Perceptual Coherence: Hearing and Seeing

STEPHEN HANDEL

OXFORD UNIVERSITY PRESS

PERCEPTUAL COHERENCE

This page intentionally left blank

Perceptual Coherence
Hearing and Seeing

STEPHEN HANDEL

OXFORD
UNIVERSITY PRESS

2006

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data
Handel, Stephen.
Perceptual coherence : hearing and seeing / Stephen Handel.
p. cm.
Includes bibliographical references and index.
ISBN-13 978-0-19-516964-5
ISBN 0-19-516964-6
1. Sensory receptors. 2. Auditory perception. 3. Visual perception. I. Title.
QP447.H36 2005
152.1—dc21 2005013750

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

To My Family, My Parents,
and the Bowling Ball Cat,
once again

This page intentionally left blank

Preface

The purpose of this book is to describe and explain some of the similarities and differences between hearing and seeing. It is written as an intermediate-level text. It is not mathematical, although it depends on mathematical and analytical thinking. I have tried to walk a line between an overly simplified and an over-the-top presentation of the material.

I think of this text as a "bridge" book in two ways.

The first bridge is between hearing and seeing. It used to be that individuals who studied hearing and seeing thought of themselves as studying perception. Perceiving, with only rare exceptions, involves making inferences and decisions based on information coming from several modalities simultaneously. The choice of using auditory, visual, or tactile input (or combinations) would be based on the particular problem studied. Audition and vision would be model systems, to be employed according to the research question. Currently, the technical expertise required to do research with either sense, and the enormous amount of information about both, have led to a distinct intellectual fissure, with separate journals and professional meetings. The research literature often makes passing references to similar outcomes in other senses, but there is little follow-up.

On top of these experimental issues, I think there is a general belief that hearing and seeing are fundamentally different. I have enumerated many of these differences in table 1.1. Nonetheless, I have always thought that beneath these differences are fundamental similarities in the ways that all modalities make sense of the external world. All events and objects (and perceivers) exist in a common space and time, and all events and objects have a sensory structure that can be picked up by the perceiver. Taken together, I believe that this implies that the internal structures for hearing and seeing are at least qualitatively the same.

There is no single way of connecting the different aspects of hearing to corresponding aspects of seeing. For example, here I connect color to timbre, but another compelling connection would be visual texture to timbre, or color to pitch. Hopefully, the material here will lead readers to consider other possibilities.

Without exception, all chapters contain information about both hearing and seeing. The two chapters that are more exclusively concerned with one sense, chapter 7 about color and chapter 8 about timbre, should be considered as a matched pair. I wrote the chapter about color thinking about timbre and vice versa.

The second bridge is between the introductory materials found in undergraduate sensation and perception, sensory physiology, or basic neuroscience courses and advanced courses covering audition or vision as well as the published literature. I have assumed that readers are not complete novices and that they have had an introductory course, so that many preliminary concepts are not explained fully. I have tried to simplify the figures to emphasize the important points.

There are many excellent introductory textbooks and many excellent advanced texts, and this is designed to slot between the two. Among the advanced texts that I have found particularly useful are Dayan and Abbott (2001), De Valois and De Valois (1988), Gegenfurtner and Sharpe (1999), C. D. Geisler (1998), Kaiser and Boynton (1996), Hartmann (1998), Palmer (1999), Rieke, Warland, de Ruyter van Steveninck, and Bialek (1997), Shevell (2003), and Wandell (1995). These are all more mathematical, and focus on either hearing or vision. My hope is that this book will make the transition to these texts and the professional literature easier.

One of the pleasures of writing a book is the ability to take time to reread books that now are considered passé. I have thoroughly enjoyed Floyd Allport's (1955) treatment of perceptual theories, Georg von Bekesy's (1967) book on sensory inhibition, Julian Hochberg's (1964) slim paperback on perception, and Wolfgang Kohler's (1969) summary of Gestalt psychology. I have also rediscovered the work of Rock (1997), which is discussed at length in chapter 9. I suggest that everyone should read these classics; they are exceptional.

On the whole, each chapter is relatively self-contained. Chapters 1, 2, and 3 cover the basic material and probably should be read first. The remaining chapters can be covered in any order, depending on the interests of the reader.

Many people have contributed to the writing of this book, often unbeknownst to themselves. I would like to thank Dr. Roy D. Patterson for allowing me to spend a sabbatical in his laboratory. Roy's ideas have been the germ for many of the themes in this book: his ideas have become so

intertwined with my own that I am afraid that I have not given him appropriate credit. I am deeply grateful to Dr. Howard Pollio, my colleague in the Psychology Department at the University of Tennessee for 30 years. Howard always has challenged my "mechanistic" explanations and he has forced me to accept the essential intentionality and creativity of perceiving. I am afraid that I will not have satisfied him or myself with what I have been able to write here about either issue. I am also deeply grateful to Dr. Molly L. Erickson and Dr. Sam Burchfield in the Audiology and Speech Pathology Department at the University of Tennessee. Molly has taught me much about acoustic analysis and voice timbre, and has good-naturedly squelched all of my outrageous analogies between hearing and seeing. Sam has been a constant support throughout.

This book has been a tremendous stretch for me and I would like to thank Drs. David Brainard, Rhodri Cusack, David Field, Jeremy Marozeau, and Mark Schmuckler for supplying data, and particularly Drs. Albert Bregman, Peter Cariani, C. D. Geisler, and Paris Smaragdis for patiently answering questions and improving the text. Hopefully they have pushed the book back from the precipice of the Peter Principle. Finally, I would like to thank the staff at the Jackson Laboratory. Doug McBeth and Ann Jordan have processed my reference needs with unfailing good humor and Jennifer Torrance and Sarah Williamson have patiently taught me the finer points of figure preparation in a fraction of the time it would have taken me to figure it out myself.

This page intentionally left blank

Contents

This page intentionally left blank

PERCEPTUAL COHERENCE

This page intentionally left blank

1

Basic Concepts

In the beginning God created the heavens and the earth
Now the earth had been wild and waste
Darkness over the face of Ocean . . .
God said: Let there be light! And there was light . . .
God separated the light from the darkness
Let there be lights in the dome of the heavens to separate the day from the night
And let them be for lights in the dome of the heavens, to provide light upon the earth
God made the two great lights,
The greater light for ruling the day and the smaller light for ruling the night, and the stars.

The beginning of Genesis is perfectly delimited; nothing missing, nothing extra. What consistently intrigues me is the second line, "Now the earth had been wild and waste, darkness over the face of Ocean" (Fox, 1983, p. 4). In the text that follows, God brings order out of chaos. God did not create order from nothingness. It is along the continuum between chaos and randomness to order and structure that our perceptual world forms. Our phenomenal world is not based on the overall level of randomness or order. Rather, our phenomenal world is created by the difference or ratio between randomness and order. Following the initial creation, God made things different: To separate the night from the day God made the greater light and the smaller light. The night is not dark; it is a lesser light. Here again, the phenomenal world is not based on the overall magnitude of light (or sound), but on the difference or ratio between the lightest and darkest or between the loudest and softest. In general terms, this contrast allows us to make sense of a physical world that varies by orders of magnitudes greater

than any single cell of our sensory systems can encode. This contrast allows us to partition the perceptual world into the objects and events that we react to. Moreover, this contrast allows us place objects and events into equivalence categories that are necessary to make sense of the world.

From this perspective and that of Genesis, the opposite of looking at, listening to, or grasping is not blackness, silence, or lack of pressure, but unstructured energy, energy that does not afford the perceiving of things or events in the world. The energy in the physical world and the energy coded by the receptors at the periphery are neutral. Perceiving is not merely attending to parts of the incoming energy, but is the abstraction of the structured energy out of the ongoing flux. It is the interpretation of the physical properties of objects and events. Hoffman (1998) described vision as an intelligent process of active construction; the world is not recovered or reconstructed. The act of looking or listening constructs objects. This is as true for seeing a tree in a snowstorm as it is for hearing a word in a thunderstorm. Perceiving is creative and not passive.

The purpose of this book is to match up auditory and visual perception. Throughout, I take the position that perception is active and that we attend to the structured parts of the world. Therefore, I do not think of *perception* as a noun, but as a gerund, *perceiving*. Looking, listening, searching, overhearing, grasping, touching, manipulating, and so on are the processes of perceiving. These processes are multifaceted. There is no doubt that biological processes exist that transform and code the firings from the peripheral receptors. But, there is no general agreement about how those firings construct the world. On the one hand, the sensory data, if taken over time and space, may have sufficient information to create unambiguous percepts (Gibson, 1966). On the other hand, sensory data may be inherently ambiguous, so that there are necessary inferential and heuristic processes to make sense of every firing pattern. The best strategy would be to make use of cues that are most likely to be correct and have the least variability (Jacobs, 2002). Following Helmholtz (1867), we would perceive what in the past would have most likely generated the sensory data (Purves, Lotto, & Nundy, 2002). It is not necessary or even appropriate to claim a predominant role for any level of processing. Rather, we make use of all levels to create the appearance of things.

