where ON inhibition appears in all OFF cells but in only 50% of ON cells [8]. Remarkably, the inhibitory asymmetries apparent in bipolar cells are also apparent in ganglion cells and in similar proportions, revealing a general organizational property of the mammalian inner retina.

Push-pull interactions are prevalent at every stage of the visual system.

Reinforcing interactions, often called push-pull interactions, were by far the most common class of interaction in cone bipolar cells, appearing in every type of OFF cell, and in about half of the ON cells, either alone, or in combination with delayed cancellation, as revealed by pharmacology in Fig 7h. Push-pull interactions have been reported in ganglion cells [17, 50-52] and in the thalamus and in early visual cortex [53]. Thus, ON and OFF pathways appear to cross-inhibit each other at every stage of visual processing, starting in the bipolar cells where the ON and OFF signals are first established.

Methods

Rabbits were sacrificed, and their eyes were removed and hemisected as described previously [8, 54, 55]. Segments of the visual streak along with their associated sclera were stored in oxygenated Ames' medium in the dark. Individual segments were then removed from the sclera, mounted on Millipore paper, and sliced with a razor into 250-μm thick slices. The slices were mounted on their sides so that a cross-section was visible through the microscope. The slices were perfused with Ames' solution at 35°C. The solution was saturated with a mixtures O_2 (95%) and CO_2 (5%) and pH buffered with $NaCO_3$ to a pH of 7.4.

Patch Clamp

Bipolar cells, identified as having cell bodies near the outer edge of the inner nuclear layer (INL), were whole-cell patch-clamped with glass electrodes of resistance between 5 and 10 MΩ. The electrodes were filled with an intracellular solution that was either potassium-based (in mM: 113 $KMeSO_4$ (Fluka), 1 Mg SO_4 (Fisher Scientific), 7.8 10^{-3} $CaCl_2$ (Fisher Scientific), 0.5 BAPTA (Fisher Scientific), 10 HEPES (Sigma), 4 ATP-Na_2 (Sigma), 0.5 GTP-Na_3 (Sigma), 5 KCl (Fisher Scientific), 7.5 Neurobiotin-Cl (Vector Labs), pH 7.2.) or cesium-based (in mM: 113 $CsMeSO_4$ (Sigma), 1 Mg SO_4 (Fisher Scientific), 7.8 10^{-3} $CaCl_2$ (Fisher Scientific), 0.5 BAPTA (Fisher Scientific), 10 HEPES (Sigma), 4 ATP-Na_2 (Sigma), 0.5 GTP-Na_3 (Sigma), 5 QX314 (Sigma), 7.5 Neurobiotin-Cl (Vector Labs), pH 7.2.). In each case Alexa Fluor 488 (Invitrogen) was also included in the intracellular solution to allow imaging of the cells after physiological measurements were complete. Excitatory currents were

measured by voltage clamping the cell at the calculated reversal potential for chloride (-60 mV). Inhibitory currents were recorded by voltage clamping the cell at the cation reversal potential (0 mV). Finally, the voltage response of the cell was recorded at a current clamp of 0 pA. Chloride resting potential was confirmed by inhibitory synaptic noise which reversed polarity at -60mV. The cation reversal potential was confirmed under pharmacological blockade of inhibition, described below, where light-evoked currents reversed at 0mV, indicating a cation reversal potential of 0mV. These results were consistent using both cesium and potassium based solutions and with previous studies [8, 56, 57]. Recordings were digitized and sampled at 10 kHz. All signals were post-analyzed in MATLAB (The Mathworks). Signals were filtered and decimated to a 60Hz sample rate (automatically eliminating line noise).

Stimulus Paradigms

For each of the clamp states, a variety of stimuli were presented against a background brightness of $3x10^6$ photons/μm^2/s The most basic stimulus was a pair of two second flashes at ± 100% contrast, specifically: $6x10^6$ photons/μm^2/s and $3x10^4$ photons/μm^2/s. The stimulus took the form of a 200-μm–wide stripe projected upon the cross-section of the retina. Full-field flashes were also used to explore possible wide-field effects. Polarity (ON vs OFF) of excitation and inhibition was identified automatically by integrating current over the 800ms immediately preceding the onset or offset of a flash and subtracting from the integral of current 800ms immediately after the onset/offset of the flash. Taking this difference for the beginning and end of light and dark flashes (denoted as: d_{BL}, d_{EL}, d_{BD}, d_{ED}) we calculated a measure:

$$P = \frac{d_{BL} - d_{EL} - d_{BD} + d_{ED}}{|d_{BL}| + |d_{EL}| + |d_{BD}| + |d_{ED}|}$$

For excitation, if P>0, the cell was considered an OFF cell, if P<0, the cell was considered an ON cell. For inhibition, if P>0, the inhibition was considered to be ON in polarity, whereas if P<, the inhibition was considered to be OFF. A cell was considered to receive crossover inhibition if the identified polarity of excitation and inhibition were of opposite type.

Stripes (200 μm wide) were stimulated with a sinusoidally varying intensity to assess temporal behavior more carefully. Sinusoids with temporal frequencies ranging from 0.3 Hz to 15 Hz (incremented by a factor of 2) and contrasts of 100% and 50% were used. Sinusoid responses were analyzed by performing a Fourier transform of the response (current or voltage) during the final 3/4 of a given frequency stimulus. Sinusoidal stimuli were in the form of 4, 8, or 12 cycles, with higher frequency stimuli having more cycles; hence, 3, 6, or 9 cycles were analyzed—the first quarter was eliminated to avoid settling effects. The magnitude and phase of the Fourier-transformed response at the stimulus frequency were extracted, and noise was estimated by taking the root-mean-squared magnitudes at the frequency bins immediately higher and lower than the fundamental frequency. For example, noise was estimated for the 1.25Hz case by averaging the signal power present at 0.8 and 1.7Hz. This method of noise estimation was chosen since it permitted noise estimates from the same time series as the response itself, and at similar frequencies. Similar analyses were performed to extract the magnitude, phase and noise for the higher harmonics of the response (i.e. for a 1.25Hz stimulus, the 2.5Hz and 3.75Hz frequency bins were analyzed to extract the 2^{nd} and 3^{rd} harmonics).

The magnitude of the voltage response across the various stimulus frequencies and cells were categorized as responding best to low frequencies (low-pass), high frequencies (high-pass), or roughly equally to both high and low frequencies (wide-band). This categorization was performed by normalizing the average response magnitude to 1 and fitting the curve with a straight line: a slope greater than 0.4 (normalizing the log of frequency to 1) was categorized as high-pass, less than -0.4 as low pass, and slopes between as wide band. Interactions were analyzed by comparing the phase of excitatory and inhibitory responses across the range of frequencies tested: a 0-degree phase difference implies that when excitation (an inward current) is maximum, inhibition (an outward current) is minimum, and when excitation was minimum, inhibition was maximum.

Morphological Reconstruction

Once physiological recordings were complete, intact neurons were imaged with Alexa Fluor 488. Digital images were captured under fluorescence and normal white light at several focal depths and superimposed. By comparing the stratification of axon terminals to the edges of the IPL, their depth of stratification could be measured, which along with their lateral width and vertical diffuseness (see Figure 4), provided a quantitative morphological description comparable with previously published results. Images were also compared directly with published examples of different identified morphological types. Finally, example cells were traced for easier comparison.

Clustering

To objectively identify physiologically distinct cell types, a clustering algorithm was employed; this used the amplitude and phase of the excitatory, inhibitory and voltage

responses at each frequency, as well as the magnitude and phase of their 2nd and 3rd harmonics. Responses were transformed from polar to Cartesian coordinates by taking amplitude R and phase ϕ and applying the equations $x = R \cdot \cos(\phi)$, $y = R \cdot \sin(\phi)$. Thus, each sinusoidal stimulus yielded two orthogonal components each (x and y) for 3 harmonics. Using 6 distinct stimulus frequencies yielded a total of 36 points for each of excitation, inhibition and voltage, or 108 total points. To account for variability in slice viability, each response type (excitation, inhibition and voltage) was normalized to an average power of 1. Morphological characteristics were also included: specifically the depth, width and vertical spread of the axonal arbor (shown in figure 4). Thus, each cell was described by 108 physiological and 3 morphological numbers, forming a 111-dimensional vector. The degree of similarity between different cells' responses was characterized by a distance metric, set by the mean squared difference between the individual 111-point vectors. To avoid artifacts due to noise, the squared difference between individual points was reduced by the summed, squared noise associated with those two points, with noise estimated as described above. Two measured responses whose differences were smaller than the noise associated with the measurements were considered to be identical. Morphological parameters were weighted to contribute approximately the same total distance as physiological parameters.

Clustering was initiated by finding the two cells separated by the smallest distance and merging them into a single cluster. The vector associated with this cluster was then taken to simply be the average of its component cells' vectors. Distances were then recalculated, and the process repeated. Merging of two clusters results in a single

larger cluster, whose vector is calculated to be the mean of all of its component cells' vectors. Repeating this process iteratively leads to increasingly large clusters, until a single final cluster is generated, containing all of the cells. Stopping the process earlier yields multiple clusters consisting of individual cells with similar physiologies and morphologies.

Although this clustering algorithm provides an objective categorization of cells types, it does not automatically provide a clear metric for how many clusters are "real" and not simply an artifact of experimental variability. Choosing an appropriate stopping point for the algorithm is especially difficult in the presence of even a small number of outlier data points. Such points can be as distant from any "real" cluster as the clusters are from one another, and so confuse the question of what distance constitutes the separation of two "real" clusters. To address this concern, we developed a test for the likelihood that a given pair of clusters actually represents a single super-cluster. As a null hypothesis, we assumed that the variation within a real cluster was due to a combination of many independent variables and so could be well described by a jointly Gaussian distribution. To assess the likelihood that a given cluster was real, we calculated its covariance matrix and then generated many sets of artificial data with the same number of points, drawn from a jointly Gaussian distribution with that same covariance and mean as the original data cluster. Performing the clustering algorithm on this artificial data yields a clustering tree. By repeating this process of artificial data generation and clustering many times we could estimate the probability of any given size clustering event occurring in the process of generating a given super-cluster. We then estimated the probability of each *actual* clustering event that lead to

a given *actual* cluster. If this probability was sufficiently small, we concluded that the two sub-clusters were distinct. We performed this process iteratively starting with the final super cluster containing the whole dataset, and progressed until each putative cluster was generated by a cluster tree of events with probability greater than 2%.

Pharmacology

Experiments were repeated in the presence of pharmacological blockers of excitation and inhibition. To block metabotropic glutamate receptors, and so selectively inactivate the ON system, 10 μM APB (L-AP4) (Tocris) was added to Ames' medium and perfused across the preparation. Recordings were repeated a third time under normal Ames' as a control. To block ionotropic GABA receptors ($GABA_A$ and $GABA_C$), 100 μM picrotoxin (Sigma) was added to Ames' medium and perfused across the preparation. Recordings were repeated a third time under normal Ames' as a control. To block ionotropic glycine receptors, 10 μM strychnine (Sigma) was added to Ames' medium and perfused across the preparation. A wash step was repeated here, but we found that strychnine was very slow to wash out, and so was rarely reversible. In order to confirm our picrotoxin findings (since picrotoxin has recently been reported to block glycine receptors in some cases [58]) and to separate GABA receptor types we performed experiments using 10μM SR95531 [59], a $GABA_A$ receptor specific blocker with little glycine receptor reactivity [58] (Sigma) and 10μM TPMPA, [60] a $GABA_C$ receptor specific blocker (Sigma) each in Ames' medium. We followed each drug with a wash step, and found that effects were completely reversed for TPMPA and more slowly washed out for SR95531. In several cases we

combined strychnine with one or another of the GABA blockers using the same concentrations described above.

References

[1] M. A. MacNeil, J. K. Heussy, R. F. Dacheux, E. Raviola, and R. H. Masland, "The population of bipolar cells in the rabbit retina," *J Comp Neurol,* vol. 472, pp. 73-86, Apr 19 2004.

[2] S. H. DeVries, "Bipolar cells use kainate and AMPA receptors to filter visual information into separate channels," *Neuron,* vol. 28, pp. 847-56, Dec 2000.

[3] C. J. Dong and F. S. Werblin, "Temporal contrast enhancement via GABAC feedback at bipolar terminals in the tiger salamander retina," *J Neurophysiol,* vol. 79, pp. 2171-80, Apr 1998.

[4] B. Volgyi, D. Xin, and S. A. Bloomfield, "Feedback inhibition in the inner plexiform layer underlies the surround-mediated responses of AII amacrine cells in the mammalian retina," *J Physiol,* vol. 539, pp. 603-14, Mar 1 2002.

[5] T. Euler and R. H. Masland, "Light-evoked responses of bipolar cells in a mammalian retina," *J Neurophysiol,* vol. 83, pp. 1817-29, Apr 2000.

[6] P. D. Lukasiewicz, B. R. Maple, and F. S. Werblin, "A novel GABA receptor on bipolar cell terminals in the tiger salamander retina," *J Neurosci,* vol. 14, pp. 1202-12, Mar 1994.

[7] P. D. Lukasiewicz and F. S. Werblin, "A novel GABA receptor modulates synaptic transmission from bipolar to ganglion and amacrine cells in the tiger salamander retina," *J Neurosci,* vol. 14, pp. 1213-23, Mar 1994.

[8] B. Roska, A. Molnar, and F. Werblin, "Parallel Processing in Retinal Ganglion Cells: How Integration of Space-Time Patterns of Excitation and Inhibition Form the Spiking Output," *Journal of Neurophysiology,* vol. 95, pp. 3810-22, June 2006.

[9] T. Euler and H. Wassle, "Different contributions of GABAA and GABAC receptors to rod and cone bipolar cells in a rat retinal slice preparation," *J Neurophysiol,* vol. 79, pp. 1384-95, Mar 1998.

[10] M. M. Slaughter and R. F. Miller, "2-amino-4-phosphonobutyric acid: a new pharmacological tool for retina research," *Science,* vol. 211, pp. 182-5, Jan 9 1981.

[11] C. Y. Yang, P. Lukasiewicz, G. Maguire, F. S. Werblin, and S. Yazulla, "Amacrine cells in the tiger salamander retina: morphology, physiology, and neurotransmitter identification," *J Comp Neurol,* vol. 312, pp. 19-32, Oct 1 1991.

[12] P. D. Lukasiewicz and F. S. Werblin, "The spatial distribution of excitatory and inhibitory inputs to ganglion cell dendrites in the tiger salamander retina," *J Neurosci,* vol. 10, pp. 210-21, Jan 1990.

[13] S. Haverkamp, U. Muller, K. Harvey, R. J. Harvey, H. Betz, and H. Wassle, "Diversity of glycine receptors in the mouse retina: localization of the alpha3 subunit," *J Comp Neurol,* vol. 465, pp. 524-39, Oct 27 2003.

[14] U. Grunert and H. Wassle, "Immunocytochemical localization of glycine receptors in the mammalian retina," *J Comp Neurol,* vol. 335, pp. 523-37, Sep 22 1993.

[15] R. Boos, H. Schneider, and H. Wassle, "Voltage- and transmitter-gated currents of all-amacrine cells in a slice preparation of the rat retina," *J Neurosci,* vol. 13, pp. 2874-88, Jul 1993.

[16] U. Greferath, F. Muller, H. Wassle, B. Shivers, and P. Seeburg, "Localization of GABAA receptors in the rat retina," *Vis Neurosci,* vol. 10, pp. 551-61, May-Jun 1993.
[17] F. Muller, H. Wassle, and T. Voigt, "Pharmacological modulation of the rod pathway in the cat retina," *J Neurophysiol,* vol. 59, pp. 1657-72, Jun 1988.
[18] J. Jager and H. Wassle, "Localization of glycine uptake and receptors in the cat retina," *Neurosci Lett,* vol. 75, pp. 147-51, Mar 31 1987.
[19] H. Wassle, I. Schafer-Trenkler, and T. Voigt, "Analysis of a glycinergic inhibitory pathway in the cat retina," *J Neurosci,* vol. 6, pp. 594-604, Feb 1986.
[20] J. Bolz, P. Thier, T. Voigt, and H. Wassle, "Action and localization of glycine and taurine in the cat retina," *J Physiol,* vol. 362, pp. 395-413, May 1985.
[21] M. A. Freed and P. Sterling, "The ON-alpha ganglion cell of the cat retina and its presynaptic cell types," *J Neurosci,* vol. 8, pp. 2303-20, Jul 1988.
[22] S. A. Bloomfield and D. Xin, "Surround inhibition of mammalian AII amacrine cells is generated in the proximal retina," *J Physiol,* vol. 523 Pt 3, pp. 771-83, Mar 15 2000.
[23] C. Zhou and R. F. Dacheux, "Glycine- and GABA-activated inhibitory currents on axon terminals of rabbit cone bipolar cells," *Vis Neurosci,* vol. 22, pp. 759-67, Nov-Dec 2005.
[24] J. G. Cueva, S. Haverkamp, R. J. Reimer, R. Edwards, H. Wassle, and N. C. Brecha, "Vesicular gamma-aminobutyric acid transporter expression in amacrine and horizontal cells," *J Comp Neurol,* vol. 445, pp. 227-37, Apr 8 2002.
[25] P. Koulen, B. Malitschek, R. Kuhn, B. Bettler, H. Wassle, and J. H. Brandstatter, "Presynaptic and postsynaptic localization of GABA(B) receptors in neurons of the rat retina," *Eur J Neurosci,* vol. 10, pp. 1446-56, Apr 1998.
[26] H. Wassle, P. Koulen, J. H. Brandstatter, E. L. Fletcher, and C. M. Becker, "Glycine and GABA receptors in the mammalian retina," *Vision Res,* vol. 38, pp. 1411-30, May 1998.
[27] E. L. Fletcher, P. Koulen, and H. Wassle, "GABAA and GABAC receptors on mammalian rod bipolar cells," *J Comp Neurol,* vol. 396, pp. 351-65, Jul 6 1998.
[28] U. Greferath, U. Grunert, J. M. Fritschy, A. Stephenson, H. Mohler, and H. Wassle, "GABAA receptor subunits have differential distributions in the rat retina: in situ hybridization and immunohistochemistry," *J Comp Neurol,* vol. 353, pp. 553-71, Mar 20 1995.
[29] M. A. Freed, Y. Nakamura, and P. Sterling, "Four types of amacrine in the cat retina that accumulate GABA," *J Comp Neurol,* vol. 219, pp. 295-304, Sep 20 1983.
[30] M. A. MacNeil, J. K. Heussy, R. F. Dacheux, E. Raviola, and R. H. Masland, "The shapes and numbers of amacrine cells: matching of photofilled with Golgi-stained cells in the rabbit retina and comparison with other mammalian species," *J Comp Neurol,* vol. 413, pp. 305-26, Oct 18 1999.
[31] M. A. MacNeil and R. H. Masland, "Extreme diversity among amacrine cells: implications for function," *Neuron,* vol. 20, pp. 971-82, May 1998.