All Sensations Belong to Things and Are Understood With Respect to Those Things

As a first guess, visual stimulation is assumed to come from one or more reflecting surfaces of rigid objects moving in three dimensions, and auditory stimulation is assumed to come from one or more continuously vibrating

sources moving in three dimensions. It may be that the visual world consists of light waves passing through transparent surfaces, or that the auditory world consists of pressure waves reflecting off passive objects, but that is not the usual way sensations arise and not the usual way we understand and integrate those sensations. We make use of these usual properties to integrate independent local excitations at the receptors (e.g., the motion of lighted dots, the variation in sound pressure, the brightness patterning of textures) into one or more coherent surfaces and objects. Visual information is "shaped" by the object: the parallel beams of light from a distant source (e.g., the sun) are reflected and shaped into a pattern that signifies the surface and shape of the object. In similar fashion, auditory information is shaped by the object: Air particles are mechanically "pushed around" and shaped into a pattern that signifies the physical properties (e.g., shape, size, material) of the vibrating surface.

Thus, I believe that the usual distinction that vision gives us objects and audition gives us events is a trap. It misleads us into thinking about vision as a spatial sense and about audition as a temporal sense. According to the *Oxford English Dictionary*, the original definition of object is "something thrown in the way," or "to stand in the way so as to obstruct or obscure." Objects are typically opaque, so they block the recognition of other objects that are behind them. In contrast, the definition of events is "to emerge out of a temporal flow." But all perceiving concerns the appearance of things, and things exist in space and time simultaneously. To Griffiths and Warren (2004), object analysis is the analysis of information that corresponds to things and their separation from the rest of the sensory world. To put it differently, all sensory input is interpreted in terms of familiar causative agents or events and not in terms of the manner and nature of sensory stimulation (R. M. Warren, 1999). Raymond (2000, p. 48) makes a similar claim: "the idea is that the brain deals in the currency of object representations, not disembodied stimulus features."

One example of our inclination to perceive sensations as bound to objects occurs with random dot kinematograms, as shown in figure 1.1. Dots are programmed to move as if each were attached to the surface of a transparent cylinder. Even though the cylinder is rotating at a constant speed (A), the observer does not see the dots moving at a constant speed. Instead the observer sees the dots slow down as they reach the edge of the cylinder, stop, and then speed up in the reverse direction as they near the center line of the cylinder (the dots also change size as they move from the front to the back of the cylinder) (B). If the dots did not change velocity or size and simply reversed direction, the perception would be that of a flat surface. Observers effortlessly see the dots moving coherently, and attached to the front or back surface of a rigid cylinder consistent with their direction of movement. What is important is that the observers infer the presence of a cylinder even if

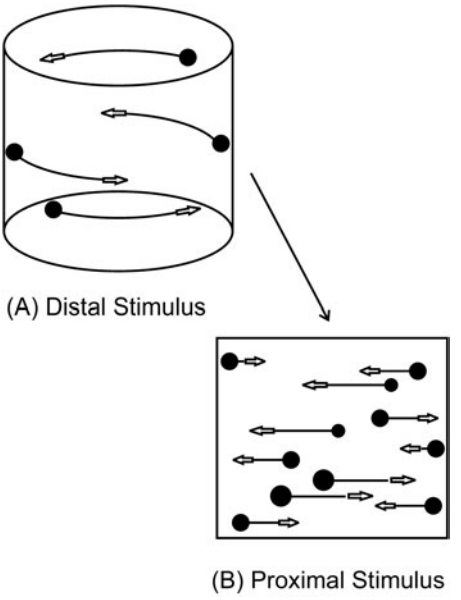

Figure 1.1. Dots are programmed to move as if each was attached to the surface of a transparent cylinder. The cylinder is rotating at a constant speed, so that each dot moves at a constant speed, the distal stimulus depicted in (A). However, the the observer sees the dots change speed and direction as indicated by the arrows attached to each dot. The observer also sees the dots change size as indicated by the size of the dots in the proximal stimulus diagrammed in (B).

individual dots disappear and new ones come into existence. Thus, the perceptual stability and existence of the cylinder surface is created and maintained by the pattern of movement of the dots, yet the temporal properties of individual dots has little effect on perception of surface; the cylinder has a perceptual existence that is independent of any single dot.

Another example of our inherent tendency to perceive elements as part of a three-dimensional object is the classic demonstration of the perception of human figures due to movements created by small lights placed on the joints (e.g., wrists, knees, angles, shoulders). Johansson (1973) dressed the actors in black so that only the lights were visible. When the lights are stationary, they appear to be randomly placed and no form is seen, but as soon as the lights begin to move it is easy to tell whether the actor is running or walking, and even the gender of the actor (Cutting, 1978). It is interesting to note that it is much harder to see the human action if the film is presented upside down (Dittrich, 1993). For both the rotating cylinder and the running person, the three-dimensionality of the immediate percept is based on the pattern of movement of the dots. If the movement stops (or does not resemble plausible biological actions), the percept collapses into a flat random collection of moving dots.

It is worthwhile to point out that the perception of a rotating cylinder or walking dots is based on at least two other implicit assumptions about the world: (1) there is only one light source and (2) it is a single rigid object

even though its appearance changes. The same sort of implicit assumptions occur for the auditory world: (1) there is a single sound source and (2) it is the same source even though its acoustical properties change. The most useful heuristic is to accept the default assumption of one source because in the natural course of time, its properties change due to a slightly different location, orientation, movement, or excitation. Pizlo (2001) argued that all perceiving should be considered as the inverse problem of going from the proximal stimulation at the receptor to the distal object in the world and that all perceiving depends on the operation of a priori probabilities and constraints such as smoothness and good continuation. In Pizlo's view, without constraints, perceptual interpretations (what the proximal stimulation tells us about the world) are not well-posed problems: There may not be a solution, or there may be several solutions. Regardless of whether you believe that the proximal stimulation is interpreted according to evolutionary tuning of the senses to the environment or according to empirical probabilities discovered with experience, or both, the interpretation is that of objects.

The Perceptual World Emerges From Processes at Many Levels

Although our auditory and visual phenomenal world is one of unified objects and happenings, the convergent and divergent auditory and visual pathways (as well as feedback loops from higher brain centers) suggest that the processing of sensory information occurs both simultaneously, in parallel at different neural locations, and successively, serially, as firing patterns converge from these locations. Furthermore, for both hearing and seeing, the initial processing of the physical energy occurs at a local level, not globally. For hearing, the acoustic wave is broken down into frequency components and the receptive cells in the inner ear fire maximally to intensity variation at specific frequencies. For seeing, cells fire to the intensity variation in small regions of the retina and moreover fire maximally to intensity variation that occurs along specific directions (i.e., black-white variation horizontally as opposed to vertically). What this means is that many mechanisms, modules, processing units, or channels (many different words are used to describe these neural "calculators") make use of the same sensory firing pattern to calculate different properties of the object and event.

Although it appears that some properties (e.g., color) are constructed in specific cortical regions, it would be a mistake, however, to think of these mechanisms as being encapsulated and strictly independent. Nakayama (1985) argues that there are several subsystems underlying the perception of

motion and that one or many could be utilized depending on the perceptual demands. Thus, the puzzle is how the various mechanisms are integrated; the problem of analysis is "solved" in terms of the neural circuitry. Each such property enters into the perceiving of many qualities. For example, a motion detection system would enter into the perception of the third dimension, the sense of one's own movement, the detection of impending collisions, and so on. For a second example, the relative intensities of the different frequencies give us pitch, instrumental and voice quality, the sense of an approaching object due to the Doppler effect, speech vowels, and so on. Moreover, there are interactions between vision and audition (see Shimojo & Shams, 2001, for a short review; see also material in chapter 9).

Still another issue is the creative intentionality in perceiving. The organization of light and sound into meaning can usually be done in several ways. The sections below describe some of the heuristics people use to make sense of stimuli. Yet, we all know of instances in which we seem to will ourselves to see or hear something differently. For example, we can make the Necker cube reverse in depth or even force it down into two dimensions; we can listen to an orchestra as a whole or pick out particular instruments; and we can listen to singing phonetically or melodically.