[32] N. Menger, D. V. Pow, and H. Wassle, "Glycinergic amacrine cells of the rat retina," *J Comp Neurol,* vol. 401, pp. 34-46, Nov 9 1998.
[33] S. L. Mills and S. C. Massey, "Differential properties of two gap junctional pathways made by AII amacrine cells," *Nature,* vol. 377, pp. 734-7, Oct 26 1995.
[34] S. L. Mills and S. C. Massey, "Labeling and distribution of AII amacrine cells in the rabbit retina," *J Comp Neurol,* vol. 304, pp. 491-501, Feb 15 1991.
[35] D. Xin and S. A. Bloomfield, "Comparison of the responses of AII amacrine cells in the dark- and light-adapted rabbit retina," *Vis Neurosci,* vol. 16, pp. 653-65, Jul-Aug 1999.
[36] E. Hartveit, "Reciprocal synaptic interactions between rod bipolar cells and amacrine cells in the rat retina," *J Neurophysiol,* vol. 81, pp. 2923-36, Jun 1999.
[37] R. Nelson and H. Kolb, "A17: a broad-field amacrine cell in the rod system of the cat retina," *J Neurophysiol,* vol. 54, pp. 592-614, Sep 1985.
[38] A. E. Chavez, J. H. Singer, and J. S. Diamond, "Fast neurotransmitter release triggered by Ca influx through AMPA-type glutamate receptors," *Nature,* vol. 443, pp. 705-8, Oct 12 2006.
[39] J. Schnapf and D. Copenhagen, "Differences in the kinetics of rod and cone synaptic transmission," *Nature,* vol. 296, pp. 862-4, Apr 29 1982.
[40] K. Rabl, L. Cadetti, and W. Thoreson, "Paired-pulse depression at photoreceptor synapses," *J Neurosci,* vol. 26, pp. 2555-63, Mar 1 2006.
[41] K. Rabl, L. Cadetti, and W. Thoreson, "Kinetics of exocytosis is faster in cones than in rods," *J Neurosci,* vol. 25, pp. 4633-40, May 4 2005.
[42] E. D. Eggers and P. D. Lukasiewicz, "GABAA, GABAC and glycine receptor-mediated inhibition differentially affects light-evoked signaling from mouse retinal rod bipolar cells," *J Physiol,* Jan 26 2006.
[43] M. J. Frech and K. H. Backus, "Characterization of inhibitory postsynaptic currents in rod bipolar cells of the mouse retina," *Vis Neurosci,* vol. 21, pp. 645-52, Jul-Aug 2004.
[44] E. D. Eggers and P. D. Lukasiewicz, "Receptor and transmitter release properties set the time course of retinal inhibition," *J Neurosci,* vol. 26, pp. 9413-25, Sep 13 2006.
[45] E. Ivanova, U. Muller, and H. Wassle, "Characterization of the glycinergic input to bipolar cells of the mouse retina," *Eur J Neurosci,* vol. 23, pp. 350-64, Jan 2006.
[46] S. C. Massey and S. L. Mills, "Gap junctions between AII amacrine cells and calbindin-positive bipolar cells in the rabbit retina," *Vis Neurosci,* vol. 16, pp. 1181-9, Nov-Dec 1999.
[47] S. A. Bloomfield and D. Xin, "A comparison of receptive-field and tracer-coupling size of amacrine and ganglion cells in the rabbit retina," *Vis Neurosci,* vol. 14, pp. 1153-65, Nov-Dec 1997.
[48] E. B. Trexler, W. Li, S. L. Mills, and S. C. Massey, "Coupling from AII amacrine cells to ON cone bipolar cells is bidirectional," *J Comp Neurol,* vol. 437, pp. 408-22, Sep 3 2001.

[49] K. A. Zaghloul, K. Boahen, and J. B. Demb, "Different circuits for ON and OFF retinal ganglion cells cause different contrast sensitivities," *J Neurosci,* vol. 23, pp. 2645-54, Apr 1 2003.
[50] J. H. Belgum, D. R. Dvorak, J. S. McReynolds, and E. Miyachi, "Push-pull effect of surround illumination on excitatory and inhibitory inputs to mudpuppy retinal ganglion cells," *J Physiol,* vol. 388, pp. 233-43, Jul 1987.
[51] M. S. Arkin and R. F. Miller, "Bipolar origin of synaptic inputs to sustained OFF-ganglion cells in the mudpuppy retina," *J Neurophysiol,* vol. 60, pp. 1122-42, Sep 1988.
[52] M. S. Arkin and R. F. Miller, "Synaptic inputs and morphology of sustained ON-ganglion cells in the mudpuppy retina," *J Neurophysiol,* vol. 60, pp. 1143-59, Sep 1988.
[53] J. A. Hirsch, "Synaptic physiology and receptive field structure in the early visual pathway of the cat," *Cereb Cortex,* vol. 13, pp. 63-9, Jan 2003.
[54] B. Roska and F. Werblin, "Vertical interactions across ten parallel, stacked representations in the mammalian retina," *Nature,* vol. 410, pp. 583-7, Mar 29 2001.
[55] B. Roska and F. Werblin, "Rapid global shifts in natural scenes block spiking in specific ganglion cell types," *Nat Neurosci,* vol. 6, pp. 600-8, Jun 2003.
[56] S. I. Fried, T. A. Munch, and F. S. Werblin, "Directional selectivity is formed at multiple levels by laterally offset inhibition in the rabbit retina," *Neuron,* vol. 46, pp. 117-27, Apr 7 2005.
[57] S. M. Wu, F. Gao, and J. J. Pang, "Synaptic circuitry mediating light-evoked signals in dark-adapted mouse retina," *Vision Res,* vol. 44, pp. 3277-88, Dec 2004.
[58] P. Wang and M. M. Slaughter, "Effects of GABA receptor antagonists on retinal glycine receptors and on homomeric glycine receptor alpha subunits," *J Neurophysiol,* vol. 93, pp. 3120-6, Jun 2005.
[59] B. D. Gynther and D. R. Curtis, "Pyridazinyl-GABA derivatives as GABA and glycine antagonists in the spinal cord of the cat," *Neurosci Lett,* vol. 68, pp. 211-5, Jul 24 1986.
[60] D. Ragozzino, R. M. Woodward, Y. Murata, F. Eusebi, L. E. Overman, and R. Miledi, "Design and in vitro pharmacology of a selective gamma-aminobutyric acidC receptor antagonist," *Mol Pharmacol,* vol. 50, pp. 1024-30, Oct 1996.

Chapter 5

Retinal Circuitry Compensates for Synaptic Distortion

Introduction

At the first synapse of the visual system, bipolar cells separate visual signals into the ON and OFF pathways[1] [2]. ON and OFF responses are found in neurons at every level of the visual system including the visual cortex. These two pathways have been shown to interact, cross-suppressing one another in retinal ganglion cells[3] [4] [5] [6], yet the function of this cross-suppression is not fully understood. At almost every stage of processing, visual signals are distorted by synaptic rectification[7] [8], yet many of the neurons along these pathways behave linearly[9] [10] [11]. We have found that at every stage of retinal processing, ON-derived signals inhibit OFF cells and OFF-derived signals inhibit ON cells, and that this "crossover" inhibition compensates for the distorting effects of rectification. We show that crossover inhibition appears to linearize the behavior of "X-type" ganglion cells, which respond

to changes in the average light intensity in their receptive field center, but do not respond to contrast-inverting gratings in the same area[9] [10] [11]. We have measured crossover inhibition (often called push-pull activity) in bipolar, amacrine, and ganglion cells, and it has also been inferred at the lateral geniculate nucleus and visual cortex[12] where it appears to play a similar role in reestablishing signal fidelity in the face of rectifying nonlinearities.

Crossover inhibition can suppress rectification introduced by synapses

Fig 5.1 illustrates an example of how ON and OFF signals can interact to correct for rectification, linearizing the response of an OFF amacrine cell. As diagrammed in Fig 5.1a, signals from cones are split into ON and OFF pathways. These pathways supply the OFF amacrine cell with excitation from OFF bipolar cells and crossover inhibition from ON amacrine cells. At light onset ON cells depolarize while OFF cells hyperpolarize, and these responses rebound with roughly equal magnitude at light OFF, as shown in Fig 5.1b. The postsynaptic conductances derived from these inputs, however, are strongly rectified. The conductance derived from the OFF bipolar cells decreases slightly at light onset but increases strongly at light offset. The postsynaptic conductance derived from ON amacrine cells increases strongly at light onset but decreases only slightly at light offset. Because of rectification, both postsynaptic conductances are larger for presynaptic depolarization than for hyperpolarization. These conductances are translated into post-synaptic currents in the OFF amacrine cell that are "in phase," as shown in Fig 5.1d. Both the excitatory and inhibitory currents show an outward change at light ON and an inward change at light

OFF. The resulting voltage response (Fig 5.1e) is composed of hyperpolarizing and depolarizing transient steps of roughly equal magnitude at light ON and OFF, indicating that the distorting effects of rectification have been suppressed. The more symmetric responses at ON and OFF reflects a more linear response.

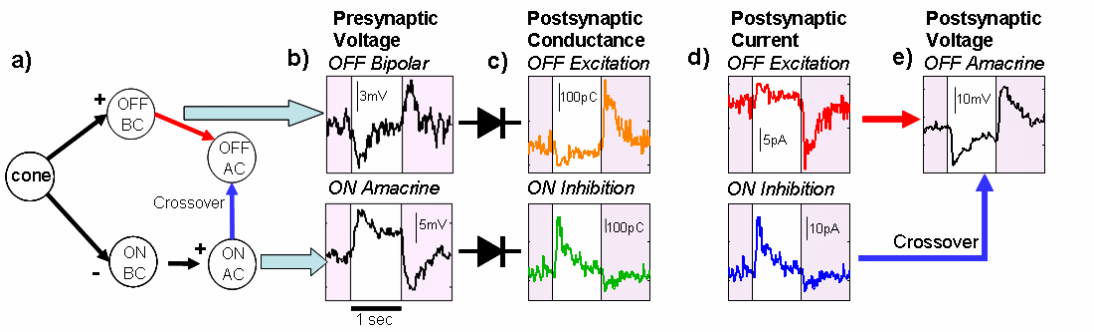


Figure 5.1. Signals showing rectification and re-linearization in generating the flash response of an OFF amacrine cell. a) Flow diagram showing convergence of typical ON and OFF signals at an OFF amacrine cell. Cones drive both ON and OFF bipolar cells (BCs), OFF bipolar cells provide excitatory input to the OFF amacrine cell (OFF AC), while ON bipolar cells drive ON amacrine cells (ON AC) that provide inhibitory input to the OFF amacrine cell. b) Example responses of an OFF bipolar and ON amacrine cell. c) Excitatory and inhibitory conductances measured from an OFF amacrine cell; both favor the depolarizing phases of the presynaptic responses. d) Synaptic currents in the OFF amacrine cell are "in phase": both show an outward change in current at light ON and an inward change at light OFF. e) These currents add to generate a voltage response in the amacrine cell with equal magnitude transients at light ON and OFF, correcting for the rectification introduced in c). White regions indicate stimulus period, grey areas indicate background brightness; stimulus intensity is 2x baseline.

Rectification and crossover inhibition are ubiquitous in the inner retina

We found similar rectification and crossover inhibition in the majority of bipolar, amacrine and ganglion cells. As shown in Fig 5.2, virtually every OFF bipolar cell (44/48), and most OFF ganglion (52/69) and OFF amacrine cells (33/51) showed crossover inhibition from the ON system in response to light flashes. These included most morphological types of OFF bipolar[13] and ganglion cells[14] and various morphologies of amacrine cells, including both mono-stratified and diffusely stratified types[15]. A similar crossover inhibition from the OFF system was dominant in subsets of ON bipolar cells (26/49), ON ganglion cells (10/67), and in most ON amacrine cells (57/76). Even in cell types where crossover inhibition was not dominant, it could still be measured. For example, we found crossover inhibition in an additional 44/67 ON ganglion cells where it appeared in conjunction with other types of inhibition.

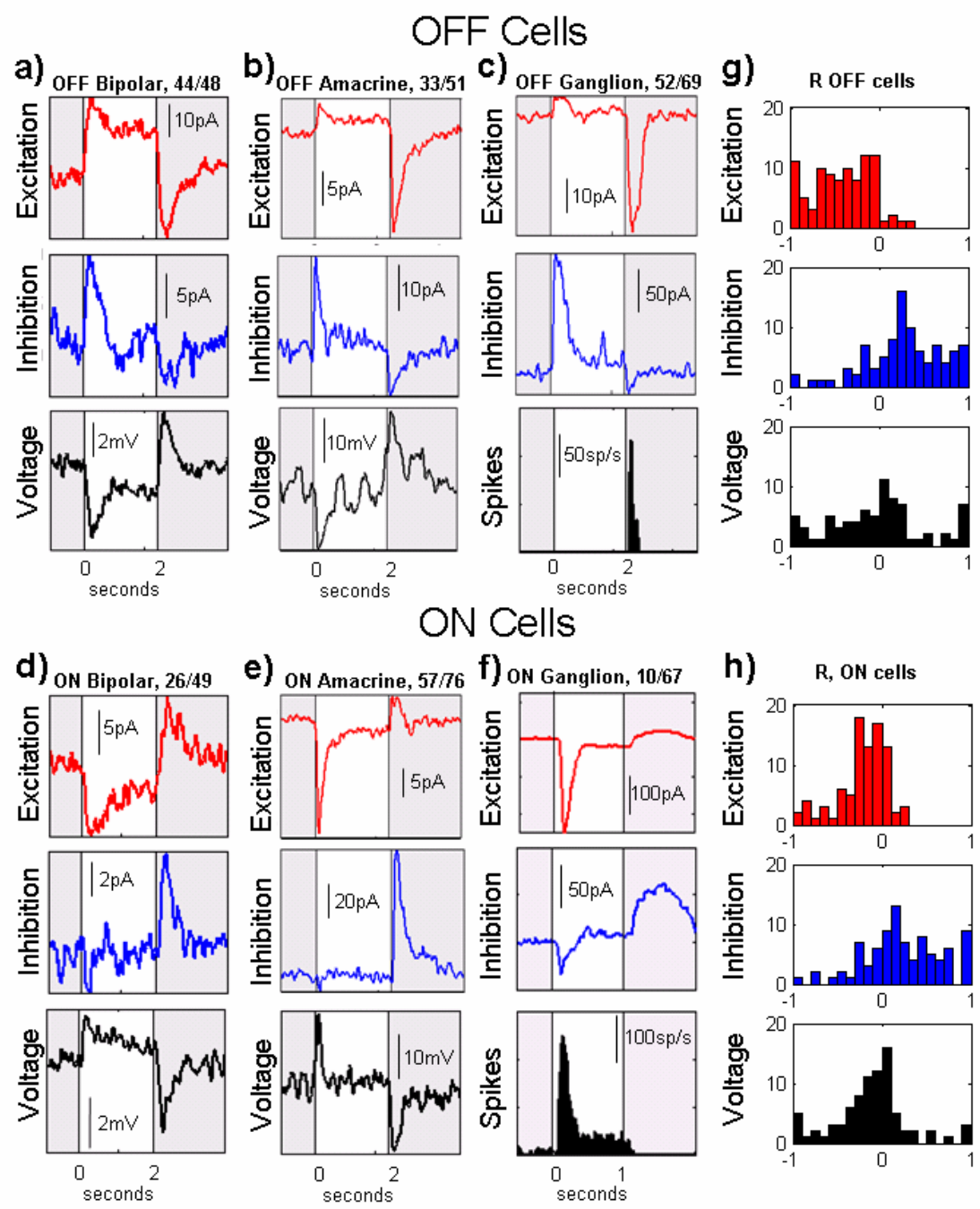


Figure 5.2 Crossover inhibition in bipolar, amacrine, and ganglion cells in response to a stepped, bright flash. a-c) Examples OFF bipolar, amacrine and ganglion cells, d-f) Examples of ON bipolar, amacrine and ganglion cells. In all cases, the excitatory and inhibitory currents are in phase. Red denotes excitatory current (measured at

Vclamp = -60mV), blue denotes inhibitory current (measured at Vclamp = 0mV), black denotes voltage or spiking rate (for ganglion cells). Numbers indicate the fraction of measured cells where crossover inhibition was dominant. g,h) Histograms of rectification, R, for excitation, inhibition and voltage in OFF and ON cells showing crossover inhibition. R values for excitation is mostly negative, indicating inward rectification, while R values for inhibition are mostly positive, indicating outward rectification, and voltage is relatively symmetric (R close to zero).

We characterized the strength and polarity of rectification with a measure, R (see Methods) for cells in which both light and dark flashes were used. Histograms of this measure are shown in Fig 5.2g for all such cells exhibiting crossover inhibition (OFF: N = 93, ON: N = 88). Excitation was predominantly inward rectifying (R_{OFF} = -0.42±0.34, p = 2×10^{-15}, R_{ON} = -0.22±0.28, p = 10^{-10}, mean ± s.d., Wilcoxon signed rank test comparing R to zero) while inhibition was predominantly outward rectifying (R_{OFF} = 0.28±0.44, p = 2×10^{-8}, R_{ON} = 0.23±0.44, p = 4×10^{-6}). Voltage responses were much more symmetric in OFF cells (R_{OFF} = -0.02±0.53, p = 0.6) while hyperpolarizing steps were slightly greater magnitude than depolarizing steps in ON cells (R_{ON} = -0.12±0.41, p = 0.001). ON bipolar cells showed the least excitatory rectification (R = .09 ± 0.19) and previous work has shown that ON ganglion cells have more linear inputs than OFF ganglion cells[7] [16] [17]. ON bipolar and ganglion cells are also the cell types that show the least crossover inhibition, perhaps indicating that cells with more linear excitatory inputs require less correction from crossover inhibition.

Crossover inhibition suppresses temporal rectification artifacts

Crossover inhibition compensates for rectification at each level of retinal processing as shown in Fig 5.2. After compensation, each cell's output is rectified again, either by succeeding synapses or spike generation (see Fig 5.2c, f). For simple light flashes, this rectification reverses the effects of crossover inhibition. Figures 5.3 and 5.4 demonstrate that for more complex stimuli rectification generates ambiguous responses that can be corrected by crossover inhibition.

The presence of rectification can confuse changes in brightness with changes in temporal contrast, as shown in Fig 5.3 for an OFF bipolar cell. The stimulus shown at the left consists of a fast 1.2Hz sinusoid that is amplitude-modulated by a much slower 0.3Hz sinusoid. The excitatory and inhibitory currents both contain an apparent 0.3 Hz brightness signal not present in the original stimulus. This low frequency component reflects distortion introduced by rectification interacting with the changing temporal contrast. The fast components of excitation and inhibition, representing the original brightness signal, are in-phase and add constructively. The low frequency components of excitation and inhibition, however, are out of phase, such that these low frequency currents cancel, suppressing the contrast artifact and yielding a voltage response that more closely resembles the original stimulus.

Crossover inhibition acted to suppress contrast artifacts in the majority of cells stimulated with modulated sinusoids, as shown in Fig 5.3d. The phase relationship between excitation and inhibition at the high frequency was close to 0 degrees (-2 ± 38, mean ± s.d.) implying that inhibition enhances excitation encoding the original fast part of the response. In contrast, the phase difference between the low frequency

artifacts in excitation and inhibition was closer to180 degrees (188 ± 48) implying that these artifacts would typically cancel. Thus, crossover inhibition maintains components of the synaptic signals that relate to brightness, but suppresses components of the synaptic signals that relate to contrast alone.

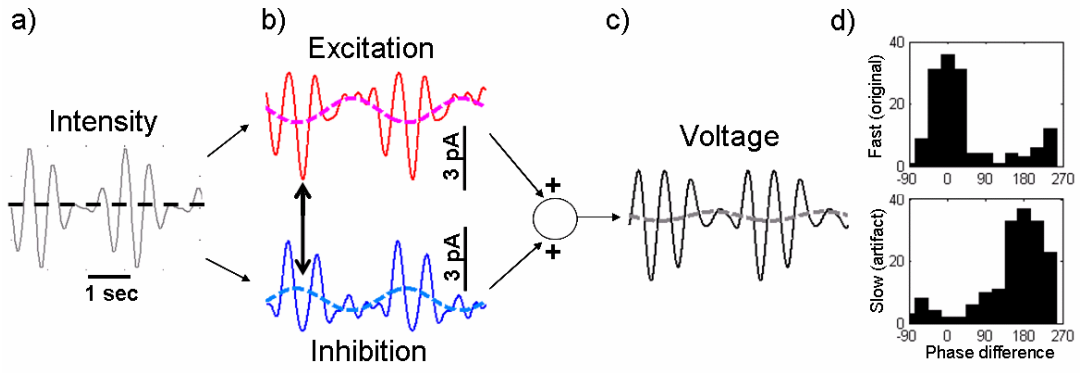


Fig 5.3. Rectification confounds temporal contrast with brightness in an OFF bipolar cell: a) The stimulus was a sinusoidally varying intensity with a slower, sinusoidal amplitude envelop, dashed lines indicate component of response at the frequency of the envelope. b) Because of rectification, both excitatory and inhibitory inputs show a strong response at the slower envelope frequency. These envelope terms are out of phase and cancel resulting in: c) a more linear voltage response. Histograms of relative phase at slow (envelope) and fast frequencies for bipolar, amacrine and ganglion cells, pooled across 3 combinations of fast- and slow-frequencies: 0.3Hz x 1.2Hz, 1.2Hz x 4.8Hz, and 0.3Hz x 4.8Hz. The phase for the fast part of the stimulus is centered at zero degrees (in phase) while the phase for the slow contrast artifact is centered at 180 degrees (out of phase).