Perceiving Occurs at Several Spatial and Temporal Scales Simultaneously

The first theme stated above explicitly links the perception of bits and pieces of objects to the overall properties of the objects themselves. All of the scales or grains are interdependent due to the fact that they are inherent in the same object or in the same scene. Wandell (1995) argued that we perceive motion with respect to broader "ideas" concerning "dense" surfaces and objects. Julesz (1971, p. 121) made the same argument that the visual system tries to find a global solution in the form of a dense surface instead of localizing points in depth and will disregard, within limits, differences in the disparity values from the two eyes that signify different depths. Bregman (1993) made an analogous assertion for hearing: The auditory system tries to find a global solution in terms of a single source. Namely, we will try to hear a single sound or sequence of sounds as coming from one object. We will break the sound wave into different sound sources only if the expected harmonic and temporal relationships among frequency components that would be created by a single source (e.g., all components should start and stop at the same time) are continuously violated. In the same way that the entire visual scene creates the percept, the rhythmic relationships among frequency components found in longer sequences of sounds will also determine our decision of whether there are one or more sound

sources. A single sound source is the default solution, and the auditory system accumulates evidence before shifting to a multiple-source global solution. Thus, both what we see and what we hear are created at several levels of perceiving. All perception occurs within such a broad context.

Simple examples that illustrate the levels of perceiving are found in photomosaic pictures. Large-scale objects are created by means of arrays of smaller photographs that have the appropriate overall color and brightness to represent features of the larger object. I have a 45×60 cm poster of the Statue of Liberty on my wall constructed from more than 1,000 little photographs. It is possible to focus on the overall shape of the head or on the individual photographs at nearly all reasonable distances from the poster. I am always overwhelmed by the creative possibilities available in perceiving.

The Aperture Problem

Although I have argued above that perceiving depends on multiple stimulus properties that can span spatial and temporal scales, typically we cannot make use of all the available properties at once due to sensory limitations, memory limitations, or even environmental obstacles. For example, cells that code orientation, motion, and shape in the vision system have small receptive fields so that each cell responds as if looking at a very small part of the visual field, and cells that code frequency in the auditory system respond to only a limited set of frequencies so that each cell responds as if hearing only a small part of the signal. It is the convergence of cells at the higher visual and auditory centers that yields cells that respond to larger parts of the field, but the success of that convergence must be due to combining corresponding parts of the field. Moreover, auditory and visual sensations occur across time, and the visual glimpse or auditory snippet at a particular instant must be interpreted by what has preceded it and what will follow it.

The aperture problem is exemplified when looking at the motion of a uniform line through a rectangular opening, as shown in figure 1.2. The problem is that one cannot determine the direction or speed of motion of the line. It could be moving along its own length at any speed, but the restriction of information through the opening makes movement in that direction ambiguous. There are no unique points on the line that allow unambiguous matching from instant to instant. Without some kind of mark on the line, it is impossible to determine if any in-line movement occurred. Regardless of the actual movement of the line, observers simply report the line as moving perpendicular to its orientation without mention of any other motion. That percept minimizes the speed and distance the line seems to move.

What we want to do is represent all possible movements of the line. We start with a straight diagonal line shown in figure 1.2(A). We can represent

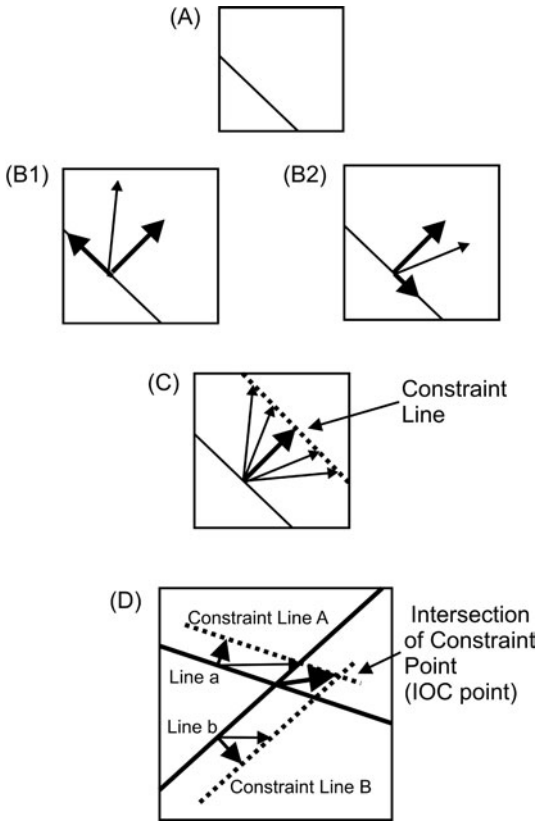

Figure 1.2. The movement of a uniform line (in A) seen through an aperture is ambiguous because it is impossible to see any in-line movement. Two possible in-line movements are shown in (B1) and (B2) and the vector sum of the perpendicular and in-line movement is depicted by the lighter vector. The sum of every possible in-line movements combined with the known perpendicular motion creates a set of vectors that will fall along the constraint line. Four such vectors are shown in (C) as well as the "pure" perpendicular movement. If two lines move simultaneously, the lines are often perceived to move together toward the intersection of the two constraint lines (abbreviated IOC point) (D). The light arrows represent the movement of each line.

any motion by the sum of two vectors at 90° in (B1 and B2): One vector is perpendicular to the line (seen) and the other is along the line (unseen). The length of each vector represents the speed along that direction, but of course, we cannot know the in-line speed. Two possible in-line movements are shown in B1 and B2 (darker lines) and the resulting sum of each of the two movements with the (known) perpendicular movement by the lighter line (a vector). All of the vectors combining the seen perpendicular movements with the possible, but unknown in-line motions, end on a single

straight line parallel to the actual line termed the *constraint line,* the dotted line shown in figure 1.2(C).[1]

Suppose that more than one line is moving within the opening. A downward-sloping line (line a in D) moving to the right would appear to move diagonally up to the right, creating the constraint line A. An upward-sloping line (line b in D) moving to the right would appear to move diagonally down to the right, creating the constraint line B. Observers report that the perceived motion of the two lines together is toward the intersection of the two constraint lines (the IOC point), directly to the right.

I want to argue that the aperture problem is ubiquitous in all perceiving. Our ability to extract the relevant information is always being obstructed in one form or another. In audition, the aperture is not spatial but temporal. In the sense of seeing a visual scene through a slit that allows viewing the scene only as a series of overlapping spatial segments, so too we hear an auditory stimulus only through a temporal slit that allows a series of overlapping temporal segments. In both hearing and seeing, we perceive things by putting together the ongoing overlapping signal. If the aperture is unduly restrictive and reduces the contrast between order and disorder, the perception changes. For example, viewing a uniformly colored surface through an aperture changes the appearance of the color. The aperture reduces the contextual information from the entire scene to brightness and hue information from small spatial areas. The color takes on a filmy appearance and does not appear to be attached to a surface.

The Correspondence Problem

The aperture problem is the cause and complement of the correspondence problem. The visual and auditory sensory worlds are in constant flux (as well as the flux due to eye movements) so that the sensations at any moment cannot unambiguously signify objects or events, and yet we perceive a stable phenomenal world by matching successive visual glimpses and successive auditory segments into stable objects. I have come to believe that the correspondence problem lies at the heart of perception.

The correspondence problem originally referred to the problem of fusing the slightly different visual images in each eye by matching their features. But in the same fashion as argued for the aperture problem, the correspondence problem can be found in nearly all instances of perceiving. Take, for example, exploring a single object using both hands. Here it is

1. If one point on the line is marked, or if one end point is shown, the actual movement can be perceived unambiguously. What happens is the movement of that point is assumed to be true of all the unmarked points on the same line (a rigidity assumption). Palmer (1999) termed this the *unique-point heuristic*.

obvious that the surfaces uncovered by each hand must be placed in registration in order to create a solid object. I can identify five types of problems.

Correspondence Between Binaural and Binocular Inputs

Due to the positioning of the two eyes, the retinal images are slightly displaced spatially with respect to each other, and similarly due to the positioning of the two ears, auditory images are slightly displaced temporally with respect to each other. Thus the problem is to match the visual features in each eye and to match the auditory features in each ear.

The traditional solution for vision was to assume that the image in each eye was analyzed first, so that the correspondence problem was reduced to matching the shapes found in each eye. However, Julesz (1971) demonstrated that binocular correspondence could occur for a wide variety of random-dot stereograms that precluded classic shape matching. The correspondence was achieved by identifying that part of the random array that was common to both eyes. Thus, shape matching is not necessary, although it may occur normally. In the natural world, the correspondence problem can be simplified by making use of the normal properties of real surfaces. Namely, continuous surfaces change slowly and gradually, while discontinuities between surfaces create sharp contrasts.