Crossover inhibition suppresses spatial artifacts introduced by rectification

Synaptic rectification can also confuse brightness with *spatial* contrast as shown in Fig 5.4. We stimulated the retina with high-contrast gratings of alternating light-and dark stripes that covered the center of the ganglion cell's measured receptive field. The intensity of each stripe was inverted every 0.5 seconds: dark stripes became bright and bright stripes became dark such that average brightness stayed constant[9] [7] [11]. This was compared to cases where only half the stripes were changed at a time. As shown in the left two columns of Fig 5.4a, the partial gratings elicited strong spiking from linear OFF ganglion cells (similar to X-cells in cats) when transitioning from light to dark. This strong response is expected, since the average brightness decreased. The full gratings, however, elicited little spiking in linear-responding ganglion cells, as shown in the third column of Fig 5.4a. This demonstrates the linearity of the cell's response to a stimulus whose average brightness did not change. Does this linear response to contrast inverting gratings derive from the kind of crossover inhibition described above? We tested this notion by eliminating ON activity originating in ON bipolar cells with APB (concentration = 20μM), a known agonist to mGluR6 receptors [18] [6]. APB had little effect on the (OFF) spiking response to partial gratings, but dramatically increased the response to each transition of the full grating as shown in Fig 5.4b. In the presence of APB, only the OFF-bipolar mediated pathways are active, leading to a nonlinear spiking response. With crossover inhibition intact, the artifacts of rectification are suppressed, restoring a linear response. This result is similar to responses previously reported in nonlinear (ie Y-

type) ganglion cells [7]. There, rectification in bipolar pathways was shown to underlie the nonlinear response.

We quantified linearity with a measure L (see methods), reflecting the relative response to partial versus full gratings. Histograms of L across 9 cells and 3 different spatial scales are shown in Fig 5.4d: L = 1 corresponds to a strongly linear response, L<0 corresponds to a strongly nonlinear response. The average linearity of OFF ganglion cells decreased under APB by 0.38 ±0.46 (mean ± s.d) $p = 6\times10^{-4}$ (Wilcoxon signed rank test), and recovered under wash to values equivalent to control (p = 0.46). Thus, linearity in OFF ganglion cells is not inherent to the feed-forward pathway that drives them, but is actively maintained by crossover inhibition from the ON system.

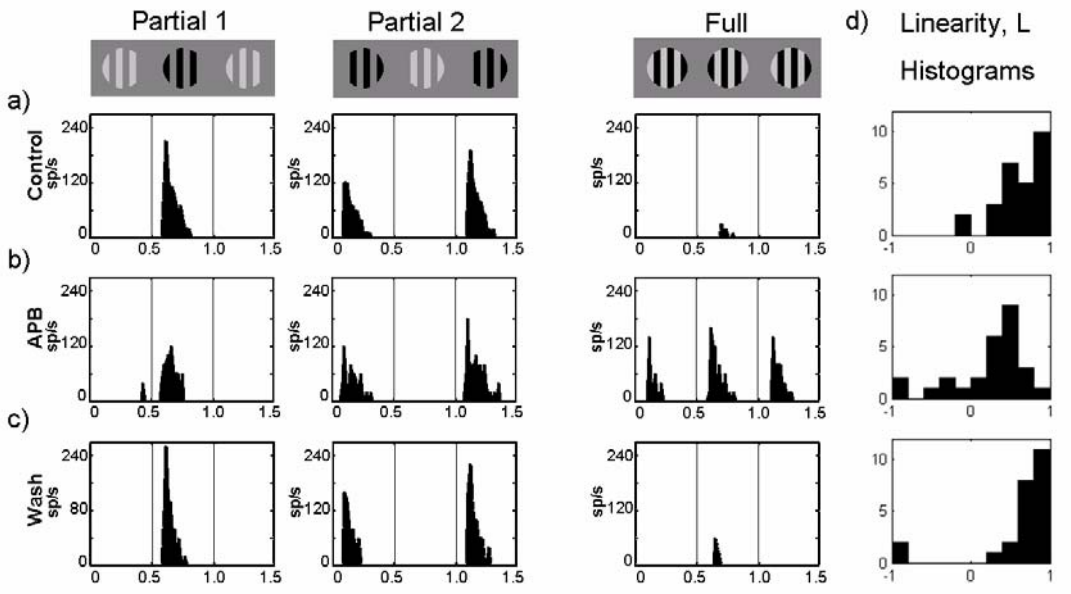


Figure 5.4. Crossover circuitry maintains linearity in the presence of high spatial contrast. Row a) Spiking histograms of a linear OFF ganglion cell which responds to partial gratings (Partial 1, Partial 2) where average brightness changes but does not respond to full gratings (Full) where average brightness is constant. Row b) APB blocks the crossover pathway, converting this to a nonlinear response, with strong spiking in response to full gratings, but little change in responses to partial gratings. Row c) washing recovers linearity. d) Histograms of linearity coefficient L with and without APB (histograms correspond to the same rows as a-c); under control/wash conditions, most responses are linear (L values close to one), under APB responses are much less linear (L values close to zero). The gray images at the top of the figure show 3 frames of the 0.5 sec/frame movie.

Visual signals are rectified by each stage of retinal processing. Between consecutive stages, interactions between rectification and filtering can degrade signal

fidelity. For example, filtering from synaptic dynamics can interact with rectification artifacts in time (Fig 5.3). Similarly, summation of rectified inputs across the processes of ganglion and amacrine cells can generate responses that fail to distinguish very different spatial scales (Fig 5.4). This distortion is suppressed by amacrine cells carrying crossover inhibition to bipolar cells, ganglion cells, and to other amacrine cells. Crossover inhibition provides a compensatory input that, when added to excitation, generates a less rectified, more linear voltage response in each cell type. This form of inhibition prevents succeeding stages of rectification and filtering from permanently corrupting the representation of temporal and spatial features in the visual world.

Methods

Eyes were extracted from rabbits and dissected as described previously[14]. Ganglion cell recordings were made in flat mount[14, 19], while amacrine and bipolar cell recordings were made from 250μm slices. Excitatory and inhibitory currents were recorded under voltage clamp at -60mV and 0mV respectively. Conductance was calculated by dividing these currents by -60mV for excitation, and 60mV for inhibition. Cell types were identified by imaging with Alexa488 and comparing with previous work[13, 20].

Slice mounted cells were stimulated by 200 μm wide light and dark flashes and sinusoidally varying intensities of ±100%, relative to a set background illumination of $3x10^5$ photons/μm^2/s. Ganglion cell receptive field centers were found by flashing spots of varying size. Sinusoids and gratings were sized to match the maximal flash

response for each cell. Gratings took the form of 50 μm, 75 μm and 100 μm stripes of ±100% relative to background.

To quantify rectification, we took the average response 200ms immediately before and immediately after the onset and offset of each flash. Finding this value at the beginning and end of light and dark flashes yields 4 numbers: d_{BL}, d_{EL}, d_{BD}, d_{ED}, describing the magnitude of response to each step in light intensity

The measure of rectification, R, was defined as:

$$R = \frac{d_{BL} + d_{EL} + d_{BD} + d_{ED}}{|d_{BL}| + |d_{EL}| + |d_{BD}| + |d_{ED}|}$$

Amplitude-modulated sine waves were analyzed by taking the Fourier transform of the response, extracting the phases for the fast and slow frequency terms, and taking the difference in phase between excitation and inhibition.

In ganglion cells, we counted total spikes in response to 6 transitions each for partial (P_1 and P_2) and full (F) gratings and subtracted the baseline number of spikes for the same period without stimulus (B). L was defined as the difference between responses to partial and full gratings, normalized by the partial response alone:

$$L = \frac{P_1 + P_2 - F - B}{P_1 + P_2 - 2B}$$

References

[1] A. Kaneko, "Physiological and morphological identification of horizontal, bipolar and amacrine cells in goldfish retina," *J Physiol,* vol. 207, pp. 623-33, May 1970.

[2] F. S. Werblin and J. E. Dowling, "Organization of the retina of the mudpuppy, Necturus maculosus. II. Intracellular recording," *J Neurophysiol,* vol. 32, pp. 339-55, May 1969.

[3] H. Wassle, I. Schafer-Trenkler, and T. Voigt, "Analysis of a glycinergic inhibitory pathway in the cat retina," *J Neurosci,* vol. 6, pp. 594-604, Feb 1986.

[4] E. D. Cohen, "Interactions of inhibition and excitation in the light-evoked currents of X type retinal ganglion cells," *J Neurophysiol,* vol. 80, pp. 2975-90, Dec 1998.
[5] R. C. Renteria, N. Tian, J. Cang, S. Nakanishi, M. P. Stryker, and D. R. Copenhagen, "Intrinsic ON responses of the retinal OFF pathway are suppressed by the ON pathway," *J Neurosci,* vol. 26, pp. 11857-69, Nov 15 2006.
[6] E. P. Chen and R. A. Linsenmeier, "Effects of 2-amino-4-phosphonobutyric acid on responsivity and spatial summation of X cells in the cat retina," *J Physiol,* vol. 419, pp. 59-75, Dec 1989.
[7] J. B. Demb, K. Zaghloul, L. Haarsma, and P. Sterling, "Bipolar cells contribute to nonlinear spatial summation in the brisk-transient (Y) ganglion cell in mammalian retina," *J Neurosci,* vol. 21, pp. 7447-54, Oct 1 2001.
[8] G. D. Field and F. Rieke, "Nonlinear signal transfer from mouse rods to bipolar cells and implications for visual sensitivity," *Neuron,* vol. 34, pp. 773-85, May 30 2002.
[9] D. I. Hamasaki, K. Tasaki, and H. Suzuki, "Properties of X- and Y-cells in the rabbit retina.," *Jpn J Physiol. ,* vol. 29, pp. 445-457, 1979.
[10] J. D. Victor and R. M. Shapley, "Receptive field mechanisms of cat X and Y retinal ganglion cells," *J Gen Physiol,* vol. 74, pp. 275-98, Aug 1979.
[11] D. I. Hamasaki and V. G. Sutija, "Classification of cat retinal ganglion cells into X- and Y-cells with a contrast reversal stimulus," *Exp Brain Res,* vol. 35, pp. 25-36, Mar 9 1979.
[12] J. A. Hirsch, "Synaptic physiology and receptive field structure in the early visual pathway of the cat," *Cereb Cortex,* vol. 13, pp. 63-9, Jan 2003.
[13] M. A. MacNeil, J. K. Heussy, R. F. Dacheux, E. Raviola, and R. H. Masland, "The population of bipolar cells in the rabbit retina," *J Comp Neurol,* vol. 472, pp. 73-86, Apr 19 2004.
[14] B. Roska, A. Molnar, and F. Werblin, "Parallel Processing in Retinal Ganglion Cells: How Integration of Space-Time Patterns of Excitation and Inhibition Form the Spiking Output," *Journal of Neurophysiology,* vol. 95, pp. 3810-22, June 2006.
[15] M. A. MacNeil and R. H. Masland, "Extreme diversity among amacrine cells: implications for function," *Neuron,* vol. 20, pp. 971-82, May 1998.
[16] K. A. Zaghloul, K. Boahen, and J. B. Demb, "Different circuits for ON and OFF retinal ganglion cells cause different contrast sensitivities," *J Neurosci,* vol. 23, pp. 2645-54, Apr 1 2003.
[17] E. J. Chichilnisky and R. S. Kalmar, "Functional Asymmetries in ON and OFF Ganglion Cells of Primate Retina," *The Journal of Neuroscience,,* vol. 22, pp. 2737–2747, 2002.
[18] M. M. Slaughter and R. F. Miller, "2-amino-4-phosphonobutyric acid: a new pharmacological tool for retina research," *Science,* vol. 211, pp. 182-5, Jan 9 1981.
[19] S. I. Fried, T. A. Munch, and F. S. Werblin, "Directional selectivity is formed at multiple levels by laterally offset inhibition in the rabbit retina," *Neuron,* vol. 46, pp. 117-27, Apr 7 2005.

[20] M. A. MacNeil, J. K. Heussy, R. F. Dacheux, E. Raviola, and R. H. Masland, "The shapes and numbers of amacrine cells: matching of photofilled with Golgi-stained cells in the rabbit retina and comparison with other mammalian species," *J Comp Neurol,* vol. 413, pp. 305-26, Oct 18 1999.

Chapter 6

Modeling of Retinal Circuitry

Introduction

This chapter will discuss a variety of modeling results relating to, and explaining the physiological results discussed in chapters 1, 4 and 5. It will start by discussing some ways of modeling components of the retina, and then discuss three general modeling efforts. The first relates to trying to make sense of the diversity of ganglion cell outputs (from chapter 1 and [1]). The second, longest part relates to modeling rectification and cross over inhibition in the retina, and trying to understand the functional benefits imparted by these aspects of retinal function, and this bears on and expands on the findings in chapters 4 and especially chapter 5. Finally there is an attempt to make sense of the asymmetry of inhibition seen in the retina, which bears on the results of chapter 4 as well as some results seen in chapter 1 and some as yet unpublished amacrine results.

A variety of approaches have been used in the modeling that follows, including simple mathematical proofs, and numerical simulation in MATLAB. The goal of this modeling has not been to completely model the behavior of the neurological components of the retina, but to use the simplest sets of assumptions possible to predict, replicate and/or explain particular aspects of retinal activity. Although these simplifications will necessarily reduce the accuracy of modeling predictions, they also reduce the number of free variables that must be included in a model, reducing ambiguity in predictions. Before going into specific modeling results, this chapter stats with several dramatic simplifications that have proven useful in modeling the retina.

Three biological components that must be addressed mathematically when modeling the retina: synaptic behavior, membrane dynamics (including spike generation) and cell morphology. Synapses and spiking are the source of most of the nonlinearity in the retina. Synapses are also thought to underlie most of the diversity of dynamics of the retina (once again the rest is due to spike generation). Cell morphology, meanwhile, relates both to the sort of connectivity that is possible and to the way signals are integrated and processed in space.

Basic synapse modeling

As discussed in chapter 1, synapses involve a strong voltage nonlinearity associated with voltage gated calcium channels, followed by several dynamic processes, involving the release, diffusion, and reception of neurotransmitter [2]. Thus modeling of synapses can be broken down into two distinct aspects: the roughly static

nonlinearity associated with the presynaptic calcium channels, and the dynamics of release and reception.

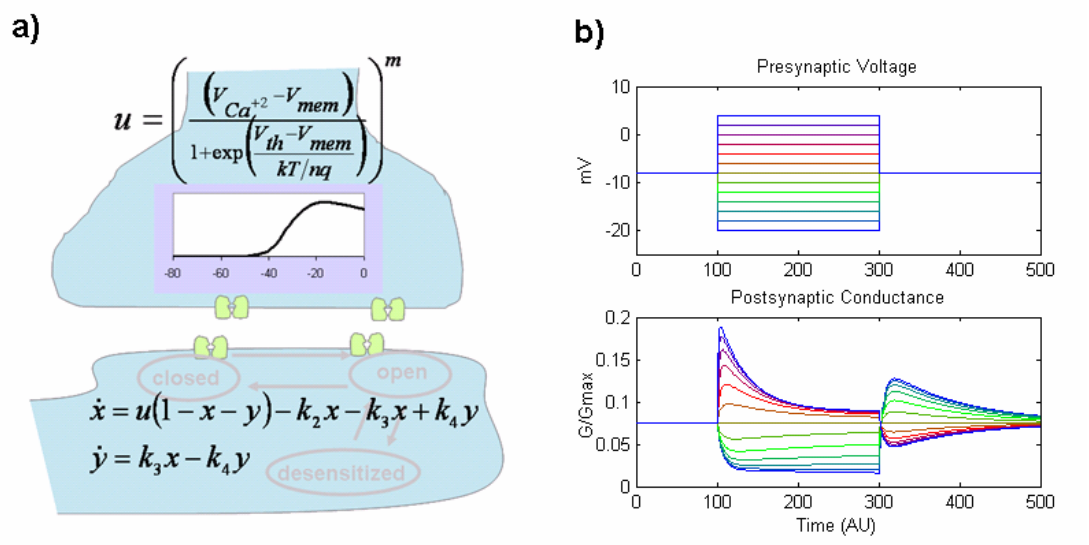


Figure 6.1 basic synapse model: a) Equations defining static nonlinearity u(V_{mem}) and dynamic behavior of receptors. b) MATLAB simulations of response to voltage steps around an operating point of Vth-8mV (n = 4, m = 2, k2 = 0.1, k3 = 0.03. k4 = 0.003).

Presynaptic voltage is translated into a calcium concentration through the Boltzman distribution associated with voltage-gated ion channels. This distribution yields a sigmoidal voltage dependence, shown in Figs 1.4 and 6.1a. This nonlinearity is then concatenated with any additional effects of cooperative binding, either of neurotransmitter to receptors [3, 4] or at other biochemical steps in the release process. Although these additional nonlinearities can change the precise shape of the sigmoidal nonlinearity, they can all be expected to follow either a simple power law or some variant of the Hill equation [5], in either case preserving the sigmoidal shape of the voltage response. A sigmoidal response can act on a signal in a variety of ways, depending on the quiescent operating point of the presynaptic neurons and magnitude of signal. If the cell's quiescent point is close to the inflection point of the sigmoid,

small signals can be treated as simply being amplified linearly, and large signals as being amplified and compressed or clipped at their peaks. If the cell operates away from the inflection point, the response will be rectified. Fig, 5.1 and Fig. 5.2 demonstrate that in the retina, most synapses appear to rectify their signal, and, furthermore, to operate "below" the inflection point, closer to the zero activity end of the curve. This rectification can be modeled by approximating the voltage-response curve either with a 2nd order power series or with a piecewise-linear half-wave rectification (see Fig. 6.2). Both of these approaches have the benefit of containing very few free variables and being computationally straightforward.

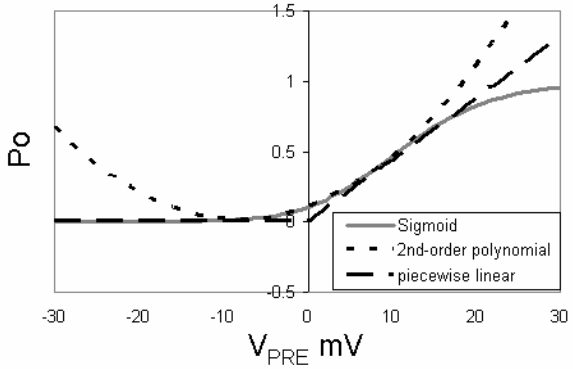


Figure 6.2 Models of synaptic rectification. The sigmoid shown is just the Boltzman distribution with n = 4, Vth = 5mV, squared. The 2nd order polynomial does a good job approximating this curve for 12mV > V_{PRE} > -12mV. The piecewise linear approximation is accurate up to V_{PRE} = 20mV, but is less accurate close to V_{PRE} = 0.

The dynamic aspect of synaptic behavior has the potential to be very complicated, since it can incorporate: 1) the binding, release and replenishment dynamics of the presynaptic vesicle pool; 2) the diffusion dynamics of released neurotransmitter across and along the synaptic cleft; and 3) the dynamics of receptor binding, opening and inactivation. For the sake of simplicity we will only focus on the

last of these effects. Even here, a reasonably complete kinetic model has many states [3, 4] and transitions. For simplicity, we will only address only simple 1st and 2nd order kinetic models, such as is illustrated in Fig 1.6 and Fig 6.1a. Even this very simple model can show behavior that would not be predicted by typical linear time invariant (LTI) models. For example, as shown in Figure 6.1b, the precise value of the input affects not just the amplitude of response, but its dynamics as well.

This behavior is not surprising: the response for a given fixed input level can be treated as a linear time-invariant system with poles at:

$$s = \frac{k_1 u + k_2 + k_3 \pm \sqrt{(k_1 u + k_2 + k_3)^2 - 4(k_1 k_4 u + k_2 k_4 + k_1 k_3 u)}}{2} \qquad \text{(eq. 6.1)}$$

And a zero at $s = k_4$

Both poles (and so, both time constants) depend on the strength of the input, *u*. Thus for different constant input levels, the output will rise and decay at different rates. This also indicates that for inputs that only change slightly, the system can be reasonably modeled as linear time invariant and still capture the essential elements of finite rise and decay of response.

Spike generation can also be well modeled as rectifying[6]. In most of the ganglion cells recorded, baseline spike rate was far below the maximum possible (see Fig 5.2 and Fig 5.4 for example). Indeed, in many OFF cells, the baseline rate was very close to zero [7]. There is also some reason to believe that the precise timing of spiking may be important for coding certain types of information, and that this timing is partially defined by the generation mechanism (for example some ganglion cells seem to fire spikes almost exclusively in bursts of 3 spikes). In most of the modeling

that follows, however, we will neglect this class of coding and look only at simple rate codes.

Morphology modeling

Morphology has two obvious effects upon retinal function. The first is that it defines the possible connections between different cell types. Since each class of cell projects its dendrites and axons to only a subset of the total depths of inner and outer plexiform layers, each cell type can only form synapses with only a subset of other cell types. This is most clearly seen in the division of ON and OFF sublaminae in the IPL, where ON and OFF cone bipolar cells' axons project to the inner and outer halves of the IPL. Not surprisingly, the depth of stratification of ganglion cell dendrites predicts the polarity of excitation those cells receive, such that those that co-stratify with ON bipolar axons receive ON excitation, and those that co-stratify with the OFF bipolar axons receive OFF excitation [1, 8]. Thus physiology can inform the modeling of presynaptic neurons' morphology, and morphology constrains the possible models of physiological connectivity.

The other important way that morphology can affect retinal activity is through the lateral spread of cells' processes. The extent of dendrites defines the area over which inputs are integrated by a given cell. Thus, a widely stratifying cell should tend to respond to inputs across a wide area, roughly corresponding to the extent of its dendrites. Conversely, a narrowly stratifying cell should respond only to inputs over a narrower area. This effect can be modeled as simple summation of inputs across the area of the dendritic tree. An alternate approach is to model this summation with a

Gaussian weighted input, such that each input contributes to the output with a weighting of:

$$\frac{1}{4\pi\sigma^2}\exp\left(\frac{-r^2}{2\sigma^2}\right) \qquad \text{(eq. 6.2)}$$

Where r is the distance of a given point on the retina from the center of the cell's receptive field. This Gaussian weighting captures the concatenation multiple random effects in the connection of synapses and the spread of cell processes. A Gaussian approximation also has the benefit that if one models multiple stages of cells with finite spread, their Gaussian distributions convolve to yield another Gaussian, keeping the math simple. Finally, since we can define $r^2 = (x-x_o)^2+(y-y_o)^2$ and

$$\frac{1}{4\pi\sigma^2}\exp\left(\frac{-(x-x_o)^2-(y-y_o)^2}{2\sigma^2}\right)=\frac{1}{\sqrt{4\pi\sigma^2}}\exp\left(\frac{-(x-x_o)^2}{2\sigma^2}\right)\cdot\frac{1}{\sqrt{4\pi\sigma^2}}\exp\left(\frac{-(y-y_o)^2}{2\sigma^2}\right) \qquad \text{(eq. 6.3)}$$

If one collapses the two-dimensional Gaussian into one linear dimension (by integrating across y, for example), the resulting 1-D receptive field is still a Gaussian with the same space constant σ.

The ease with which Gaussians can be convolved and collapsed is especially useful when dealing with interactions across both the outer and inner plexiform layers (OPL and IPL). Specifically, the effect of horizontal cells interacting with cones can be well modeled by a difference of Gaussians (DOG) [6], where a narrow Gaussian represents the receptive field of a given cone without feedback, and a wider, antagonistic Gaussian models the receptive field of the horizontal cell network at the location of that cone. Since both of these cell types are relatively linear, simply subtracting one Gaussian from the other is reasonable. If this DOG is then convolved with the spread of later cells' processes (also modeled as Gaussian), the result is just

another difference of Gaussians, whose spreads are just the root-summed-squares of the original Gaussians.

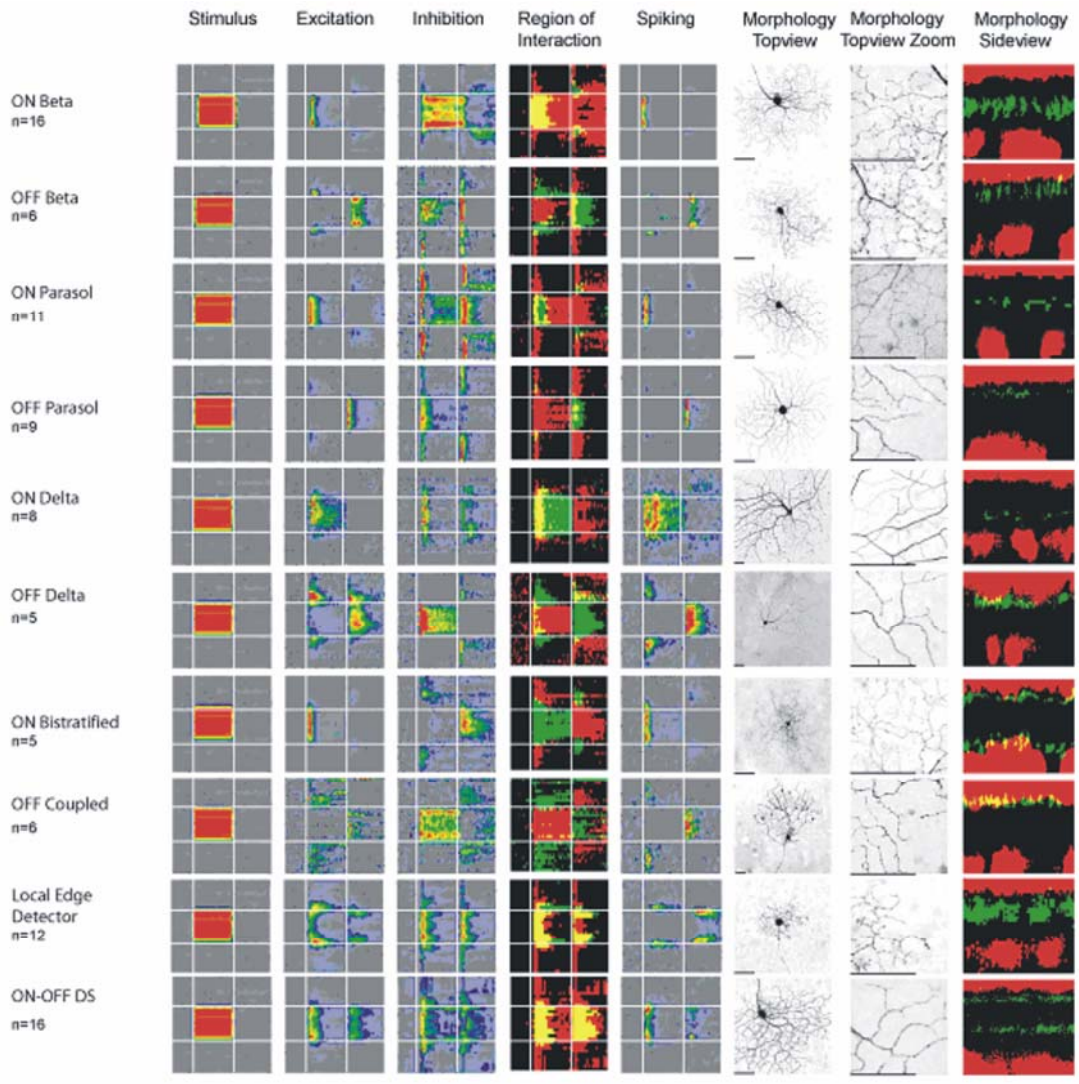


Figure 6.3 Shifted-square (spatiotemporal step) responses for 10 types of ganglion cells and their morphology (from [1]). White lines define the boundaries of the stimulus in columns 1 thru 5. Column 2 is excitation, column 3 is inhibition and column 4 an overlay of excitation and inhibition. Note that for most OFF cell, inhibition looks like an ON response, with little overlap, but that in many ON cells, the

inhibition also has a clear ON component (this asymmetry will be discussed at the end of this chapter). Column 5 is spiking response. Column 6 is the top view of each cell's morphology (scale bar = 50 μm). Column 7 is a close-up of dendrite morphology, and column 8 shows each cells' dendritic depth (green) relative to the INL and GCL (the upper and lower red zones).

Extracting spatiotemporal receptive fields from shifted square data

Previous work in our lab [1, 8] used 1 second flashes of a large (600um) square stepped in 60um increments to characterize the spatiotemporal behavior of Ganglion cells. Plotting the resulting excitation and inhibition and spike-rates yields plots such as shown in Fig 6.3. One thing that is clear from these plots is that different cells types respond at different spatiotemporal scales, and it seems likely that these different cells should responds to different features of the visual scene. What these different features are, however, is less clear, since the responses to the shifted squares overlap in both space and time. We therefore sought to extract simple quantitative models for each cell type.

Ideally, the response properties of each cell would be captured with only a small number of parameters, and then different aspects of the physiology could be correlated with each other and with morphology. A 7-parameter description was used, wherein the cell's spiking rate in response to a light stimulus was assumed to be separable in space and time. Thus, the response of a given cell type to a given stimulus could be found by convolving a time and a space kernel with the stimulus, and introducing nonlinearity only in the form of a final rectification. That is,

$$R(t,s) = rect\left(\left(S(t,s) * k_s\right) * k_t\right) \quad \text{(eq. 6.4)}$$

Where R is the response, S is the stimulus, and k_s and k_t are the space and time kernels respectively

More explicitly, this means

$$R(t,s) = rect\left(\int_{-\infty}^{t}\int_{-\infty}^{\infty}\left(S(t-\tau, s-x)\cdot k_s(x)\right)k_t(\tau)dxd\tau\right) \quad \text{(eq. 6.5)}$$

The parametric description of each cell type, then, lies in the properties of k_s and k_t.

The spatial receptive field (which is the space kernel k_s) of each cell was approximated as a difference of Gaussians:

$$k_s(x) = \frac{1}{\sigma_C\sqrt{2\pi}}e^{-\left(x^2/2\sigma_C{}^2\right)} - \frac{R_s}{\sigma_S\sqrt{2\pi}}e^{-\left(x^2/2\sigma_S{}^2\right)} \quad \text{(eq. 6.6)}$$

Thus special response was described in terms of only three parameters: σ_C, the radius of the center response, σ_S, the radius of the antagonistic surround, and R_S, the ratio of the total effect of the surround to that of the center. Note that the second spatial dimension of the response has been collapsed since the shifted square does not move in that dimension.

The temporal response (that is, the kernel k_t) was approximated for each cell by a delayed difference of exponentials:

$$k_t(t) = \frac{1}{\tau_1}e^{-(t-t_d)/\tau_1} - R_t\,\frac{1}{\tau_2}e^{-(t-t_d)/\tau_2} \quad \text{(eq. 6.7)}$$

Thus temporal responses were described in terms of only four parameters: t_d, the time delay from stimulus to first response, τ_1, the rising time constant, which relates to the rate at which the cells rises toward its maximal response, τ_2, the falling time constant, which relates to the rate at which the cell adapts to a constant stimulus and R_t, the ratio

of the total effect of the rising exponent to that of the decaying exponent. This reflects the LTI synapse model described earlier, with 2 poles and one zero.

To find the 7 parameters that described each cell's response, a gradient descent algorithm was used. An idealized response, found by using equation (6.4) and an S(t,s) of a 600μm, 1 second square flash, was compared to the actual response of a given cell type (both responses were normalized to have equal summed squared responses). The mean-square-error (MSE) was calculated by subtracting the predicted response from the actual response at each point on the space-time map (1240 points in all), and these individual errors were squared and averaged. The parameters were then each varied slightly, and the MSE recalculated. If this new MSE was smaller than the previous, the new parameter values were kept; otherwise, they were reset and a new combination of variations was used. After a number of iterations, starting from a heuristically derived initial guess, this algorithm will tend to settle on the combination parameters that minimize the MSE between the predicted response and actual data.

One benefit of this parameterization was that it allowed us to quantify the different ways in which each cell encodes visual input, and the degree to which these codes overlap. With this in mind, the space- and time- kernels were transformed into the frequency domain (using the Fourier transform). By combining the resulting frequency-domain time- and space- kernels and thresholding at half the maximum gain, the regions of the space-time frequency plane to which each cell type responded best could be plotted. The various steps of the described parameterization are illustrated in Fig 6.4.

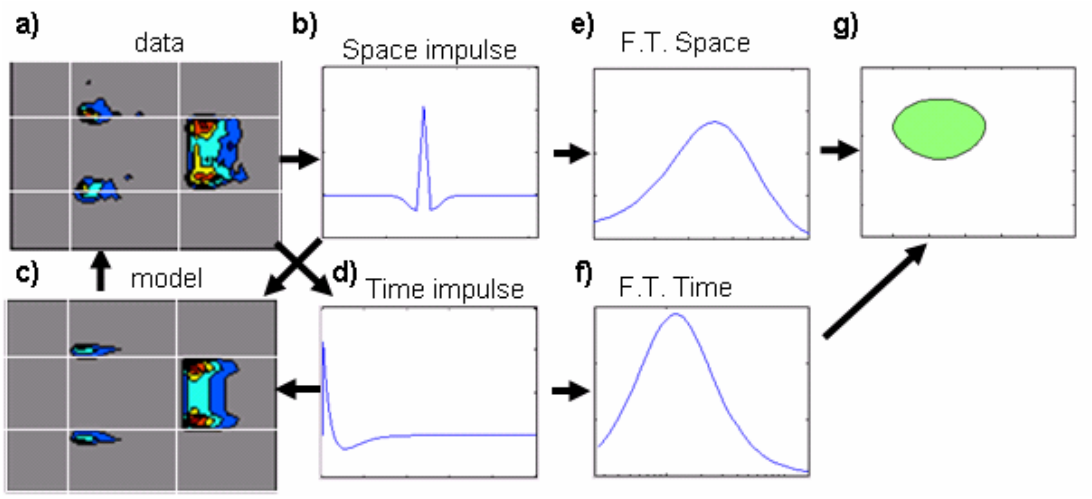


Figure 6.4 Parameter fitting and frequency response plotter. Initial guess at parameters from shifted-square spike data (a) yields space- (b) and time- (d) impulse functions which are convolved with the stimulus and rectified (piecewise linear approximation) to yield the model output (c) which is compared to the data (a) to generate an error signal and improve parameter estimates. Once an error minimum is found, the impulse responses are Fourier transformed to yield time (f) and space (e) frequencies responses. Plotting the 50% power contour yields a frequency response area (g).

This method was performed for all cell types described in [1] except the ON-OFF DS cell (whose response was so nonlinear as to make this approach completely inappropriate). The resulting parameters are shown in table 6.1, and modeled responses are shown in Fig 6.5.

	σ_C	σ_S	R_S	τ_d	τ_1	τ_2	R_t	Err %
OFF coupled	234	446	0.76	91	38	130	1.00	45
OFF delta	137	348	0.94	91	82	174	1.00	12
OFF parasol	94	210	0.69	98	11	52	0.98	19
OFF beta	38	97	0.87	127	51	198	0.95	17
LED	38	71	0.94	173	259	259	1.00	32
ON beta	63	117	0.66	98	29	79	1.02	4
ON parasol	96	297	0.42	88	28	57	1.00	9
ON delta	194	195	0.24	50	51	334	0.75	22
ON bistrat	112	119	0.05	90	44	67	0.91	16

Table 6.1 Fitted parameters for cells shown in figure 6.4.

For most cell types, the resulting fit was quite good, accounting for more than 80% of the "energy" in the data (Error is defined as the final MSE/MSE_0, where MSE_0 is the MSE for a prediction of zero response). Those cells with worse fits generally showed one of two phenomenon: either they showed very noisy spiking (such as the OFF coupled cell) or they showed some overlap between ON and OFF responses, (such as the LED and ON delta), such overlap is inherently nonlinear, and cannot be accounted for in our parameterization.

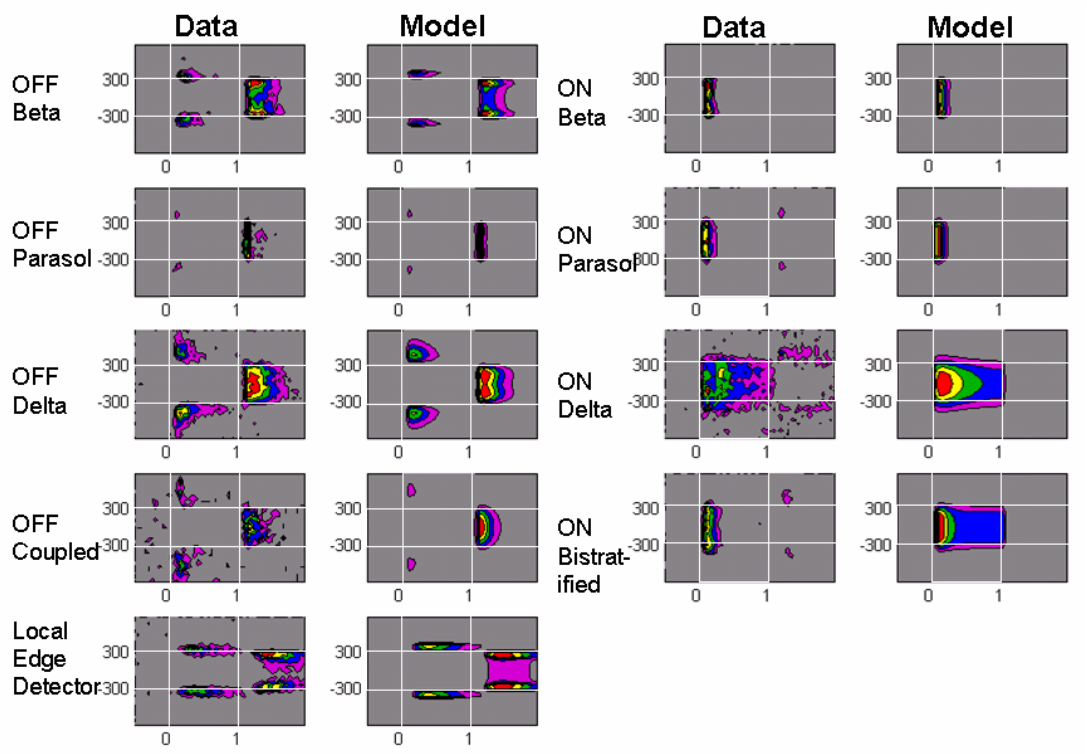


Figure 6.5 Comparison of data and model results for parameters shown in table 6.1.

The center and surround widths were strongly correlated (compare σ_C and σ_S, ρ^2=0.63) as one would expect since both are likely a consequence of convolving the outer retinal response with the cell's dendritic spread, as shown in Fig 6.6a. The diameter of each cell types' dendritic tree is also correlated with the center dimension σ_C (ρ^2 = 0.50) and surround σ_S (ρ^2 = 0.31) of each cells' space impulse response.