The traditional solution for hearing is to assume that there are cells sensitive to various time delays created by the outputs from the two ears. Imagine that the neural signal from the near ear is transmitted along parallel neurons so that the signal in each neuron is delayed by an increasing amount of time. Then, each delayed signal is matched against the far ear signal. The match (i.e., the coincidence of the firings) will be maximized at one delay and that delay will signify a direction in space based on head size. Simultaneously, the two firings will become fused into a unified percept.

Correspondence Between Patterns Repeated in Space or Time

Imagine a sequence in which a set of identical but randomly placed dots changes position. We can think of this as a sequence of images, such as the frames of a motion picture. If the motion is rigid, the relative positions of the dots do not change and the correspondence problem becomes matching the dots in one image with those in a later image that represents the same pattern that could have been rotated or translated. If the motion is nonrigid, then the correspondence problem becomes finding the match that represents the most likely transformation. Similarly, imagine a segment of a random sound that is repeated without interruption so that the listener hears a

continuous sound. The correspondence problem is to isolate the repeating segments so that the amplitudes at corresponding time points in each segment are perfectly correlated.

As found for the binaural and binocular correspondences discussed above, the proposed explanations make use of heuristics that reflect the highly probable characteristics of the environment to reduce and simplify the matching problem. For example, one such visual heuristic is that most objects are rigid so that correspondences requiring deformations are given low probabilities, and one such auditory heuristic is that most sounds come from a single sound source that changes frequency and amplitude slowly so that correspondences requiring large changes are given low probabilities. One unresolved issue is what units are being matched. The match could be based on simple elements such as lines, blobs, and individual sounds, or based on geometric figures and rhythmic or melodic phrases.

Correspondences Within One Interrupted Visual Image or Auditory Segment

In our cluttered environment, one visual object is often partially occluded by other objects, yielding a set of disconnected parts, and a single sound is often masked by partially overlapping competing sounds, yielding a sequence of interrupted parts. Here the correspondence problem is whether the parts are separate objects themselves or come from one auditory or visual object.

Correspondences Between Auditory and Visual Information

We see and hear a ball bounce, a person speaking, or a violinist playing. In all such cases, the energy in each modality must be kept in correspondence in space and time. If the information is deliberately misaligned in space (ventriloquism) or time (flashing lights that are not synchronous with sound rhythms), sometimes the information in one modality dominates (we "listen" to the visual dummy and see the lights as synchronous with the auditory rhythm) and sometimes there is a compromise. On the whole, observers are biased toward the more reliable information, irrespective of modality.

Correspondences Between Objects and Events at Different Orientations, Intensities, Pitches, Rhythms, and So On

It is extremely rare that any object or event reoccurs in exactly the same way. The perceptual problem is to decide whether the new stimulus is the reoccurrence of the previous one or a new stimulus. Sometimes, an observer

must judge whether two shapes can be matched by simple rigid rotations or reflections. But often the new stimulus is a more complex transformation of the original one, such as matching baby to adult pictures or matching an instrument or singer at one pitch to an instrument or singer at a different pitch. In both of these cases, the perception of whether the two pictures or two sounds came from the same source must depend on the creation of a trajectory that allows the observer to predict how people age or how a novel note would sound. I would argue that the correspondence problem is harder for listening because sounds at different pitches and loudness often change in nonmonotonic ways due to simultaneous variation in the excitation and resonant filters. The transformation simultaneously defines inclusion and exclusion: the set of pictures and sounds that come from one object and those that come from other objects.

Inherent Limitations on Certainty

Heisenberg's uncertainty principle states that there is an inevitable trade-off between precision in the knowledge of a particle's position and precision in the knowledge of the momentum of the same particle. Niels Bohr broadened this concept by arguing that two perspectives may be necessary to understand a phenomenon, and yet the measurement of those two perspectives may require fundamentally incompatible experimental procedures (Greenspan, 2001). These ideas can be understood to set limits on the resolution of sensory systems. For vision, there is a reciprocal limitation for space and time (and, as illustrated in chapter 2, a reciprocal limitation between spatial frequency and spatial orientation). Resolution is equivalent to the reliability or uncertainty of the measurement; increasing the resolution reduces the "blur" of the property. The resolution can be defined as the square root of the variance of repeated measurements.[2]

For audition, there is a reciprocal limitation between resolution in frequency and in time. To simultaneously measure the duration and frequency of a short segment, the resolution of duration restricts the resolution of the spectral components and vice versa. Suppose we define the resolution of frequency and time so that $(\Delta F)(\Delta T) = 1$.[3] Thus, a temporal resolution of 1/100 s restricts our frequency resolution to 100 Hz so that it would be impossible to distinguish between two sounds that differ by less than 100 Hz. Gabor (1946) has discussed how to achieve an optimal balance between frequency and space or time uncertainty in the sense of minimizing the

2. In chapter 9, we will see that resolution also determines the optimal way to combine auditory and visual information.

3. In general, $(\Delta F)(\Delta T) = \text{constant}$, and the value of the constant is determined by the shape of the distributions and the definitions of the width (i.e., the resolution) of the frequency and time distributions.

overall uncertainty. Gabor argued that Gaussian (sinusoidal) distributions of frequency and time are optimal because the product of their uncertainties is a minimum: $(\Delta F)(\Delta T) \geq .07$. Actually, human performance can be a bit better than this physical limit (Hartmann, 1998).

One way to conceptualize inherent uncertainty is to imagine a simple *x*-*y* coordinate system in which the *x* axis represents frequency and the *y* axis represents duration. If there was no uncertainty, then any tone could be represented by a single point in the *x*-*y* space. But because there is uncertainty, the best we can do is create a minimum rectangular area in the space so that the width along the *x* axis represents the frequency resolution with the height along the *y* axis representing the duration resolution. If we want to measure both frequency and duration with equal resolution, then the area becomes a square. The receptor will not be able to resolve combinations of tones within that square. If we want to measure frequency with greater resolution, then the square becomes a vertical rectangle so that the *x* width gets smaller, but the *y* height (i.e., resolution) must increase to maintain the same area. Similarly, if we want to increase the resolution for duration by making the *y* height smaller, we must necessarily decrease the resolution for frequency by making the *x* width longer to maintain the same rectangular area (Daugman, 1985).

Figure 1.3A illustrates the joint uncertainty arising from spatial frequency and spatial orientation as discussed in chapter 2. Figure 1.3B illustrates that to increase frequency resolution by elongating the frequency axis to encompass more cycles, it is necessary to reduce the length of the orientation axis, thereby decreasing orientation resolution. Figure 1.3C illustrates that to increase spatial orientation resolution by elongating the orientation axis, it is necessary to reduce the length of the frequency axis.

The solution to the resolution problem is to construct a perceptual system with multiple levels so that there is a distribution of resolution trade-offs at each level and so that there is also a trade-off of resolutions between lower and higher levels. This solution returns us to the second theme: Perceiving is the interplay of several levels at once. For the visual system, we can imagine an initial level composed of receptors with small receptive fields, some optimized for frequency resolution and some optimized for orientation resolution. Each receptor is sensitive to changes in a tiny part of the visual scene. The problem is to convert this local information into coherent global percepts. We can further imagine that this first level feeds into a second level that integrates sets of spatially adjacent receptors so that the receptive field is larger but the resolution is less. The second level feeds into a third level that integrates sets of adjacent second-level receptors and so on. By combining all lower and higher levels in parallel, the perceptual system gets the best of two worlds: spatial detail from the initial level embedded in the global shapes from the higher levels. For the auditory system, the initial level would respond to individual frequencies; the next level would integrate the firings