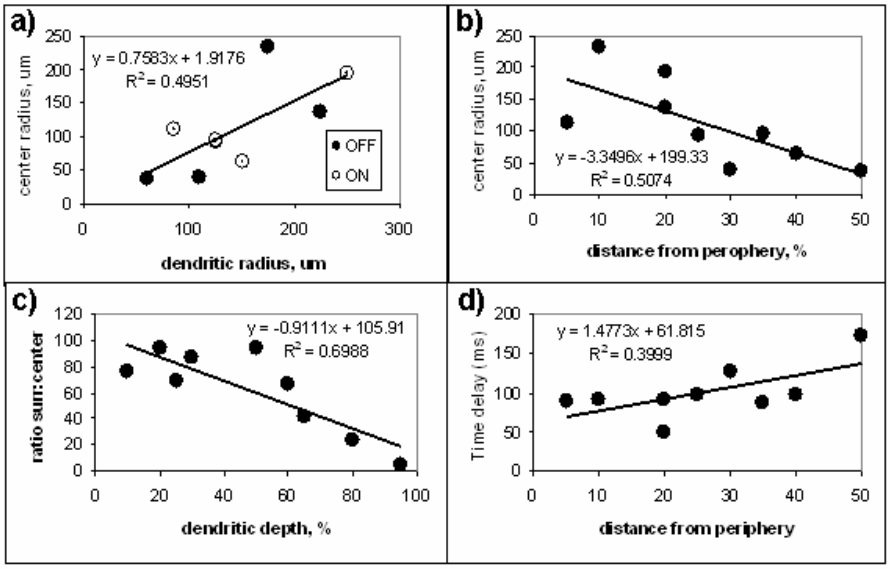


Figure 6.6. Correlations between morphological and physiological responses of ganglion cells.

The center width σ_C of each cell is also correlated with its distance from the edge of the IPL ($\rho^2 = 0.51$) as shown in Fig 6.6b. This is a bit surprising given that the diameter of the dendritic tree is only weakly correlated with this distance ($\rho^2 = 0.13$). Thus, it would appear that something about the architecture of the retina tends to restrict the spatial response of cells near the center of the IPL besides their simple physical dimensions. Also, the strength of the surround relative to the center response, Rs is correlated with absolute dendritic depth, such that in general, the deeper a cell stratifies in the IPL, the weaker its surround, as shown in Fig 6.6c. This may partly reflect an asymmetry in how the ON and OFF systems respond to our stimulus, but it is especially striking in the two deepest stratifying cells, the ON Delta and ON bistratified, where the surround seems almost completely absent. One possible explanation for this asymmetry is that it is due to direct input from Rod Bipolar cells,

which stratify deeply in the IPL and which receive very different surround signal in the OPL than the cone system.

The rise- and fall- time constants were weakly correlated with each other (compare τ_1 and τ_2, ρ^2=0.28), but there was essentially no correlation between the spatial parameters and the rise and fall time. That is, knowing that a cell is wide or narrow field tells you nothing about whether it is sustained or transient. Interestingly, however time delay *was* correlated with the size of the center of the receptive field (compare σ_C and t_d, ρ^2=0.48 and see Fig 6.6d). For the most part, the temporal properties of the physiological responses were also uncorrelated with morphology, showing no particular relationship with either dendritic spread or depth. The only exception is that the delay associated with responses are correlated with the closeness of the dendrites to the center of the IPL (ρ^2=0.4). For the most part Rt was close to one, indicating that responses to sustained stimuli tended to decay back to zero.

Thus, overall, except for receiving some additional delay, probably from transient amacrine cells that stratify centrally in the IPL, the nature of temporal responses is independent of morphology. Spatial responses are correlated with morphology, but it is interesting that this is as much a function of depth as it is of dendritic spread.

Turning now to the frequency responses derived from the above parameters, shown in figure 6.7, several aspects of the results are striking: for OFF cells at least, the different regions of response overlap only partly. They also generally fill in a region of the space-time frequency plane corresponding to frequencies from 1/4Hz to 16Hz and 75μm to 1200μm. Hence, the OFF system, at least seems to be recoding the

visual world into parallel pathways that span the different temporal and spatial scales of the visual world in a relatively complete but non-overlapping way.

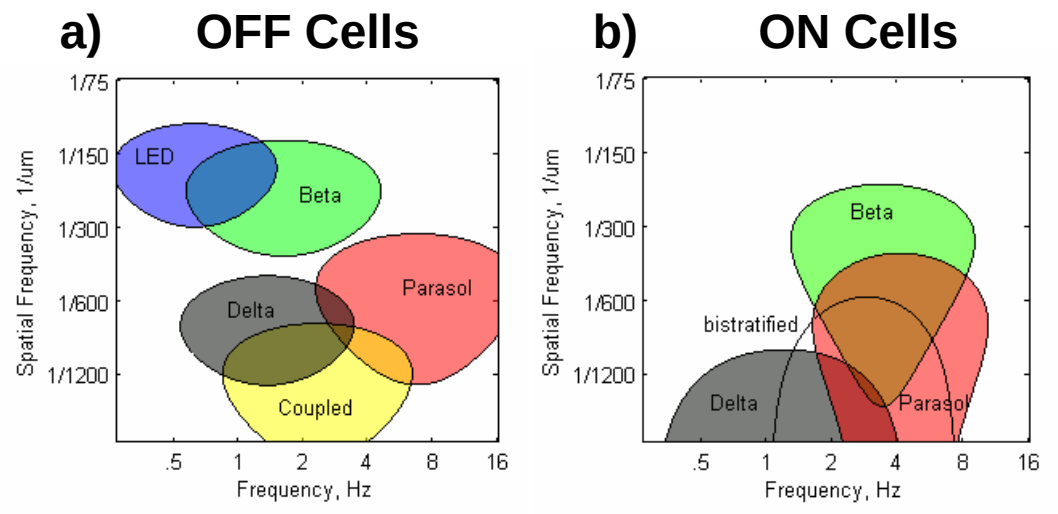


Figure 6.7 Zones of maximal frequency response for OFF (a) and ON (b) ganglion cells.

The ON system is somewhat less straightforward. For one thing, there is a great deal more overlap between response regions. This is especially true of the ON Bistratified cell, whose response seems largely redundant. Given its rather unusual morphology, it seems reasonable to suggest that the ON Bistratified cell actually performs some other, unknown function which is not captured in this analysis, similar to that of the ON-OFF DS cell. The other striking aspect of the ON-cell plot is the large gap in the high-spatial/low temporal frequency corner (the upper-left). This may be explained by noting that for many stimuli, the LED acts ON-OFF and so may fill in this gap. Another possibility is that this gap indicates the existence of an un-characterized cell type.

Modeling cross-over inhibition and rectification.

As discussed in chapters 4 and 5, one of the most common forms of inhibition seen in the retina is "crossover inhibition" wherein inhibition originating in the ON system acts upon OFF cells, and vice versa. This kind of interaction appears at every level of processing, leading us to posit a basic circuit connectivity shown in Fig 6.8. As shown in chapter 4 for bipolar cells, this kind of interaction does not depend on the time scale of stimuli, since excitation stay complementary (zero degrees apart) across a wide range of temporal frequencies. Similar measurements in amacrine and ganglion cells show a similar phase relationship between excitation and inhibition across frequency (see Fig 6.9). Also, excitation and crossover inhibition tend to show similar spatial receptive field sizes in ganglion cells. Indeed, since such inhibition must be carried between ON and OFF layers, it is presumably carried by diffusely stratified amacrine cells, which span both subamina. Previous morphological studies[9, 10] indicate that such diffusely stratified cells are generally not wide field. Also, crossover signals seem to usually be glycinergic in bipolar (chapter 4) and amacrine cells (unpublished work by Ann Hsueh in our lab), and glycinergic amacrine cells tend to be narrow-field [11]. Thus it seems very likely that crossover inhibition is primarily carried by narrow field amacrine cells. The narrow field, time locked nature of this inhibition seems to indicate that its function depends upon it matching the excitatory pathways reasonably tightly. The problem, then, has been to figure out what signal processing function it performs.

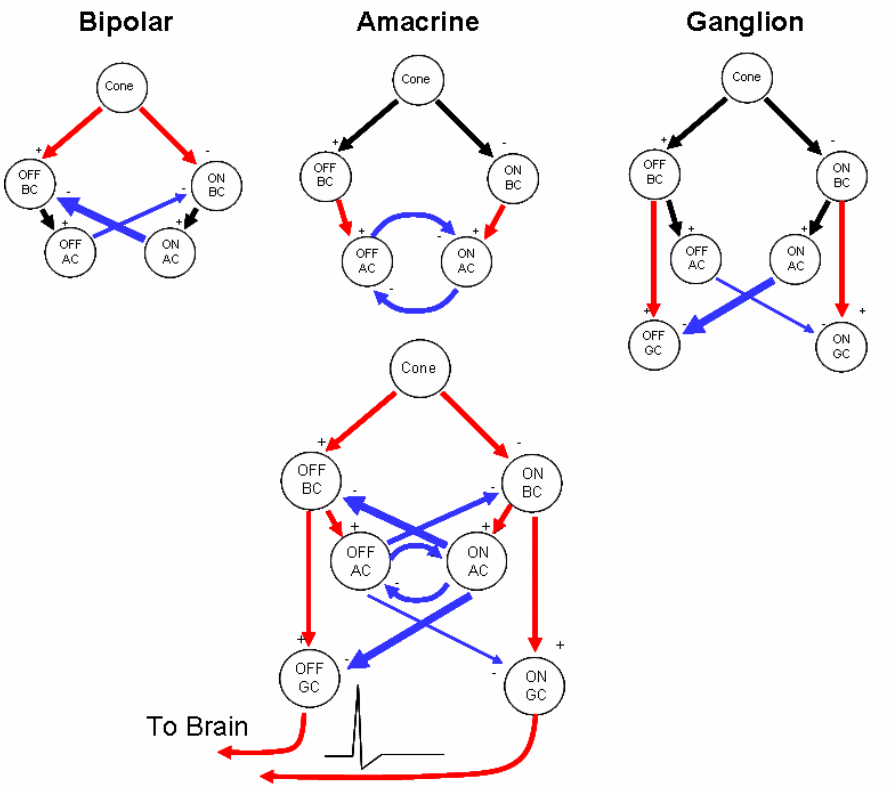


Figure 6.8 Crossover connections in the retina at every level. a) Bipolar b) amacrine and c) ganglion cells receive inhibition from the opposite pathway. In bipolar and ganglion cells, much more inhibition flows from ON to OFF than vice versa. In amacrine cells the flow is more balanced. These forms of inhibition combine (d) so that a given output is likely to have been shaped by multiple rounds of inhibition from the other system.

This structure of having two tightly synchronized pathways of opposite polarity that cross subtract at each layer is very similar to fully differential circuits in analog electronics. What defines such circuits is that they are inherently redundant, and in many cases are not used to perform specific signal processing functions. Instead, differential circuits are used because they are much more robust than single-ended circuits. In general, fully differential circuits act to suppress common mode

signals. Common-mode signals are any signal that appear with the same polarity on both pathways in a differential circuit. Because these pathways are subtracted from each other at each stage, the output reflects the difference between the inputs to that stage, automatically removing any common signal. Common mode suppression is useful for preventing interference from coupled signals, variations in battery voltage, and even-order nonlinearity. This suppression of even order nonlinearity is analogous to the rectification suppression showed in chapter 5.

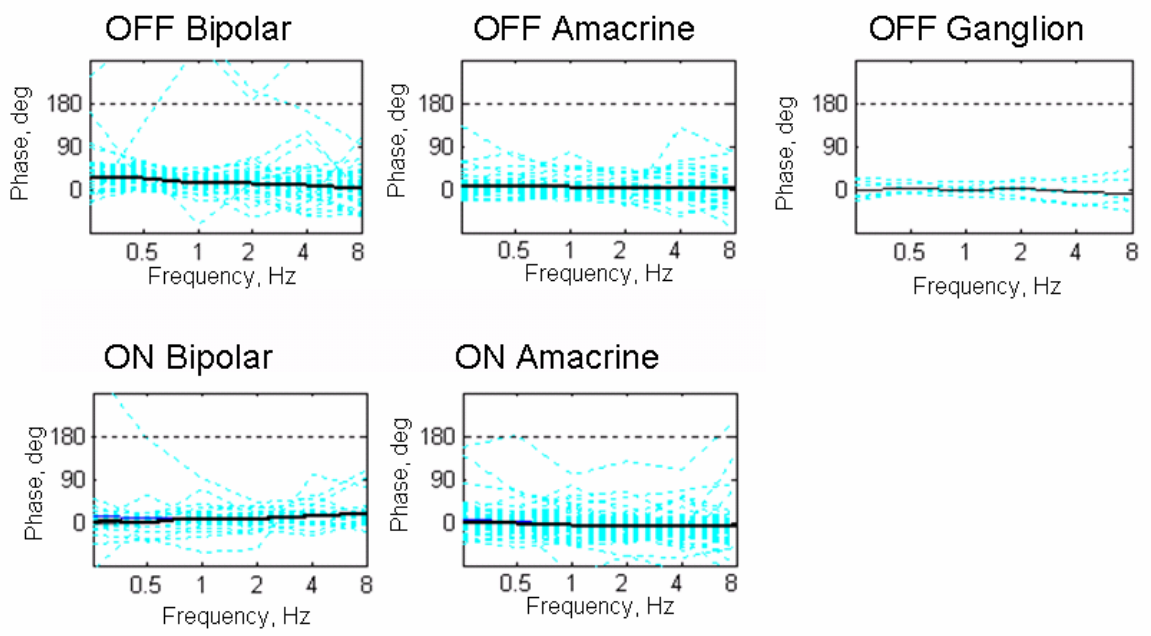


Figure 6.9 Crossover inhibition maintains a zero degree relationship across frequency for bipolar, amacrine and ganglion cells. Blue dashed lines are results for individual cells, black lines are the average phase across all cells shown.

Cross-over inhibition suppresses rectification in simple mathematical models:

The basic idea of how crossover inhibition can suppress simple rectification is shown in figure 6.10. Two half-wave rectifying synapses acting on opposite polarity

signals from the ON and OFF pathways reconstruct the original linear signal when subtracted from each other. This holds for any of the simple rectification models described above, including the piecewise linear, power series and sigmoid approximations, as shown in figure 6.10b-d.

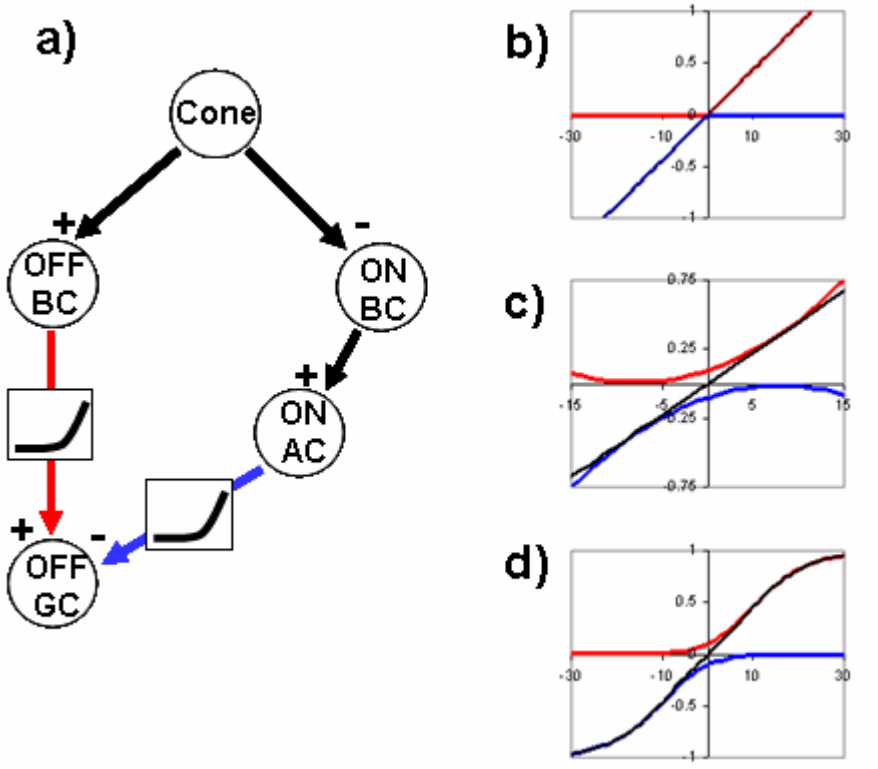


Figure 6.10. Crossover inhibition corrects rectification. a) Schematic of converging rectified pathways: in each model, equal magnitude, opposite-sign inputs are applied to the same rectification function and then subtracted. Piecewise linear (b) and 2nd-order polynomials (c) yield ideal linear combinations. Sigmoids (d) yield another sigmoid with a wider linear range.

This cancellation can be shown analytically for the power-series and piecewise linear cases as follows:

For the piecewise linear case, define:

$x = rect(u) \approx u + |u|$ (eq. 6.8)

In this case, the ON and OFF pathways can be modeled as:

$x_{OFF} = rect(u) \approx u + |u|$ (eq. 6.9)

$x_{ON} = rect(-u) \approx -u + |-u|$ (eq. 6.10)

And the voltage of a cell receiving crossover inhibition is:

$y = x_{OFF} - x_{ON} = (u + |u|) - (-u + |u|) = 2u$ (eq. 6.11)

so the absolute value terms cancel, leaving just the linear terms.

Similarly, for the power series case, define:

$x = rect(u) \approx a_0 + a_1u + a_2u^2$ (eq. 6.12)

In this case, the ON and OFF pathways can be modeled as:

$x_{OFF} = rect(u) \approx a_0 + a_1u + a_2u^2$ (eq. 6.14)

$x_{ON} = rect(-u) \approx a_0 - a_1u + a_2(-u)^2$ (eq. 6.15)

And the voltage of a cell receiving crossover inhibition is:

$y = x_{OFF} - x_{ON} = (a_0 + a_1u + a_2u^2) - (a_0 - a_1u + a_2u^2) = 2a_1u$ (eq. 6.16)

here, the offset (a_0) and second order (a_2x^2) terms completely cancel, leaving just the linear term. This effect is one of the reasons that most radio receivers, which are highly sensitive to second order nonlinerity [12] use differential circuits at every stage of signal processing, since such differential circuits have exactly he same effect. This suppression is contingent upon the ON and OFF pathways being matched (that is, a_2 for both pathways must take the same value). Even with mismatched pathways one still sees suppression, as seen in figure 6.11

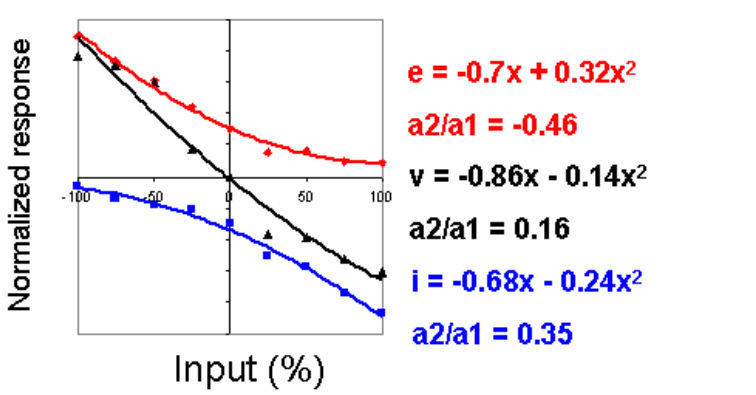


Figure 6.11 Magnitude of excitatory, inhibitory and voltage responses to different magnitude light steps (as a percent of background). Results are normalized such that |peak response|=1 and averaged across 5 amacrine cells.

The 2nd order power series approximation also encompasses the contrast versus brightness dichotomy describe in chapter 5. If one defines "brightness" as the mean light intensity, and contrast as the variance in that intensity, then the average output of a second order power series inherently contains both mean and variance:

$$mean(x) = mean(a_0 + a_1u + a_2u^2) = a_0 + a_1mean(u) + a_2var(u) \quad \text{(eq. 6.17)}$$

Thus, by suppressing even order nonlinearity, crossover inhibition inherently suppresses cells' response to stimulus variance. Indeed, cells which exclusively respond to contrast, such as the ON-OFF amacrine cells that suppress responses to saccades [13] take an output:

$$y = x_{OFF} + x_{ON} = 2(a_0 + a_2u^2) \quad \text{(eq. 6.18)}$$

Suppressing the linear, brightness term and selectively responding to the variance.

Simple rectification models predict temporal contrast/brightness result.

In Chapter 5, Fig 5.3 demonstrated that synaptic inputs to bipolar, amacrine and ganglion cells extract the envelope term of an amplitude modulated sinusoidal

input, confusing contrast and brightness. This result is predicted by both classes of simple rectification, as shown in figure 6.12. This result can also be predicted analytically for the 2nd order power series approximation. If we take the input to be:

$u(t) = \frac{1}{2}(1+\sin(\omega_{env}t))\cdot\sin(\omega_{fast}t)$ (eq. 6.19)

then we know that

$$x = a_0 + \tfrac{1}{2}\cdot a_1 (1+\sin(\omega_{env}t))\cdot\sin(\omega_{fast}t)$$

$$+ \tfrac{1}{4}\cdot a_2(1.5 + 2\sin(\omega_{env}t) + \tfrac{1}{2}\sin(2\omega_{env}t))\cdot(\tfrac{1}{2}+\tfrac{1}{2}\sin(2\omega_{fast}t)) \quad \text{(eq. 6.20)}$$

Which contains the term $\frac{1}{4}\cdot a_2\sin(\omega_{env}t)$, the envelope term shown in Fig 5.3. In this context, crossover inhibition simply combines to signals (excitation and inhibition) whose a_2 coefficients have opposite signs and so act to cancel each other.

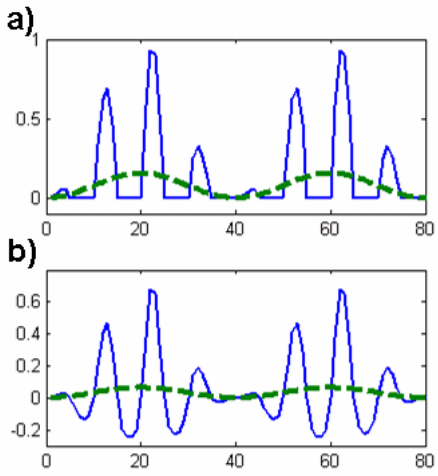


Figure 6.12 Simple models of rectification generate envelope terms. Piecewise linear (a) rectification generates a stronger envelope term (dashed green) than to a 2nd order (b) approximation confined to its monotonic region.

Simple rectification models predict grating contrast/brightness result.

The grating result shown in Figure 5.4 can be seen as a specific sub-case of the mean/variance combination predicted in equation 6.16. Here the mean is being taken across the spatial extent of the ganglion cell by summing up individual bipolar cell inputs. The responses to inverting gratings represent the variance aspect of the response, since the mean is held constant. In order to replicate this result in simulation, a slightly more complicated model is required. If synapses are modeled as simple rectifiers, then they must be preceded by some form of high- or band-pass filtering, which can be taken to represent the dynamics of preceding synapses and/or of light adaptation in the photoreceptors themselves. This effect was modeled by including a 2^{nd} order linear bandpass filter in both the ON and OFF pathways. For simplicity, rectification was only included in the final synapse of each pathway. Although not strictly correct, this approximation can be made because we know it is likely that each earlier stage receives crossover inhibition which corrects its rectified input. Indeed, this assumption was checked by modeling rectification and crossover in the OFF bipolar cell, and got very similar results to those shown below.

The spatial extent of the ganglion cell and of its input cells must also be modeled. Conceptually, this just means that each one receives convergent inputs from multiple presynaptic cells as shown in Fig 6.13a. Mathematically this was implemented by convolving each synaptic input with a Gaussian of the appropriately scaled dimension. Thus, the stimulus was described by an X by T matrix, where X is the number of discrete points modeled in space, and T is the number of discrete points modeled in time. This dimensionality was maintained at each modeled layer of the

retina. Each time step, the effect of spatial filtering was calculated by convolving with an appropriate Gaussian and temporal filtering by updating simple difference equations. A simple difference-of-Gaussians model of the outer retina was included to reproduce basic center-surround behavior. Rectification was modeled as piecewise linear half-wave rectification with a threshold at zero. Thus the final model is as shown in Fig 6.13b.

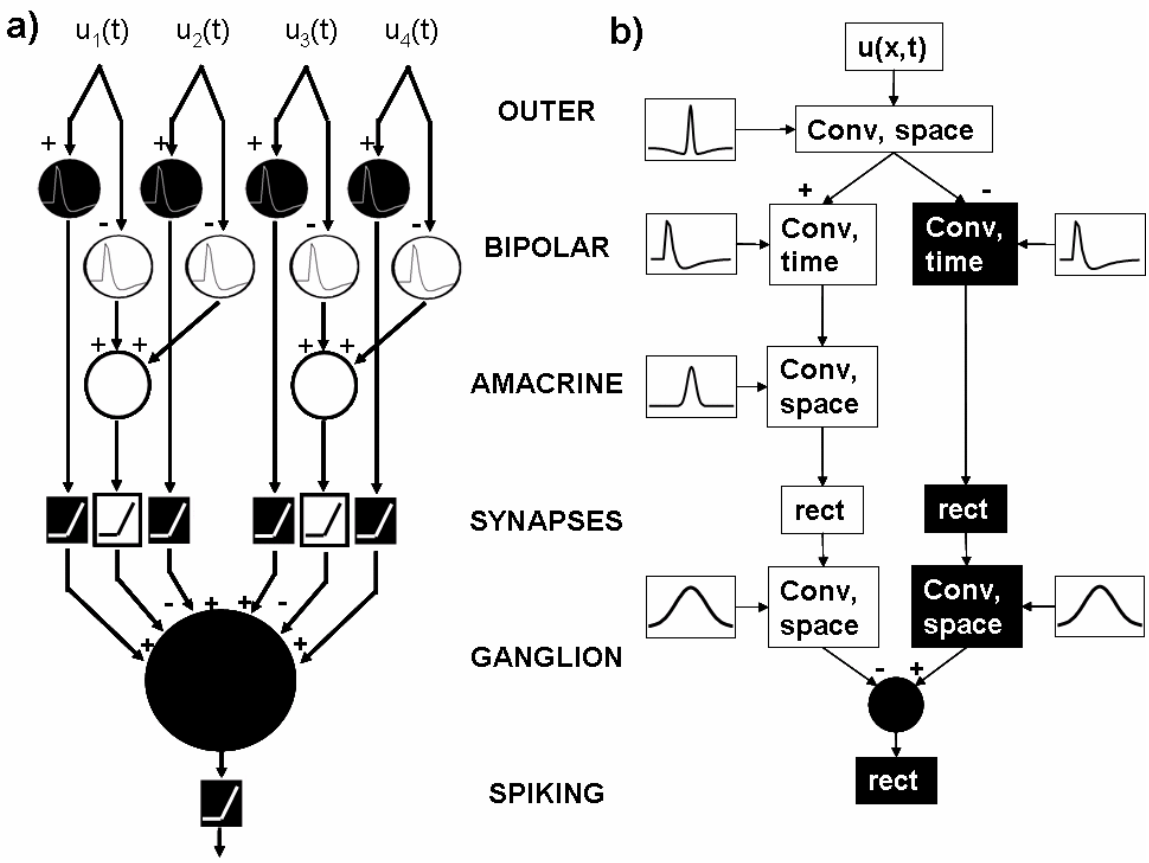


Figure 6.13 Feed-forward models of nonlinear spatial summation. a) Explicit flow diagram of individual cone inputs (u_1-u_4), which are temporally band-pass filtered by ON and OFF bipolar cells and then summed in amacrine and ganglion cells, with the individual ganglion cell inputs and outputs each being rectified. b) Flow diagram using spatial convolution and fine-grain spatial vectors at each stage.

This model predicts both the response to contrast inverting gratings in the absence of crossover inhibition (simply the "excitatory" input), and its suppression in the presence of crossover. It also predicts a second-order effect observed in many cells, which was that not all size stripes of gratings were equally suppressed by crossover inhibition. The degree of cancellation depends upon the size of the dendritic fields of the bipolar and amacrine cells that drive the ganglion cell relative to the grating periodicity. This makes sense, since if those cells extend over an area at the scale of the individual stripes or more, they will tend to blur them together, reducing their response to those stripes. If the amacrine cells carrying crossover inhibition are wider field than the bipolar cells supplying excitation (as they generally are) then one can expect that at grating scales between those of the bipolar and amacrine cells, crossover inhibition will imperfectly cancel the effects of rectification. In this simulation each spatial division was modeled as being ~6μm on the retina. Based on this the diameter, (2σ in the Gaussian) of the ganglion cell and horizontal cells were modelled to be ~200μm, the bipolar cells to be 24 μm, and amacrine cells to be 48 μm. All of which are reasonable based upon morphological structure [10, 14, 15] of these cells. Stimuli with stripe widths equivalent to 48 μm, 72 μm, and 96 μm were used. These results are shown in Fig 6.14 a-f.

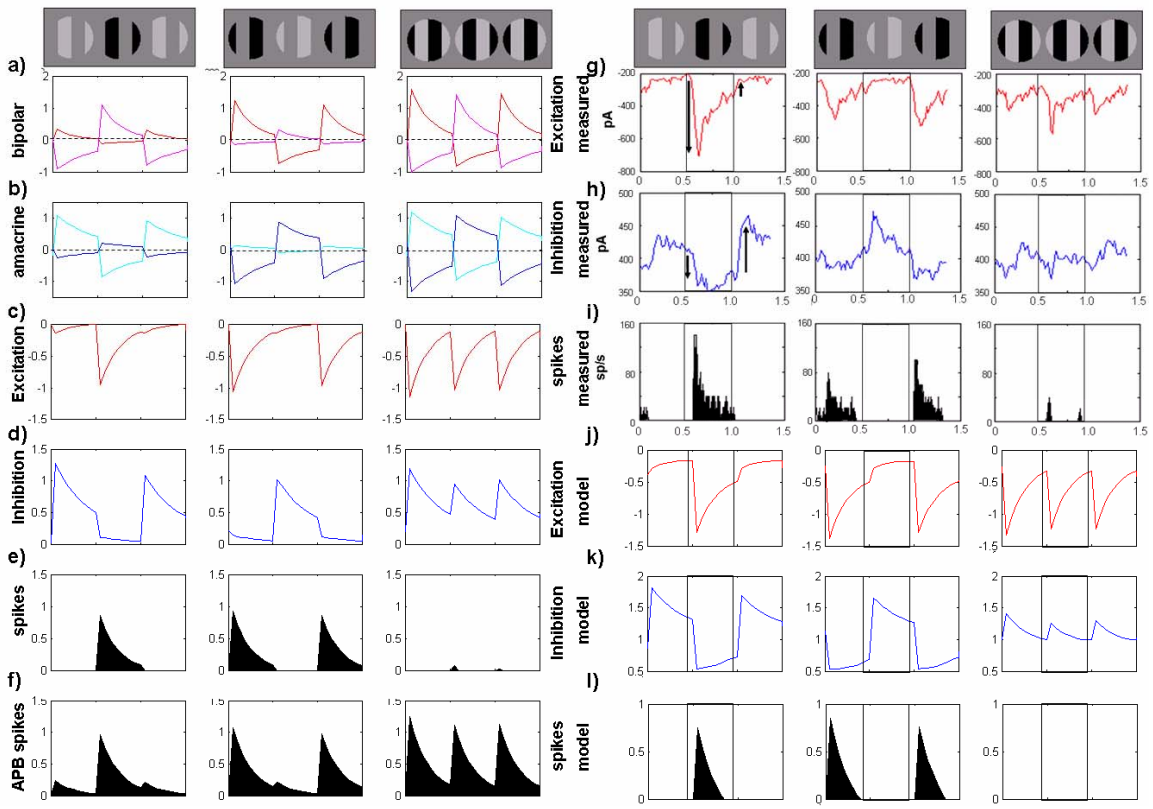


Figure 6.14 Modeling of grating results. Simulations were performed on the model described in figure 6.13b. a)Bipolar and b) amacrine simulated responses show opposite polarities depending which polarity of stripe they are centered on (only two example shown in each case). c)Excitation and d) inhibition take many of these inputs, rectify and sum them with a Gaussian weighting. Thus, each follows roughly the maximum response from (a) and (b). e) Spiking is modeled by summing excitation and inhibition and rectifying again, such that when both are active (3rd column: full grating), they suppress each other. f) When cross-over inhibition is eliminated, modeling the presence of APB, spiking reflects excitation, replicating the measured result if figure 5.4. Actual measurements of g) excitation and h) inhibition show some rectification, (compare arrows for rising and falling edges) but not as strongly as in the ideal case shown in (c) and (d). Nonetheless, i) spiking shows linearization, and

this result can be closely replicated in simulation by including a DC offset before rectification in excitation and inhibition, as shown in j-l). Note that simulation time and magnitude scales are arbitrary and set to provide 10 time steps between stimulus transitions, and yield response magnitudes close to 1.

The simulation results were also compared with actual recordings of excitation and inhibition in an OFF Beta cell (shown in Fig 6.14 g-i). The main observation here is that rectification is not as complete as in the simple model described above. However, simply adding an offset before rectification (effectively weakening it) were simulated results were generated that were very similar to direct measurement (Fig. 6.14, j-l) and still showed linearization of spiking.

Rectification also confuses edge location in both simulation and measurements

As mentioned earlier, one of the sources of error when modeling shifted spot stimuli with a linear kernel is that sometimes the spiking response will tend to bleed across the edge of the square at both light ON and OFF. This apparent ON-OFF response at the edge cannot be modeled by a linear kernel, but can be replicated by including rectification between the initial outer retina difference of Gaussians and the subsequent convolution of that response with the ganglion cell's dendritic spread (modeled as another Gaussian). This was demonstrated by using the same model as shown in figure 6.13, but stimulating with a 600 μm square, as shown in figure 6.15. This result also leads to the prediction that if a ganglion cell receives crossover inhibition, then the elimination of that inhibition will tend to exacerbate this nonlinear

bleeding across the edge. And indeed simple experiments with APB do seem to show this.

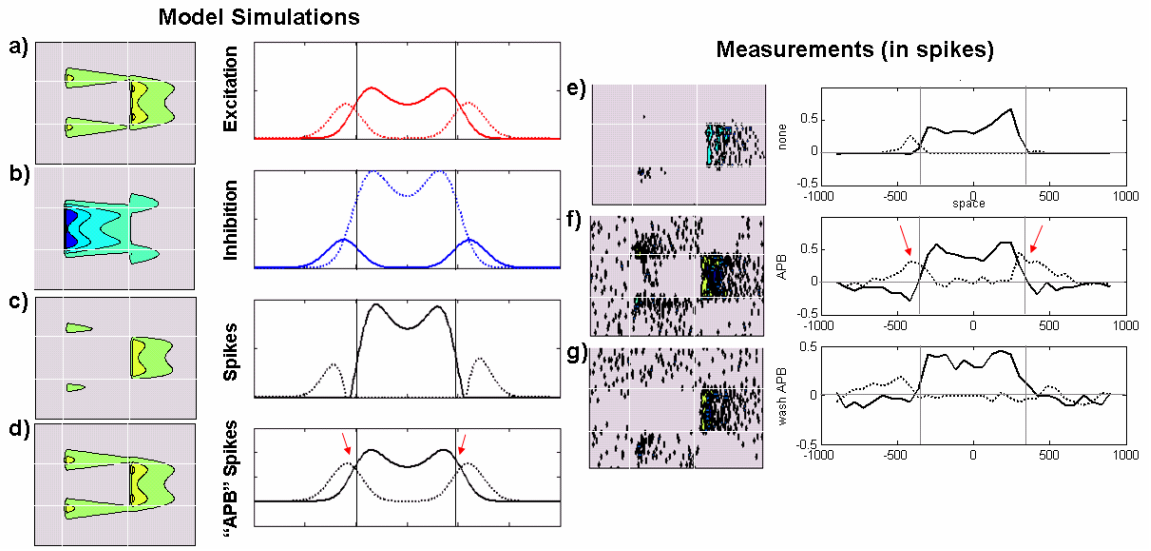


Figure 6.15 Crossover inhibition maintains edge coding. Using the model in Figure 6.13b, stepped-square stimuli of the sort used in [1, 8] and shown in figure 6.3 were simulated leading to the simulated a) excitatory, b) inhibitory and c) spiking responses shown in the left column. Extracting just the spatial aspect of this response gives the adjacent traces, dotted lines for the onset of the flash, solid for the end. Both the excitatory and inhibitory inputs bleed across the edge of the stimulus (marked with vertical lines) but suppress each other, yielding a more confined spiking response (c). Removing the inhibitory input results in spiking bleeding across the edges. These simulation results were confirmed in spiking measurements. Under normal conditions (e) the onset and offset responses do not much overlap, f) but under APB, which blocks the ON inhibition to OFF cells, these responses bleed and overlap (red arrows). This effect washed out (g) under normal Ames' solution. Note that response is adjusted to account for baseline spiking, resulting in some negative responses in the spatial profiles.

Rectification combined with high-pass filtering can destroy information in pseudo-differential systems

It is the nature of rectification to destroy information in a signal. Put simply, piecewise linear half-wave rectification preserves the positive half of a signal, but destroys the negative half. This loss of information is compensated by maintaining both ON and OFF pathways, since one preserves the positive half of the signal, the other the negative half. Furthermore, multiple stages of this sort of rectification would seem not to have any additional effect, once one half of the signal or the other has been extracted. However, as shown above (and in chapter 5) interspersing filtering with rectification can lead to extra artifacts not present in the original signal. This does not necessarily mean, however, that interspersing of filtering and rectification actually *destroys* information when both ON and OFF systems are present. All of the artifacts of rectification described above have involved the effect of low pass filtering (averaging) on rectified signals. Although the combination of rectification and lowpass filtering can lead to artifacts in the response, these artifacts tend to add redundant components to the signal without necessarily destroying the original response. In general, if we describe rectification as r(x) = x + |x| then the interspacing of filtering between two stages of rectification can be described by:

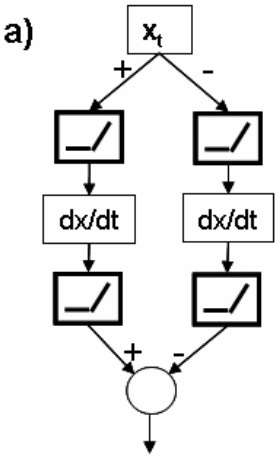
$$y(t) = r\left(\sum_k a_k r\left(x\left(t-k\right)\right)\right) = \tfrac{1}{2}\sum_k a_k\left(x(t-k)+\left|x(t-k)\right|\right)+\tfrac{1}{2}\left|\sum_k a_k\left(x(t-k)+\left|x(t-k)\right|\right)\right| \quad \text{(eq. 6.21)}$$

if $a_k > 0$ for all k, then the sum of rectified signals ≥ 0 for all possible values of x and the subsequent rectification has no additional effect.

If one assumes that the ON and OFF systems are identical and symmetric, artifacts introduced by rectification and low pass filtering can be eliminated by cross-subtracting the two pathways at any point. The implication, then, is that rectification artifacts may make the system inefficient by introducing extra responses that need to be suppressed by cross inhibition at some point in the system, but lack of crossover inhibition would not necessarily do any permanent damage to the signal.