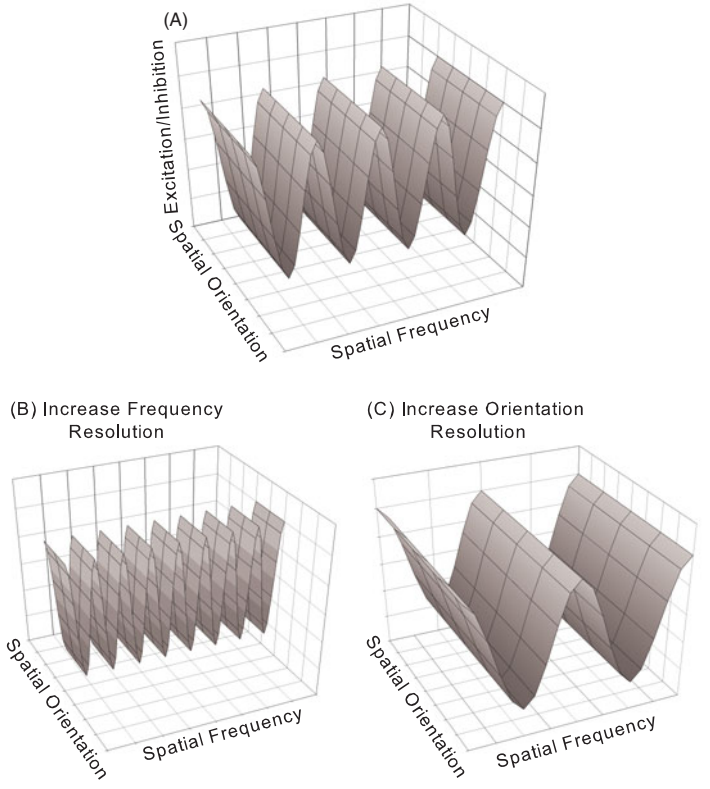

Figure 1.3. The joint uncertainty arises from spatial frequency and spatial orientation resolution, as discussed in chapter 2. Panel (B) illustrates that increasing frequency resolution by elongating the frequency axis necessarily reduces the orientation resolution. Conversely, panel (C) illustrates that increasing spatial orientation resolution by elongating the orientation axis necessarily reduces the spatial frequency resolution.

of adjacent frequencies. Still higher levels would integrate lower levels to create tone quality (i.e., timbre and pitch), and temporal organizations such as rhythm that extend over longer time spans.

Aperture, Correspondence, and Inherent Uncertainty

The aperture, correspondence, and inherent uncertainty issues are all interrelated. The inherent trade-offs in resolution force us to create "tight" apertures in space and time to capture the rapidly changing light and sound energy that signify the boundaries of objects and events. The necessity for apertures to maintain the fine-grain information in turn creates the correspondence

problem. The "snapshots" in space and time must be fused to create a useful perceptual world.

What will emerge in the following chapters is that the correspondence problem is solved in two ways. The first may be termed *effortless and passive*. Here the correspondences are found without conscious effort, before higher-order processes involved with shape analysis or figure-ground segmentation occur. The second may be termed *effortful and attentive*. In this second case, the correspondences are found by actively searching the stimuli to seek out the matches. As a first approximation, the first type of correspondence occurs in the short range, across small displacements in space or time, while the second type occurs in the long range, across large displacements. Perhaps the best strategy is to choose the process that minimizes the correspondence uncertainty.

Perceiving the World Occurs in Real Time

Given the immediacy and transparency of perceiving, it is easy to forget that perceiving is based on the patterning of neural spikes (Rieke et al., 1997). The spike train is not a static image; it is a running commentary or simultaneous translation of the objects and contrasts in the world. Here is yet another trade-off hinging on temporal integration: between combining the spike trains in time to average out inherent errors or maintaining correspondence with the ongoing changes. As I will argue throughout the book, the solution entails a continuum of neural mechanisms that cover the range from short temporal periods necessary for responding to rapid changes to long temporal periods necessary for averaging responses.

Rieke et al. (1997) persuasively argued that the neural spike code must be understood in the context of the natural timing of external events and in the context of what alternative events could occur. In many natural environments, stimulus variation may occur within intervals of 100 ms (e.g., speech sounds) so that given typical neuron firing rates from 10/s to 50/s, the stimulus change may be signaled by as few as one to five spikes. Thus, there may be sparse coding in the temporal domain in which there is but one spike for each change in the environment. (I return to the issue of sparse coding in chapters 2 and 3.) The interpretation of such a neural code cannot be made without some a priori knowledge of the possible stimulus changes, and our interpretation of the information and redundancy of the signals cannot be done without defining such alternatives. The auditory and visual worlds are not random, and there should be strong internal correlations in the neural spike train that match the internal structure of objects and events.

Rieke et al. (1997) went on to point out that the classic dichotomy between neural coding based on spike rate and that based on the timing

between spikes (e.g., phase-locking to specific parts of the signal) should be understood in terms of the rate of change of the stimulus. If the stimulus is not varying (e.g., a static visual image), rate coding provides the usable information, and the timing information is nonexistent. If the stimulus is constantly changing, then the timing between spikes provides the useful information and the average firing rate may be unimportant. But if the stimulus is changing very rapidly, then the neural system may not be able to fire rapidly enough to synchronize to each change, and then only rate coding would be possible. In sum, the interpretation and usability of the neural code can be investigated only in terms of the intentionality of the perceiver, be it a fly, bat, or human, in a probabilistic environment.

Perceptions Evolve Over Time

Previously I argued that perception is the construction of the distal world from the proximal stimulation. What we often find is that perception of an event evolves over time. Initially, the percept is based purely on the proximal stimulus, but over time that percept is superceded by one that takes into account the overall context, previous stimuli, prior knowledge, and so on that result in a more accurate rendition of the distal world.

One example of this occurs if two lines with slightly different orientations are viewed through an aperture. Suppose the two lines are moving perpendicularly at very different velocities that are represented by the lengths of the two vectors. There are two possible perceptions here. The first, shown in figure 1.4A, which I term the *proximal motion*, is simply the vector sum of the two line vectors and therefore is an upward motion that is between the two individual motions, a sum. The second, shown in figure 1.4B, which I term the *distal motion*, is in the direction of the intersection of the two lines of constraint. That motion is up to the right, outside the individual motions. Observers report that the initial perception is that of the vector sum (less than 90 ms of presentation), but that percept soon gives way to motion toward the intersection of constraints (Yo & Wilson, 1992). The vector sum motion will still bias the perceived motion, pulling it toward the sum direction and away from the constraints' motion. I discuss other examples of this in chapter 9.

Pack and Born (2001) have shown that the response of individual cells of alert monkeys in the middle temporal visual area (MT or V5) of the visual pathway, which has been shown to integrate directional motion from lower levels, mirrors this perceptual transition. Early in the visual pathways, direction-sensitive neurons have only small receptive fields, so that they can respond to but a small region of a moving object. Thus they are likely to "send up" the visual pathways incorrect or conflicting information about motion. The stimuli used by Pack and Born were short parallel line

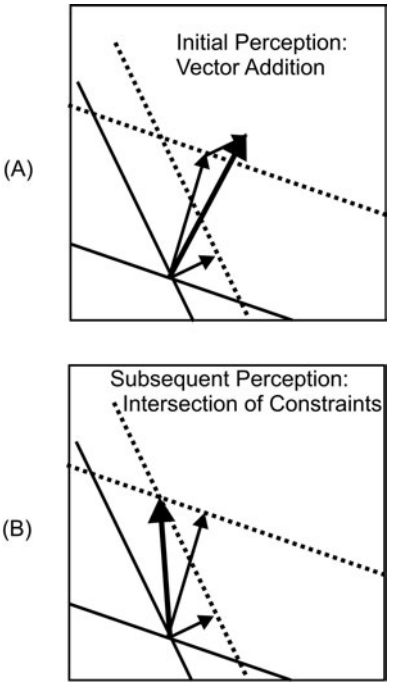

Figure 1.4. If two lines (or two gratings) move at different speeds, the initial percept is that both lines move together in a direction between the perpendicular movements of the two lines, as pictured in (A). Within a short time, the percept switches and now the two lines appear to move toward the intersection of constraints. As shown in (B), paradoxically that motion can be outside the angle formed by the two individual movements.

segments at different orientations. The lines moved either strictly perpendicular or 45° off from perpendicular. They found that the initial direction-specific responses (70 ms after movement onset) were affected by the original orientation of the bar, with the majority of responses perpendicular to orientation. But over time, the effect of the orientation decreased and the MT cells began to encode only the actual stimulus direction. Thus, by integrating responses that individually are spatially limited, the MT region can derive a relatively accurate picture of motion.

Perceptual Variables Are Those of Contrast and Change

The fundamental problem in beginning to understand perceiving is to isolate the important physical variables that create our perceptual world and to discover how to measure those variables to create simple relationships. We need to know which properties affect our construction of objects and events in the world, and which properties provide background and context. We could precede either empirically by manipulating the levels of the properties to determine their effects on perception, or we could proceed rationally by considering how such properties could affect perception in natural conditions.

From the latter perspective, we are asking a joint question about the ecology of the environment, the organism's goals in that environment, and the properties of the sensory systems. Consider overall intensity and the resulting perception of brightness and loudness. For both hearing and seeing, the range of intensities from the lowest (e.g., dim evening, whispers) to highest (e.g., sunny noon, rock music concerts) values can exceed the ratio of 1,000,000:1. However, individual neurons can only signal changes of intensity across the much smaller range of 100:1 or 1,000:1 because the firing rate saturates and cannot increase beyond that range. Yet we need to optimize our sensitivities at all light levels.

Clearly we need sensory energy in order to perceive at all, and overall intensity can provide information about such things as the size and distance of objects. But opaque solid or vibrating objects are characterized by their contrast to the overall level of energy. What is important for seeing is the ability to take the neutral mosaic of different light intensities reaching the retina and assign the bits to opaque objects interspersed and overlapped in space and time. What is important for hearing is taking the neutral pattern of air pressure variation and assigning parts to one or more vibrating objects interleaved in time and space. What characterizes all such objects is that the variation in intensity (i.e., their contrast) at the boundaries occurs more rapidly across time and space than variation in the background environment. Thus, we should expect the auditory and visual neural systems to maintain the correct response to contrast variation and to sacrifice an accurate response to overall illumination and loudness. In fact, the majority of cortical visual cells do not respond to blank scenes of any illumination. Moreover, the firing rates of many neurons in the auditory and visual pathways have a sort of gain control. As the background intensity increases, the average firing rate remains constant (instead of increasing) so that the neuron still can increase its firing rate to increases in intensity above the background. Without such a gain control, the firing rate would saturate at even modest background intensity levels.

There are many ways to demonstrate that contrast determines our perceptual world. Imagine a scene in which a black piece of paper is situated in a region of bright sunlight while a white piece of paper is situated in a region of dim light created by shadows. The black piece of paper would reflect more light energy overall. However, the black paper is seen as black while the white paper is seen as white. Thus, the amount of light energy per se does not determine brightness. The brightness is based on the ratio of the reflectance from the paper to the reflectance from the background. The visual system partials out the overall level of illumination (possibly to avoid saturation). The ratios are calculated in terms of the light in the local areas surrounding each piece of paper, and not in terms of the overall light across the entire scene. For hearing, we can construct a tone that oscillates

in amplitude across time (analogous to a visual stimulus that oscillates in brightness across space). The threshold for detecting the amplitude changes is nearly identical across a 100-fold change in overall intensity. In similar fashion, we can construct a complex tone by summing together a set of frequency components such that each component has a different amplitude. The threshold for detecting a change in amplitude of just one of the components also is relatively constant across a wide range of overall amplitudes (D. M. Green, 1988). Thus, the important auditory properties are those that signify changes in the relative vibration patterning that characterizes objects.

If we proceed empirically, then we would look for dependent variables that change smoothly, optimally in linear fashion, to changes in independent variables. Given the ecological properties described above, we should not expect a linear function. In fact, simple relationships are not found; the functional relationships change smoothly but not in linear fashion. At lower intensities, it appears that all the energy is integrated to detect the object at the cost of perceiving fine details. But at higher intensities, inhibitory processes emerge that limit the neural response in order to achieve a sharper auditory or visual image. Thus, auditory and visual adaptation at higher intensities maximizes object detection based on contrast. This makes intensity a nonlinear property that is not scalable. The functional relationships that exist for small changes in intensity at lower magnitudes are not the same ones that exist for the identical changes at higher magnitudes.

Perception Is the Balance of Structure and Noise

Above I have argued that the perceptual variables are those of change. But obviously that is just one part of the answer. The change must be predictable and that predictability must be able to be derived by the observer. Barlow (1990) put it differently: Perception converts possibly hidden statistical regularities into explicit recognizable forms to prepare for the figure-ground segregation necessary for learning. Perception is the construction of a representation that enables us to make reliable inferences about associations among sensations in the world around us.

At one end, there is noise in which there is no predictability among elements. For auditory noise, the pressure amplitudes are not predictable from one instant to another. For visual noise, the brightness of elements (e.g., points of different grayness levels) is not predictable from one spatial location to another. At the other end are periodic auditory and visual events in which there is perfect predictability between elements separated by a specific time interval or spatial distance. We might say that the combination of the predictable and nonpredictable parts is the "stuff," and that the

abstraction of the predictable parts yields the "things" of perceiving. (The stuff is really defined in terms of the things that result from the stuff.)

When there is a mixture of unpredictable and predictable elements, there is a normally irresistible perceptual segregation that isolates the predictable parts. If we mix a predictable tonal component together with a nonpredictable noise component, the perception is that of listening through the noise to hear the tone. Similarly, if we look at an object through a snowstorm, the perception is that of looking at an object whose parts are being covered and uncovered. I find it impossible to put the noise back into the tone to hear an integrated sound or to put the snowflakes back onto the object. I believe that the auditory and visual segregation is obligatory and represents the first step in achieving the objects of perceiving. Bregman (1990) has termed this process *primitive segregation*.

As argued above for single properties, it is the contrast or ratio between the amount of structure and amount of noise that is the important perceptual variable. For auditory stimuli that can be conceptualized as lying on the continuum between tone and noise, listeners reliably can judge the perceptual tone:noise ratio even when the overall levels of the stimuli are varied randomly (Patterson, Handel, Yost, & Datta, 1996). Perceiving is contextual and relativistic.

Rules of Perceiving Should Be True for All Senses

All of the above implies that listening, seeing, grasping, smelling, and tasting are fundamentally the same. Although the sensory inputs and sensory receptors are quite different in structure and operation, and the actual contrasts may be different, all function by partitioning and contrasting structure and noise. All senses have been optimized through evolution to provide animals with information about survival: predators, conspecifics, and food and water. But all senses must simultaneously be general-purpose systems that can respond to an ever-changing environment.

Often it is difficult to find the best way to illustrate correspondences between the senses. It is possible to attempt to match the basic dimensions of auditory and visual experience and then compare their psychophysical properties. I have implicitly compared loudness to brightness above, and pointed out that the range of perceptible physical energy is relatively equivalent. At this level, the comparisons would tend to focus on the parity of discrimination (e.g., ranges of discriminability, difference thresholds and Weber ratios, time and space integration windows). It is also possible to match the gestalt (for lack of a better word) properties of auditory experience (such as timbre, pitch, noise, roughness, texture, vibrato, location, motion, consonance, repetition, melody, and rhythm) to the gestalt properties of

visual experience (such as shape, motion, color, brightness, texture, symmetry, transparency, opacity, and location in three-dimensional space). For example, is the perception of temporal auditory noise equivalent to the perception of spatial visual noise? Finally, it is possible to compare the segregation of auditory scenes into sound-producing objects to the partitioning of a visual scene into light-reflecting objects. Figure-ground visual organization assigns a contour line to one and only one object. Does figure-ground auditory organization similarly assign a frequency component to one and only one object? Is there a generalized time-space representation into which all sensory experience is intertwined?

At first, the differences between hearing and seeing seem huge. Is it possible to use the same conceptualizations for listening and looking, given the vast differences in their normal functioning? Light energy is electromagnetic. Light waves travel nearly instantaneously, so that interocular temporal differences cannot exist. The wavelengths are miniscule (400–700 nm), which allows excellent spatial resolution, while the frequency is very high, which disallows phase-locking of the neurons to individual cycles. Sound energy is mechanical pressure. Pressure waves travel slowly, so that interaural temporal differences can be used for localization. The wavelengths can be body size, which minimizes the ability to determine object size and shape, while the frequency is relatively low, so that neurons can phase-lock to individual cycles. The physiological differences reflect these differences. The visual system has 120 million spatial sensors per eye (every rod and cone in each eye can be thought to represent one spatial point), while the auditory system has but 2,000 inner hair cells per ear that cannot represent spatial direction. However, the 2,000 auditory inner hair cells have different frequency sensitivities, whereas the visual system has but three different cone sensitivities and just one rod sensitivity. These differences are summarized in table 1.1.

On this basis, Kubovy and Van Valkenburg (2001) claimed that audition and vision serve very different spatial functions: "listening to" serves to orient "looking at." Caelli (1981) suggested that it is impossible to meaningfully compare the different kinds of perception, and Julesz and Hirsh (1972) argued that analogies between vision and audition might, at best, not be very deep because visual perception has to do with spatial objects while auditory perception has to do with temporal events.