Most synapses actually act as temporal band-pass filters, implying a high-pass aspect to each stage of retinal processing as well as the lowpass and rectification already discussed. What effect does interleaving highpass processing with rectification have? A high-pass aspect implies that an approximation such as in equation 6.21 will by necessity have coefficients with both positive and negative values. We can start with the simplest model of a high pass filter possible, which is a discrete time derivative. If we perform this (linear) operation between two successive rounds of rectification, as illustrated in Fig 6.16, we find that information can actually be destroyed even if both ON and OFF pathways are present.

a)



b)

	x1>0						x1<0					
	x2>0				x2<0		x2>0		x2<0			
	x1>x2		x2>x1		x1>x2		x1<x2		x2>x1		x1>x2	
	ON	*OFF*	*ON*	*OFF*	*ON*	*OFF*	*ON*	*OFF*	*ON*	*OFF*	*ON*	*OFF*
start	x1	-x1	x1	-x1	x1	-x1	x1	-x1	x1	-x1	x1	-x1
	x2	-x2	x2	-x2	x2	-x2	x2	-x2	x2	-x2	x2	-x2
rect 1	x1	0	x1	0	x1	0	0	-x1	0	-x1	0	-x1
	x2	0	x2	0	0	-x2	x2	0	0	-x2	0	-x2
difference	x2-x1	0	x2-x1	0	-x1	-x2	x2	x1	0	x1-x2	0	x1-x2
rect 2	0	0	x2-x1	0	0	-x2	x2	0	0	0	0	x1-x2
reconstruct	0		x2-x1		x2		x2		0		x2-x1	

Figure 6.16 Interleaving rectification with high-pass filtering destroys information. a) Simple model: signal is split into positive and negative (ON and OFF) pathways, each is rectified, differentiated and rectified again. Reconstruction is performed by subtracting the resulting responses. b) 6 possible cases result, depending upon the signs of x1, x2 and their difference. In "rect 1" negative inputs go to zero. The difference is then calculated and if this is less than zero, set to zero by "rect 2". In two of these cases, (marked red) the output is always zero, regardless of the specific values of x1 and x2.

To see is, look at the input for two consecutive time steps. There are four distinct cases that describe how these two inputs can interact with rectification:

1) $x_1>0$, $x_2>0$ → only ON system responds
2) $x_1>0$, $x_2<0$ → ON systems responds, then OFF system
3) $x_1<0$, $x_2>0$ → OFF systems responds, then ON system
4) $x_1<0$, $x_2<0$ → only OFF system responds

now if we assume the effect of filtering is to generate two outputs

$y_{ON} = x_{ON1} - x_{ON1}$

$y_{OFF} = x_{OFF2} - x_{OFF1}$

and we follow this with another round of rectification we find that if $x_{ON2} < x_{ON1}$ then we get no response for the "ON derivative" output, and if this occured when both x_1 and x_2 were positive, then the OFF system won't respond either. Put another way, if the input is positive but decreasing, there is no response in either system, and information is lost. Similarly if the input is negative but increasing, there is no

response in either system. Thus uncorrected rectification can lose information if interleaved with high pass filtering.

This simple model, is, of course, rather over-simplified, especially in its assumption of a hard rectification at each stage. We therefore also tested this idea in the somewhat more realistic model shown in Fig 6.17. An amplitude modulated sinewave demonstrates that without crossover correction, the fast cycles at the later part of the each slow modulation cycle are lost. This result can be understood by seeing that the high-pass aspect of each synapse tends to suppress the low-frequency modulation artifact, but with some phase shift, such that for the later half of each modulation cycle, the synaptic output is actually driven below baseline such that the next round of rectification tends to suppress fast signals in the time period. Since this artifact appears with the same sign in both the ON and OFF pathways, it is suppressed in both, so that the response is entirely lost. Including crossover inhibition suppresses the common mode and recovers the response.

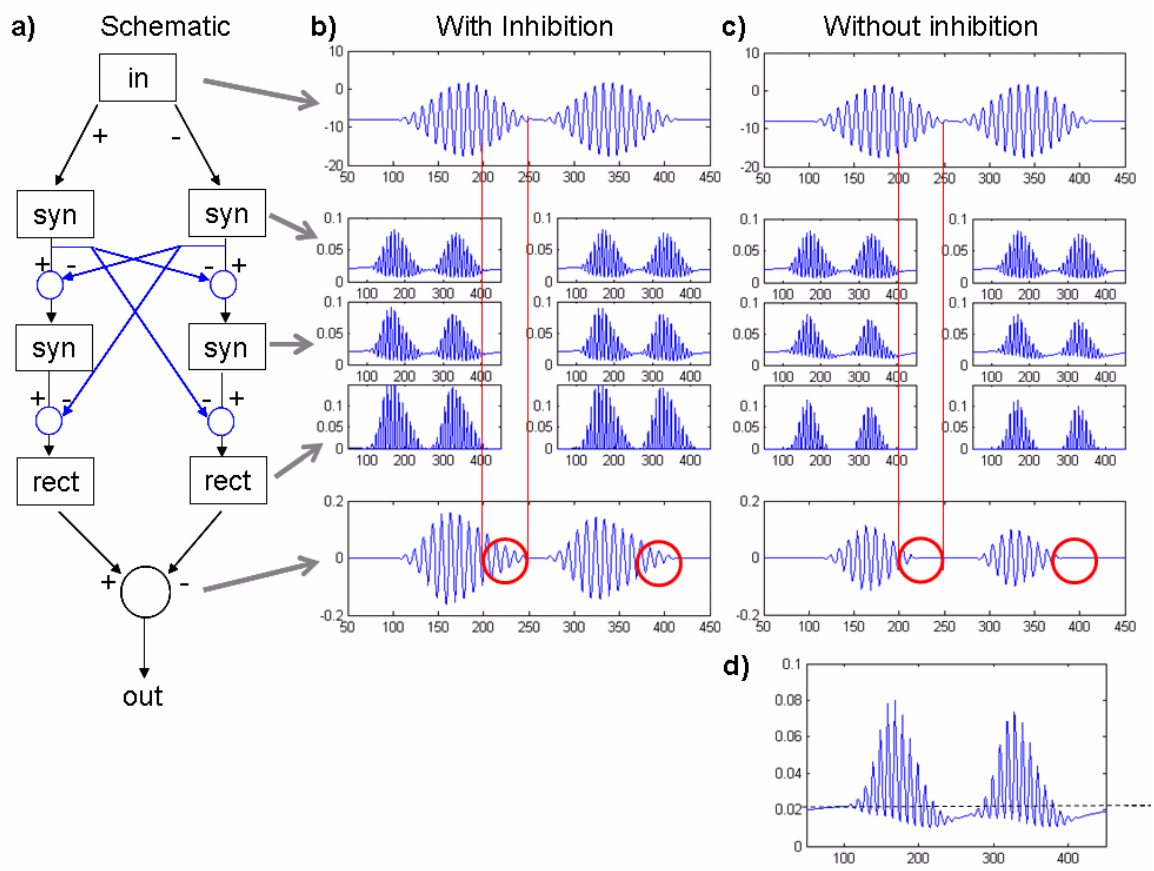


Figure 6.17 High pass filtering combined with rectification destroys information in more realistic models of synapses. a) Model setup: each "syn" block incorporates a static nonlinearity and LTV model as in Figure 6.1, "rect" block model spiking rectification. Blue lines/summing junctions are (idealized, linear) crossover inhibition pathways. b) AM-modulated input simulation results. Each small subplot is the output of a given synapse or rectifier. Larger plots are input and reconstructed output. c) Simulation results with cross-over inhibition turned off. Note that areas in red circles in the output no longer reflect cycles in the input, implying information loss. d) The mechanism of information loss: high-pass filtering of envelope artifact (see figures 6.12, 5.3) reduces response below baseline, causing rectification to eliminate some cycles.

Why rectify?

In chapter 5, it was established that rectification is introduced at nearly every synapse in the retina, and is then corrected by crossover inhibition. The last section established that rectification will introduce a wide variety of artifacts when it interacts with the spatiotemporal filtering inherent to the retina, and that this can cause permanent loss of information if not corrected. Why does each synapse rectify if additional retinal circuitry is then required to correct it? This question is especially of interest since synapse's transfer function is reasonably well described by a sigmoid, which has a region of linear response (see figure 6.1) implying that existing synapses could be deployed in a non-rectifying state of operation. Several mutually compatible explanations are possible.

A synapse with a quiescent operating point close to zero activity, (where it is rectifying) should be more efficient than one operating closer to half-maximum activity. Specifically, the synapse will release less neurotransmitter tonically when in a rectifying state than if it was operating at its linear, inflection point. Unfortunately, this idea is difficult to prove experimentally or with modeling since relinearization after rectification seems to require additional circuitry to maintain linearity, burning a lot of extra energy.

A second explanation is that synapses are general structures that can be used to generate both strongly nonlinear and linear responses depending upon the circuitry they are embedded in. It is certainly the case that certain nonlinear operations (such as saccadic suppression and direction selectivity) require rectification. If the bipolar cell

pathways that underlie these functions are also used in more linear pathways, then rectifying synapses may be the best general component to serve both pathways, and crossover inhibition may be the retina's way of using necessarily nonlinear components to build linear circuits.

It is also possible that rectifying synapses provide inherently superior signaling properties, and this benefit outweighs the additional cost of relinearizing their response with crossover inhibition. It can (and now will) be proven mathematically that for certain reasonable assumptions, a synapse with a sigmoidal release curve will achieve optimal noise performance when operating in a rectifying state.

Signal-to-noise ratio is optimized when rectifiying

To see this, start with the assumption that the dominant noise source in a synapse is the process of vesicle fusion (see Figure 1.5). If we take this fusion to be a Poisson process with the probability of fusion being proportional to calcium concentration, set by equation 1.15, then the variance (noise squared) in release will be proportional to the level of release, or:

$$n^2 \propto \frac{\exp\left(\left(V - V_{TH}\right)nq/\left(kT\right)\right)}{1+\exp\left(\left(V - V_{TH}\right)nq/\left(kT\right)\right)} \qquad \text{(eq. 6.22)}$$

Next we look at the gain of the synapse, which will be proportional to the derivative of release with respect to presynaptic voltage:

$$G \propto \frac{d}{dV}\left(\frac{\exp\left(\left(V - V_{TH}\right)nq/\left(kT\right)\right)}{1+\exp\left(\left(V - V_{TH}\right)nq/\left(kT\right)\right)}\right) = \frac{nq}{kT}\frac{\exp\left(\left(V - V_{TH}\right)nq/\left(kT\right)\right)}{\left(1+\exp\left(\left(V - V_{TH}\right)nq/\left(kT\right)\right)\right)^2} \qquad \text{(eq. 6.23)}$$

The signal-to noise ratio (SNR) for the synapse will then be proportional to:

$$SNR \propto \left(\frac{G}{n}\right)^2 \propto \left(\frac{nq}{kT}\right)^2 \frac{\exp((V - V_{TH})nq/(kT))}{(1+\exp((V - V_{TH})nq/(kT)))^3} \quad \text{(eq. 6.24)}$$

The presynaptic voltage that gives the optimal SNR can be found by taking the derivative of SNR with respect to voltage and setting it to zero:

$$\frac{d}{dV}SNR = 0 = \frac{d}{dV}\left(\frac{\exp((V - V_{TH})nq/(kT))}{(1+\exp((V - V_{TH})nq/(kT)))^3}\right) \Leftrightarrow$$

$$0 = \exp\left(\frac{(V - V_{TH})nq}{kT}\right)\left(1+\exp\left(\frac{(V - V_{TH})nq}{kT}\right)\right)^3 - 3\left(\exp\left(\frac{(V - V_{TH})nq}{kT}\right)\right)^2\left(1+\exp\left(\frac{(V - V_{TH})nq}{kT}\right)\right)^2$$

$$0 = 1 - 3\exp\left(\frac{(V - V_{TH})nq}{kT}\right) \Rightarrow V = \frac{kT}{nq}\ln\left(\frac{1}{3}\right) + V_{TH} \approx V_{TH} - 1.1\frac{kT}{nq}$$

(eq. 6.25)

Which also happens to be the level where equation 1.15 = 0.25, below the inflection point (where V=V_{TH}, and equation 1.15 = 0.5) implying outward rectification.

This argument can be generalized for any sigmoid, provided there is only one inflection point, and P(V) > 0 for all V. In this case,

$$n^2(V) = P(V)$$
$$G(V) = \frac{dP(V)}{dV} \quad \text{(eq. 6.26)}$$

$$0 = \frac{d}{dV}SNR \propto \frac{2\frac{d^2P(V)}{dV^2}\frac{dP(V)}{dV}P(V) - \left(\frac{dP(V)}{dV}\right)^3}{P(V)^2} \Rightarrow \frac{d^2P(V)}{dV^2} = \frac{\left(\frac{dP(V)}{dV}\right)^2}{2P(V)} \quad \text{(eq. 6.27)}$$

Now, we know that for a sigmoid describing the probability of release, the probability will be positive, implying, from equation 6.27, that the second derivative must also be positive. Thus, optimal synaptic SNR will be achieved when the synapse

is rectifying with a positive second derivative, where V is less than the inflection point. This is exactly the region of operation in which we find most synapses operating. Thus combining rectification and crossover inhibition will dramatically increase the dynamic range of a given cell's inputs by simultaneously increasing the operating range of the cell and decreasing noise.

Rectification permits contrast gain control

One place that rectification is used in electronic circuits is in gain control of dynamic signals. In such cases, a rectifier is used to detect the amplitude of signals and this amplitude signal is used to control to a variable gain amplifier. A similar type of gain control is also seen in the retina, where it is generally known as "contrast gain control" [16-18]. This type of gain control responds to changes in the variance of light stimuli. This is distinct from the basic brightness gain control that allows the eye to adapt to changing light levels and which occurs in photoreceptors. It seems reasonable to postulate that synaptic rectification plays a similar role in contrast gain control as it does in volume control in electronics. Indeed, when basic synaptic rectification is combined with the 3-state LTV receptor model described earlier, contrast gain control appears in simulations. This can be understood by first looking at the instantaneous gain inherent to the receptor dynamics:

$$0 = u(1-x-y) - k_2 x \Rightarrow x = \frac{1-y}{u+k_2} \qquad \text{(eq. 6.28)}$$

$$\frac{dx}{du} = \frac{(1-y)(u+k_2) - u(1-y)}{(u+k_2)^2} = \frac{(1-y)k_2}{(u+k_2)^2} \qquad \text{(eq. 6.29)}$$

which is proportional to (1-Iy). y, meanwhile is simply a low-passed version of x, and therefore of u. Thus, gain varies inversely with the average recent level of activity. If u is rectified, then y will reflect not just recent activity, but the variance in that activity, and so will reduce synaptic gain accordingly. This will also introduce a change in baseline activity, of course. However, if this change in baseline is compensated by crossover inhibition, the result will be contrast gain control with relatively little offset artifact. This idea can be tested using a similar model as used in Fig 6.17, and gain control is shown in Fig 6.18, as is the effects of artifacts without crossover inhibition. Furthermore by changing the cell's simulate operating point to the inflection point of their synapses' sigmoids, we can eliminate rectification, and see that contrast gain control is also eliminated. Thus, rectification interspersed with crossover inhibition permits contrast gain control without information loss.

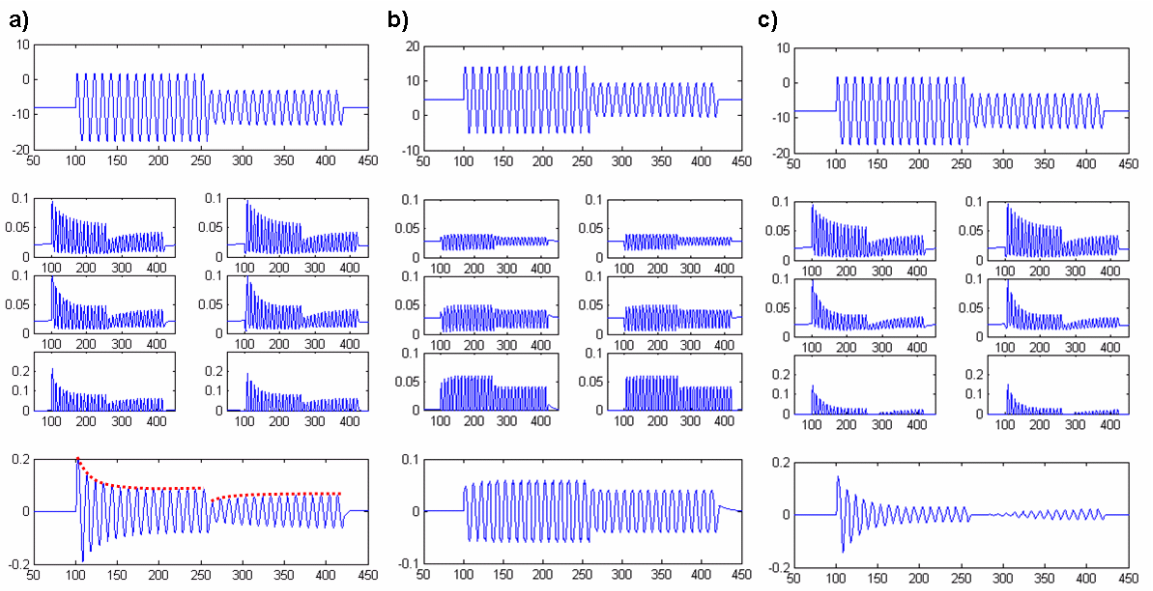


Figure 6.18 Rectification plus inactivating receptors generates contrast gain control. a) Model from Figure 6.17 responds to steps in temporal contrast with gain control. Red dashed lines show changes in amplitude over time. When contrast is first increased, gain decreases, but when contrast is decreased (by 50%) gain increases again. b) Removal of rectification (by setting the operating point of each cell to the inflection point of the static nonlinearity) also removes contrast gain control. c) Removal of crossover inhibition (as in figure 6.17c) keeps gain control, but loses information after step-down in contrast due to shifting baseline.

These results are at least partially borne out by [17] where contrast gain artifacts (offsets in baseline) dramatically increase in the presence of pharmacological blockers of inhibition (eliminating crossover inhibition). Also it has been observed that contrast gain control is stronger in the OFF system where both rectification[7] and crossover inhibition (chapters 4, 5) are more predominant, supporting the idea that these phenomenon are interrelated.

Thus we can see that rectification permits contrast gain control, while crossover inhibition acts to remove gain control artifacts that can completely obliterate the signal.

Reasons for asymmetry in the retina

In chapter 4 it was shown that inhibition to bipolar cells is not symmetric between the ON and OFF systems. Each receives inhibition from the other, but, in addition the ON system receives inhibition from within the ON system. This asymmetry also appears in ganglion cells, where most OFF cells receive ON inhibition, while most ON cells receive a mixture of ON and OFF inhibition (see Fig 6.3). Interestingly, in amacrine cells, this pattern is reversed, such that while nearly all ON amacrine cells receive OFF inhibition, and most OFF amacrine cells receive ON inhibition, many OFF cells also receive OFF inhibition (9/41, vs 2/55 ON amacrine cells receiving ON inhibition). These additional connections are shown in the schematic in Fig 6.19.

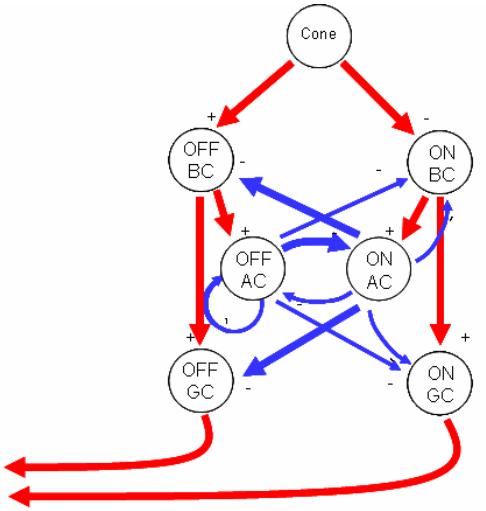


Figure 6.19 Schematic of all of the primary synaptic connections found in the inner retina. Red indicates excitation, blue, inhibition. Thick arrows are thought to e nearly universal (ie, all OFF BCs receive ON inhibition) while thin arrows indicate connections seen in only a subset of post synaptic cells of a given type. Note that while cross over inhibition is fairly ubiquitous, within-layer inhibition appears assymmetricaly between the ON and OFF systems.