Nonetheless, I would argue that perceiving in all sensory domains is finding structure in the energy flux and that deriving equivalences among the domains can deepen our understanding of how we create the external world. For example, one kind of equivalence is that the cortical representation of all senses tends to be arranged into discrete processing areas. Nearly always, adjacent cells represent slightly different values of the same feature (e.g., acoustic frequencies or spatial orientations). In each of these cortical

Table 1.1 Comparison of Hearing and Seeing

Property	Hearing	Seeing
Type of Energy	Mechanical Pressure Waves	Electromagnetic Waves
Speed of transmission	a. Relatively slow—(340 m/s) b. Allows for interaural temporal differences to judge direction	a. Nearly instantaneous—(3×10^8 m/s) b. No interocular temporal differences
Wavelength	a. Relatively long—(.02–10 m) b. Poor spatial discrimination	a. Very short—(400–700 nm) b. Excellent spatial resolution (light shadows)
Frequency	a. Relatively slow—(30–20000 Hz) b. Allows phase-locking to individual cycles c. Excellent temporal resolution	a. Very high—($4.3–7.5 \times 10^{14}$ Hz) b. Phase-locking impossible c. Poorer temporal resolution
Physiological sensors	a. Mechanical process b. Rapid regeneration c. Rapid adaptation	a. Chemical process b. Slow regeneration c. Slow adaptation
Number of receptors	Relatively small number—(2,000 hair cells/ear)	Large number—(120,000,000/eye)
Cerebral cortical area	8%	20–30%
Sensitivity	Distributed across frequency range	Three types of cones plus one type of rod
Object properties	Tend to be intermittent	Tend to be stable
Additivity	Sound pressure waves are transparent and add together	Light waves reflect off opaque objects and usually block each other

zones, an environmental stimulus or movement becomes represented by an isomorphic pattern of firing in the cortex (DeCharms & Zador, 2000). There is no necessity for this type of organization and yet all systems have evolved to this arrangement.

To represent the auditory and visual worlds, I make use of the concept of autocorrelation in space for vision (co-occurrences of brightness or color patterns separated by a fixed distance) and autocorrelation in time for audition (co-occurrences of intensity patterns separated by a fixed interval). By thinking in terms of autocorrelation to find order, I shift the explanation for perception to the global space-time properties of the ongoing stimulus array (Uttal, 1975). It is in same tradition as the efforts of J. J. Gibson to describe what there is to perceive in the world.

To represent the correspondences between the physical world, neurological codes, and perceptual experience, I will again use the correlation. Here,

we would expect the correlation to be between stimulus contrasts and neurological contrasts (differences in rate or timing of the spikes). Both experimental data and mathematical simulations (Panzei & Schultz, 2001) indicate that the nature of the correlation depends on the timing of the stimulus contrasts, the presumed time in which the nervous system integrates the firings, and the variability in the noise of the neurons (this is the same argument made by Rieke et al., 1997, described previously). The correlation should not make use of a simple physical description of the stimulus. The nervous system does not create a perfect recording or photograph of the stimulus, and may exaggerate or disregard certain physical correlations and properties. Moreover, the perceptual representation is malleable as the person shifts attention. Julian Hochberg (1982, p. 214) argued, "the attributes that we perceive do not in general exist in some internal model of the object waiting to be retrieved. They are the results of our intention to perceive, and they appear in the context of the perceptual task that calls upon them." Thus, there may be no single kind of correlation that always is used, but we might expect that the auditory and visual systems will use the same neural contrasts when faced with equivalent stimulus contrasts (DeCharms & Zador, 2000).

Summary

The many interrelated concepts discussed in this chapter shape the intent of this book. Namely, I search for correspondences in the construction of the external world achieved by abstracting the structure of auditory and visual sensations across space and time. This is not to argue that there is consensus as to how sensory systems create a percept. There is not such a consensus and I would suspect that this lack is due to the diverse ways in which a percept could be constructed. Formulating the correspondences is slippery, and the bases for the correspondence can change from instance to instance. Nonetheless, the consistent goal is to compare the textures of the auditory and visual phenomenal worlds.

2

Transformation of Sensory Information Into Perceptual Information

If we take the reasonable position that perceptual systems evolved to perceive the spatial and temporal properties of objects in the world, then the place to begin is with an analysis of the characteristics of that physical world.[1] For some species, the perceptual world may consist of specific objects necessary for survival, and therefore we might look for physiological mechanisms that uniquely detect those objects (e.g., specific cells in the frog's tectum, colloquially termed *bug detectors* by Lettvin, Maturana, McCulloch, and Pitts (1959) that fire to small dark convex objects moving relative to the background). For other species including humans, the perceptual world is ever expanding in terms of novelty and complexity and therefore we might look for physiological mechanisms that detect statistical regularities and relationships, rather than specific things. This suggestion is analogous to Shepard's (1981) theory of psychophysical complementarity that physiological mechanisms and perceptual heuristics evolved in response to physical regularities. It may be possible to predict the characteristics of peripheral and central processes by figuring out how such regularities could be coded optimally.

We should ask a variety of questions:

1. Are there physical regularities in the scenes we normally encounter (excluding man-made objects that produce sounds at particular frequencies or that are made up of vertical and horizontal straight lines meeting at right angles)?

1. It is possible to take a different theoretical stance and argue that the function of sensory systems is to enable appropriate behavior with or without a conscious percept.

2. Are the sensitivities and functioning of the physiological mechanisms and perceptual systems optimally constructed to encode physical regularities in the world? Do these systems make use of the prior probabilities of objects and events?
3. Do the perceptual organizations mirror the physical properties of the world in terms of the physical actions necessary to survive (breaking through the camouflage of predators and prey)?

There are many reasons for an optimal code:

1. An optimal code will compensate for the rather limited range of firing rates for individual cells in the retina and inner ear in the face of much wider variation of physical properties in the world.
2. In the vertebrate visual system, the number of optic nerve fibers creates a bottleneck for the transmission of retinal signals to the brain. The human eye contains about 5 times more cones, and 100 times more rods, than optic nerve fibers (Thibos, 2000). For each eye, there are approximately 100 million receptor cells in the retina but only 1 million fibers in the optic nerves so that the retinal signal must be compressed to achieve the necessary transmission rate (the number of cells does increase again to more than 500 million cells in the cortex). The purely spatial retinal information of the rods and cones is transformed into a localized receptor-based analysis based on frequency and orientation that can sacrifice the part of the retinal information that is redundant and that does not help capture the object causing the sensations.
3. An optimal code at the receptor level will minimize the propagation and amplification of intrinsic error as the signal progresses through the nervous system.
4. An optimal code will match the output of the perceptual mechanism to the distribution of the independent energy in the external world. An important fact about natural time-varying auditory and visual scenes is that they do not change randomly across time or space. Due to the physical properties of objects, the brightness and color of any single visual object and the frequency and loudness of any single sound object change very gradually across space and time. Nonpredictable, sharp, and abrupt changes signify different visual and different sound-producing objects (Dong & Atick, 1995). Therefore, removing the predictable parts or making them explicit (Barlow, 2001) can lead to a concise and nonpredictable description.

We need to be cautious about embracing any optimality argument because it is impossible to state definitively just what should be optimized. As stated in chapter 1, perceptual systems need to be optimized in two conflicting ways: (1) for those relatively static properties involved in specific

tasks and contexts (e.g., identification of mating calls and displays); and (2) for those emergent properties that identify auditory and visual objects in changing situations. A fixed set of feature detectors would be best for the former but unable to encode novel properties, while a dynamic nervous system that can pick up correlated neural responses would be best for the latter but unable to rapidly encode fixed properties. As described in this chapter, the auditory and visual systems are organized into tracts that are selective to particular stimulus dimensions, but there is an immense amount of interconnection among the tracts. What you hear or see has been modified by those interactions among the neural tracts.

In what follows, I consider two interrelated issues. The first issue is the neurological transformations that convert the sensory excitations that result only in increases in firing rate at the receptors into excitatory or inhibitory codes that represent objects in the world. Every neuron in the auditory and visual pathways is maximally sensitive (selective) to combinations of stimulus dimensions. For example, an auditory neuron might respond to particular combinations of frequency and amplitude, while a visual neuron might respond to particular combinations of frequency and spatial position. In general, farther up the pathways, the neurons become more diverse and selective and respond only to particular combinations of stimulus dimensions. It does not seem to be that perception occurs only at the end of the auditory or visual pathways; rather, the brain selects and alters the neural firings throughout the pathways.

The second issue is the match between the above transformations and the structured energy in the auditory and visual worlds. This entire book is predicated on the assumption that there is a close match between the two. It is more logical to proceed from stimulus energy to neurological transformation to reflect the role of evolution. However, I have found it easier to work in the reverse direction, first understanding the neural transformations and then matching those transformations to the properties of stimulus energy.

Neurological Transformations: The Concept of Receptive Fields

The receptive field of a neuron is the physical energy that affects the activity of that neuron. The receptive fields of nearly all cells past the receptor level contain both excitatory and inhibitory regions. The receptive field concept was first used in vision by Hartline (1940) to describe the ganglion retinal cells in the frog's retina, but it is so general that it has been used for all modalities and at all levels of the nervous system. Once the receptive field is known, it becomes a description of the transformation of some property of the sensory energy into a sequence of neural firings. Colloquially,

we think of that property as being a feature of the visual and auditory stimulus and imagine that the identification of an object is based on the collection of such features. But we should not be trapped by that metaphor; the neurons really are filters, not feature detectors.