This is not the only asymmetry observed between the ON and OFF systems. It has also been observed that in ganglion cells, OFF cells are more strongly rectifying and higher gain [7], while ON cells show a stronger level of baseline spiking, larger receptive fields and generally more transient responses [19]. Although the functional significance of the asymmetry in inhibitory circuitry remains a mystery, it does align with the above observations. Both lower gain and greater transience of response in the ON system make sense in light of the ON to ON inhibition present in ON bipolar cells, which will both suppress response, decreasing gain, and making that response more transient (as shown in Chapter 4).

The brightness distribution in a natural scene is generally not symmetric around its mean. This is reasonable since a given input can at most be completely dark (-100% relative to the mean) but can, in principle, be arbitrarily bright. Thus, for video clips taken in "natural" parts of Berkeley (ie among trees and bushes instead of buildings), the brightness distribution takes forms similar to that shown in figure 6.20a, showing a long tail on the bright end of the range, such that the distribution is skewed. This asymmetry is not necessarily a good representation of what the inner retina will be processing, however, since it is known that the outer retina, and especially photoreceptors will tend to adapt to changes in brightness, in effect whitening the response (removing autocorrelation). This whitening filter, if linear, should take the form of

$$y = u(t) - \sum_k a_k u(t-k) \qquad \text{(eq. 6.30)}$$

where $a_k > 0$, since brightness across successive time steps is generally positively correlated. Such filtering tends to reduce the asymmetry seen by the inner retina, since a give input appear first as a positive and then negative number. This approximation, however, misrepresents the nature of adaptation, which takes the form of gain control, such that a better model would be:

$$y = \frac{u(t)}{\sum_k a_k u(t-k)} \qquad \text{(eq. 6.31)}$$

Where $\sum_k a_k = 1$

such that the basic operating point of y is 1 when u is constant. In this case the asymmetry is preserved, and generally is, as shown in figure 6.20b for two cases:

a1=1, and a1, a2, …a10 = 0.1. Thus, it is reasonable to assume that the visual signal being passed to the inner retina is asymmetric.

This asymmetry in signal distribution may explain several of the asymmetries describe in ganglion cells. It has been demonstrated that if information encoded in a random variable x which is distributed with a density f(x) is recoded to another random variable y with a bounded range of response (say -1 to +1), information is transferred most accurately when y(x) = F(x), the cumulative probability distribution of x [20]. This implies that for the information shown in figure 6.20b would best be recoded by a function like that shown in figure 6.20c. If this signal is then coded by two separate pathways, the ON and OFF pathways, we could reasonably expect that they would distribute this information evenly, which is to divide it at the median. The slope (and thus gain) for the transform below the median is steeper than above, implying that for maximum information transfer, the OFF system should be higher gain than the ON system, which, of course, it is. Also, since the median of the distribution is below both its mean, and more importantly, is lower than its quiescent point (y = 1) then we could expect that quiescent (baseline) level is actually in the part of the distribution coded by the ON system, implying some baseline activity for the ON system, but not the OFF system, which is also born out by measurement [18]. Thus the basic asymmetries in gain and baseline activity between ON and OFF ganglion cells make sense in terms of optimal coding of visual information.

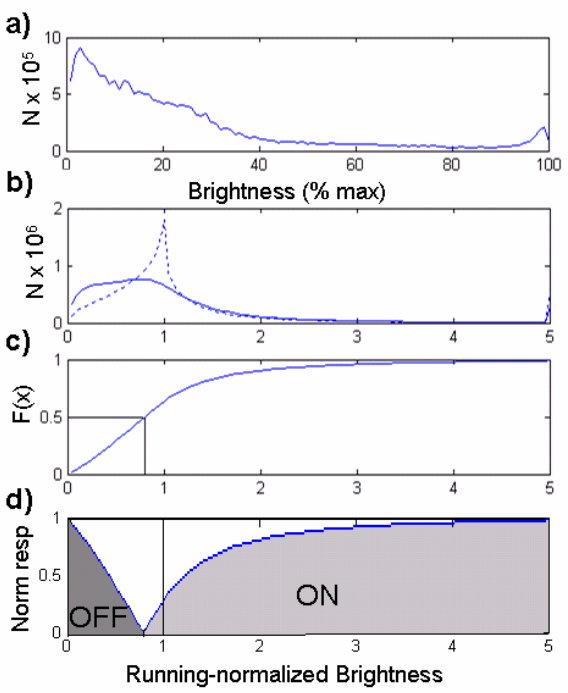


Figure 6.20. Brightness distribution of pixels in a natural scene, and implications for asymmetry in the retina. a) Histogram of brightness from pixels (in time and space) as a percent of maximum. b) Histograms of running-normalized brightness, normalization relative to 1 previous time point (dotted line) or 10 previous points (solid line). c) Cumulative probability distribution for the 10-point case, lies indicate the median, which is clearly less than the quiescent point of 1. d) Idealized ON and OFF response curves based on c. The OFF slope is generally steeper than the ON, but the ON system already shows some response at the quiescent point.

How do these asymmetries relate to the asymmetries seen in retinal circuitry? There are, after all, many ways of attaining this asymmetry in gain and baseline activity. Indeed the basic asymmetries described are already present in the excitatory inputs to the bipolar cells. ON cells receive less rectified excitation the OFF cells (chapter 5), which implies that they must have a stronger baseline level of excitation

relative to excitatory gain. In other words, some difference in the receptors and/or synaptic morphology between ON and OFF bipolar cells will tend to make the ON pathway have lower gain and/or higher baseline activity than the OFF pathway from the outset. On top of this, the actual circuitry seen (as diagramed in Fig 6.19) will enhance this difference: the inhibition to bipolar and ganglion cells from within the ON sublamina will tend to reduce that pathway's gain. At the same time, this feedback will also tend to stabilize the ON pathway's baseline activity since that pathway can only receive ON inhibition if the ON bipolar cells are actively releasing neurotransmitter, implying that ON pathway should stably settle to some level of baseline activity. In contrast, the OFF system, which receives its inhibition from the ON system, is not self-stabilizing. Thus the, the topology present makes sense in terms of the asymmetry seen in ganglion cells, but this does not answer the question of whether this topology is actually any "better" than other possible topologies.

Comparing possible retinal circuit topologies for robustness

This question can be addressed by first considering the basic topologies available to the inner retina. For simplicity this will be limited to the case of two bipolar cells, two amacrine cells and two ganglion cells (one ON and one OFF each). Furthermore, we will assume that excitatory connectivity is fixed, this leaves the possible amacrine connections, of which there are 12 possible connections, or 2^12 possible distinct topologies. The main questions that will be addressed are the relative robustness of these different possible topologies in terms of gain, stability and sensitivity to changes in global activity.

In order to asses these basic properties of these various possible structures, we can describe each connection as linear, all with equal gain. Assuming linearity obviously misses a lot of detail (such as most of the results of chapter 5), and assuming equal gain is a huge oversimplification, but implies that we are only looking at the "strongest" connections, which should all at least have the same order of magnitude of response within a given cell. Because many of the connections are feedback connections, we need to account for the effect of recursion in these circuits. We can do this using a simple matrix description of the system [21]. We define the states of the bipolar and amacrine cells by a vector **x**, and of the ganglion cells by an output vector **y**. Three matrices define the evolution **x** and **y** as a function of the input u. These are b, which defines the direct connections from cones to bipolar and amacrine cells, C which defines the feedforward connections from bipolar and amacrine cell to ganglion cells, and A which defines the connections between bipolar and amacrine cells. The assignment of each of these variables is shown in figure 2.21a. Thus, for example, the circuit shown in figure 6.19 would have the matrices:

$$A = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & -1 \\ 1 & 0 & -1 & -1 \\ 0 & 1 & -1 & 0 \end{bmatrix} \qquad b = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} \qquad c = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & -1 \end{bmatrix}$$

The update equations can be described as:

$$\boldsymbol{x}_{+1} = A\boldsymbol{x} + bu \qquad \text{(eq. 6.31)}$$

$$\boldsymbol{y} = C\boldsymbol{x}$$

It is A that defines the stability of the systems, since it describes the recursive aspect of the system. Since we are fixing the bipolar to amacrine connections, only 8 entries

in A are variable (the 3rd and 4th columns), allowing for 256 distinct arrangements. We can now ask questions about the relative stability of each of these possible topologies. This system will only settle to a stable solution if the eigenvalues of A all have magnitude less than 1. Thus, for a given inhibitory configuration, we can find the largest eigenvalue of A and force the synaptic gain to be $g < 1/\lambda_{max}$. We can then find the steady state gain of the system, which is

$$\boldsymbol{y}_{\infty} = C(I-Ag)^{-1}bg^2 \qquad \text{(eq. 6.31)}$$

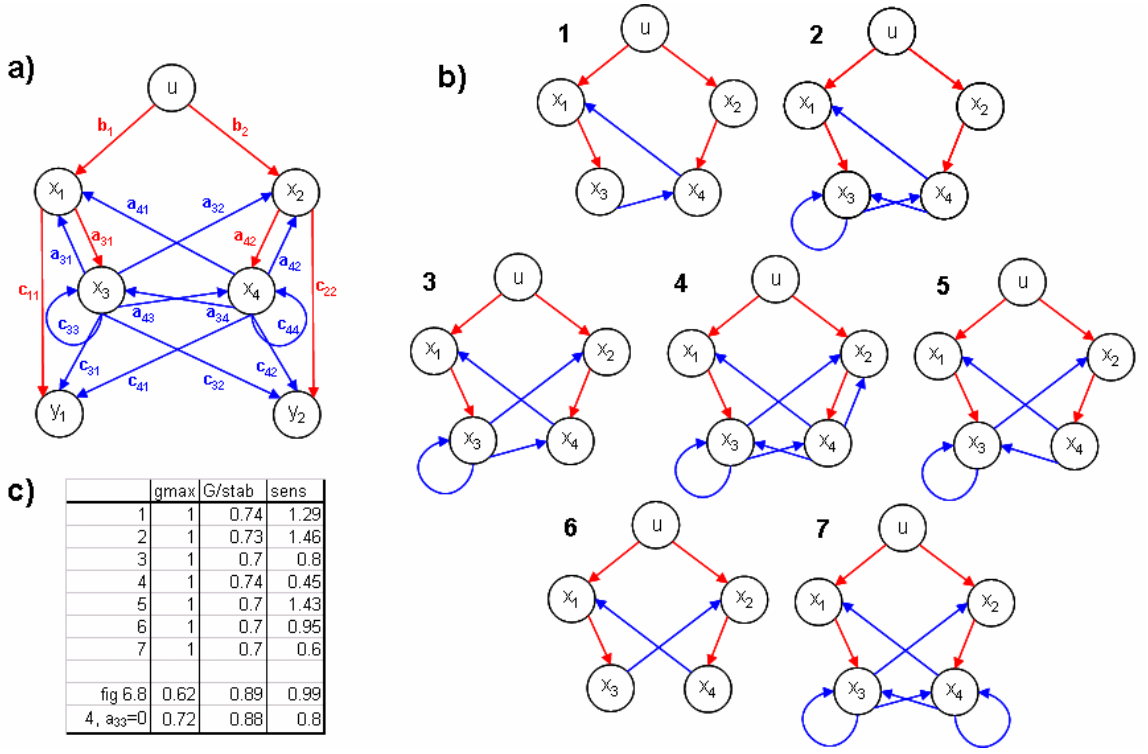
We can also ask for what synaptic gain the topology actually achieves gain beyond the first synapse, such that y1-y2 >2 as the system goes to steady state. In particular, we can ask what the ratio is between the synaptic gain that provides system gain =2 and that which drives the system unstable (we will ignore feedforward inhibitory connections for now). From a robustness point of view, the smaller this ratio is, the better, as this permits the system to have gain and remain stable even with some amount of synaptic gain variation. Of all of the 256 possible topologies, 7 have no recursion and will not be considered further right now. Of the remaining topologies only 12 have a gain ratio of 0.75 or less. And of these, two are symmetric structures and the other 10 appear as pairs of mirror image structures (swapping ON for OFF), of which we need only consider the half with lower ON gain than OFF. Thus we need only consider 7 possible topologies, which are shown in Fig 6.21b. Happily, this set of possible topologies contains one, #4, which is exactly the topology we actually see in Fig 6.19, as well as several sub circuits, #1, #2, and #3, that are also likely to be present (since not all cells receive all of the connections shown in figure 6.19: only the thickest lines shown are probably universal). Note that a pure, complete cross-over

topology (such as shown in Fig 6.8) is not present. This isn't really surprising since multiple levels of feedback crossover inhibition generates positive feedback, very similar to that responsible for the oscillator in Fig 3.2, which is a decidedly unstable structure. It seems that some amount of in-layer inhibition may be necessary to stabilize the generalized crossover we see, and that since some asymmetry is desirable for encoding visual information efficiently, this in-layer feedback is introduced in such a way as to imbalance gain in favor of the OFF system.

We can also use this model to compare the robustness of each of these topologies with respect to variations in the overall excitability of the system. This is done by changing b to be all ones, such that the input is now introduced to all of the bipolar and amacrine cells equally, with the same sign. This is equivalent to shifting their operating pints due to, for example, shifts in metabolic state, mutations in a common ion channel. Next we look at the resulting changes in x and divide this by the steady state gain of the system for each of these cells (for a given synaptic gain, g=.75). We find that of those topologies shown in Fig 6.21 #4, the actual topology of the retina (and its mirror image) provide the best resistance to global shifts in activity level. Thus the architecture of the retina actually provides the most robust structures in terms of resistance to perturbation in both synaptic gain and baseline activity. This also explains the presence of the feedback among OFF amacrine cells, since it enhances robustness in terms of both system stability and sensitivity to baseline. Fig 6.21c lists the gain/stability ratio and sensitivity to global shifts for each of the topologies shown as well as for that universal crossover topology shown in Fig 6.8

and for a topology identical to #4, but with recursive feedback among OFF amacrine cells removed.

So we see that while the polarity of the asymmetry between ON and OFF systems in the retina is probably a consequence of the statistics of visual scenes, the presence of the specific asymmetry in the general connectivity of the retina is probably a consequence of optimizing the robustness of the retina with respect to perturbations in its synapses and cells.



c)

	gmax	G/stab	sens
1	1	0.74	1.29
2	1	0.73	1.46
3	1	0.7	0.8
4	1	0.74	0.45
5	1	0.7	1.43
6	1	0.7	0.95
7	1	0.7	0.6
fig 6.8	0.62	0.89	0.99
4, a_{33}=0	0.72	0.88	0.8

Figure 6.21 Simple linear models of different retinal topologies. a) All possible connections between the ON and OFF systems, each labeled according to their entry in the matrices of an $x = Ax + b$, $y = Cx$ description. b) 7 of the 12 most stable topologies (the other 5 are just mirror images of topologies 1-5. c) Properties of these and two ather topologies, gmax is the synaptic gain for which the system becomes

unstable, G/stab is the ratio of the synaptic gain for which the system has gain = 2, to gmax "sens" is the sensitivity of the topology to global shifts in operating point.

Conclusion

Thus we see that the basic common symmetries and asymmetries in signal flow in the retina can be explained in terms of two basic ideas: linearization of nonlinear components, and robustness in the face of varying parameters. Put another way, the architecture of the retina, just like the architecture of most integrated circuits is constructed such that it is maximally insensitive to the imperfections and variations of its component parts.

References

[1] B. Roska, A. Molnar, and F. Werblin, "Parallel Processing in Retinal Ganglion Cells: How Integration of Space-Time Patterns of Excitation and Inhibition Form the Spiking Output," *Journal of Neurophysiology,* vol. 95, pp. 3810-22, June 2006.

[2] M. J. Zigmond, F. E. Bloom, S. C. Landis, J. L. Roberts, and L. R. Squire, *Fundamentral Neuroscience*: Academic Press, 1999.

[3] A. Destexhe, "Membrane Excitability and Synaptic Interactions," in *Computational Modeling of Genetic and Biochemical Networks*, J. M. Bower and H. Bolouri, Eds. Cambridge, MA: MIT Press, 2001.

[4] B. Hille, *Ion Channels of Excitable Membranes*, 3rd ed. Sunderland Massachusettes: Sinauer Associates, Inc., 2001.

[5] D. L. Nelson and M. M. Cox, *Lehinger Principles of Biochemistry*, 3rd ed. New York, NY: Worth Publishers, 2000.

[6] P. Dayan and L. F. Abbott, *Theoretical Neuroscience*. Cambridge MA: MIT Press, 2001.

[7] K. A. Zaghloul, K. Boahen, and J. B. Demb, "Different circuits for ON and OFF retinal ganglion cells cause different contrast sensitivities," *J Neurosci,* vol. 23, pp. 2645-54, Apr 1 2003.

[8] B. Roska and F. Werblin, "Vertical interactions across ten parallel, stacked representations in the mammalian retina," *Nature,* vol. 410, pp. 583-7, Mar 29 2001.

[9] M. A. MacNeil and R. H. Masland, "Extreme diversity among amacrine cells: implications for function," *Neuron,* vol. 20, pp. 971-82, May 1998.
[10] M. A. MacNeil, J. K. Heussy, R. F. Dacheux, E. Raviola, and R. H. Masland, "The shapes and numbers of amacrine cells: matching of photofilled with Golgi-stained cells in the rabbit retina and comparison with other mammalian species," *J Comp Neurol,* vol. 413, pp. 305-26, Oct 18 1999.
[11] J. Bolz, P. Thier, T. Voigt, and H. Wassle, "Action and localization of glycine and taurine in the cat retina," *J Physiol,* vol. 362, pp. 395-413, May 1985.
[12] A. Molnar, R. Magoon, G. Hatcher, J. Zachan, W. Rhee, M. Damgaard, W. Domino, and N. Vakilian, "A single-chip quad-band (850/900/1800/ 1900MHz) direct-conversion GSM/GPRS RF transceiver with integrated VCOs and Fractional-N synthesizer," in *ISSCC*, San Francisco, 2002, pp. 232, 233.
[13] B. Roska and F. Werblin, "Rapid global shifts in natural scenes block spiking in specific ganglion cell types," *Nat Neurosci,* vol. 6, pp. 600-8, Jun 2003.
[14] M. A. MacNeil, J. K. Heussy, R. F. Dacheux, E. Raviola, and R. H. Masland, "The population of bipolar cells in the rabbit retina," *J Comp Neurol,* vol. 472, pp. 73-86, Apr 19 2004.
[15] R. L. Rockhill, F. J. Daly, M. A. MacNeil, S. P. Brown, and R. H. Masland, "The diversity of ganglion cells in a mammalian retina," *J Neurosci,* vol. 22, pp. 3831-43, May 1 2002.
[16] S. M. Smirnakis, M. J. Berry, D. K. Warland, W. Bialek, and M. Meister, "Adaptation of retinal processing to image contrast and spatial scale," *Nature,* vol. 386, pp. 69-73, Mar 6 1997.
[17] M. B. Manookin and J. B. Demb, "Presynaptic mechanism for slow contrast adaptation in mammalian retinal ganglion cells," *Neuron,* vol. 50, pp. 453-64, May 4 2006.
[18] K. A. Zaghloul, K. Boahen, and J. B. Demb, "Contrast adaptation in subthreshold and spiking responses of mammalian Y-type retinal ganglion cells," *J Neurosci,* vol. 25, pp. 860-8, Jan 26 2005.
[19] E. J. Chichilnisky and R. S. Kalmar, "Functional Asymmetries in ON and OFF Ganglion Cells of Primate Retina," *The Journal of Neuroscience,,* vol. 22, pp. 2737–2747, 2002.
[20] B. A.J. and S. T.J., "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation,* vol. 7, pp. 1129-1159, 1995.
[21] F. M. Callier and C. A. Desoer, *Linear System Theory*. New York: Springer Verlag, 1991.