In vision, the receptive field is defined as the retinal area in which an increase or decrease in illumination changes the firing rate of the ganglion neuron (or cortical neuron) above or below the average rate of firing found in the absence of stimulation (Kuffler, 1953). The receptive field of the ganglion or cortical cell will be determined by the sensory receptors to which it is connected. To determine the retinal location and the spatial and temporal properties of the receptive field, flashing small lights, moving bars, or more complex configurations are presented at different retinal locations to identify the retinal positions and the light/dark patterns that maximally excite and inhibit the cell. In audition, the receptive field is defined as the frequencies, intensities, and durations of the acoustical wave that increase or decrease the firing rate of the neuron (identical to that for vision) and it is identified in the same way as in vision. Receptive fields imply specialization in firing. For vision, the receptive field of a neuron is localized at a particular retinal location and differentiated in terms of the spatial and temporal pattern of the light energy that fires that cell. For audition, the receptive field is localized at a position on the basilar membrane and is differentiated in terms of the temporal pattern of the acoustic energy that fires the cell.

Intuitively, the way to identify the receptive field is to present a wide array of visual and auditory stimuli and pick out those stimuli that increase the firing rate of the cell and those stimuli that decrease the firing rate. If you are smart (and lucky), then it will be possible to construct such a set. However, given the innumerable configurations in space, white-and-black contrast, frequency, intensity, and frequency and intensity oscillations that might uniquely trigger an auditory or visual cell, a more formalized procedure often is necessary.

The procedure that has evolved has been termed *reverse* or *inverse correlation*. In essence, the experimenter presents a sequence of randomly varying stimuli and then averages the stimulus energy that precedes a neural spike. Imagine a very short duration, very small pinpoint of light that is either brighter or darker than the surround. Furthermore, imagine that any response immediately following the presentation of the pinpoint simply increases the firing rate by one spike. Next, the experimenter presents the lighter and darker light many times at each spatial position and counts the number of spikes for each light (clearly the responses will not be identical at a single point to either light due to chance factors in the nervous system or in the light emitted). After measuring the probability of firing to each light at every position, the experimenter can identify excitatory regions where an increase in intensity generates a spike, inhibitory regions where a

decrease in intensity generates a spike, and neutral regions where neither an increase nor decrease in intensity change generates a spike. In effect, he or she is correlating the input (light intensity) to the output (spike probability). The responses of the neuron define its own receptive field.

Now consider a more complex case in which the relevant stimuli are unknown. We might try using natural stimuli. However, it can be difficult to describe the characteristics of a neuron using natural stimuli because natural stimuli have internal correlations of energy, so that it may be impossible to link the spikes to a specific feature of the stimulus. For this reason, white noise has often been used as the stimulus to identify the receptive field. White noise can be simply understood as a pattern or sequence of light or sound stimuli such that the amplitudes vary randomly so that no correlation or prediction is possible between any two amplitudes separated in space or time.

We present the random white noise continuously. The intensity of the stimulus prior to each spike is measured and cumulated in say 100 sequential 1 ms time bins. Then, the intensities in each bin are averaged separately. The stimulus feature (intensity pattern) that triggers the spike will occur consistently in the time bins prior to the spike and therefore create high average amplitudes (or high probabilities), while the nonrelevant features will vary randomly (being essentially error) and average toward zero. This outcome is termed the *spike-triggered average stimulus* (Dayan & Abbott, 2001). The spike-triggered average stimulus is mathematically equivalent to calculating the correlation between the stimulus amplitude at each prior time point and the probability of a spike. It also has been termed the *fast Weiner kernel*, or the *reverse correlation function*. It is the receptive field of the cell.

In table 2.1, I generated a series of 60 random numbers (0–9 with an average of 4.5) and indicated the 18 spikes by the symbol *. I then averaged the intensities in the five time periods preceding the spike and plotted the averages in figure 2.1.

We could classify the receptive field of this hypothetical cell as an "on" cell that fires when the intensity at −20 ms and 0 ms is high. (I constructed the sequence so that spikes occurred if the sum of two successive intensities was 12 or greater.)

A more complex case occurs when the stimuli consist of multiple frequencies and the problem is to induce the receptive field, which may consist of several excitatory and inhibitory regions. I constructed a simplified example in table 2.2 in which four frequencies were presented (16 possibilities). As above, there were 60 presentations, spikes are indicated by *, and the probability that each frequency occurred in the four time bins preceding the spike is shown in table 2.3. The probabilities for F_1 and F_4 are close to the expected value; the probabilities for F_3 are above the expected value (excitation) particularly for −20 ms; and the probabilities for F_2 are below the expected value (inhibition), particularly for −20 ms.

Table 2.1 Derivation of the Receptive Field

(A) Stimulus Sequence and Resulting Spikes

Stimulus	4	6	7	0	3	2	1	2	4	0	5	1	5	9	5	4	1	6	6	8	4	5	9	6	3	3	8	3	7	7	9
Spike		*												*	*				*	*	*		*	*						*	*

Stimulus	0	5	1	5	4	0	4	3	3	4	4	4	8	9	2	0	6	9	3	1	9	7	0	5	5	9	0	2	3	5	9
Spike													*	*				*	*			*				*					*

(B) Derivation of Reception Field (Assume Stimuli Are Presented at 20 ms Intervals)

	Time Before Spike				
Spikes	80	60	40	20	Spike
1			4	6	7
2	0	5	1	5	9
3	5	1	5	9	5
4	5	4	1	6	6
5	4	1	6	6	8
6	1	6	6	8	4
7	6	8	4	5	9
8	8	4	5	9	6
9	3	8	3	7	7
10	8	3	7	7	9
11	3	4	4	4	8
12	4	4	4	8	9
13	9	2	0	6	9
14	2	0	6	9	3
15	9	3	1	9	7
17	7	0	5	5	9
18	0	2	3	5	9
Mean	4.6	3.4	3.8	6.7	7.2
SD	3.0	2.5	2.1	1.7	1.9

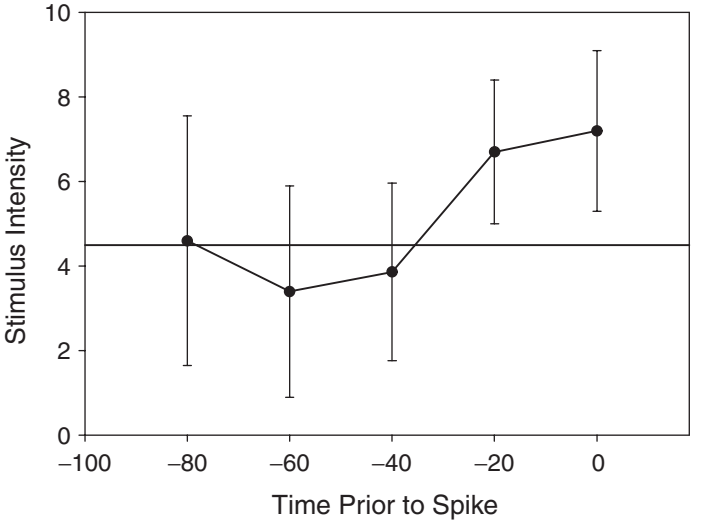

Figure 2.1. Average and standard deviation of the stimulus intensity before a spike (derived from table 2.1).

We can represent this cell from two perspectives. The response is depicted in figure 2.2A, measured from the tone onset at time 0. It portrays the receptive field as a filter. This simplified representation illustrates that 20 ms after the presentation of F_3 the firing rate decreases (shown in black), that 20 ms after the presentation of F_2 the firing rate increases (shown in white), and that the presentation of other frequencies does not change the baseline rate. If both F_2 and F_3 were presented, the resulting firing rate would be the difference between the two effects. The response is depicted in figure 2.2B, measured backward from the spike at time 0, as for reverse correlation. The frequency response of the cell can be found by drawing a vertical line through the region of maximum excitation (shown to the right). The temporal response can be determined by drawing a horizontal line through the region of maximum excitation, shown below the receptive field. This cell will fire with the highest probability 20–40 ms following the F_2 stimulus. It "detects" F_2.

It is useful to conceptualize the receptive field as a linear filter. As the auditory or visual stimulus energy evolves over time, the receptive field allows certain energy configurations through. An auditory receptive field could fire only when a specific range of frequencies occurs (a band-pass filter), or it could respond only to an upward (or downward) frequency glide within a set time period. We can test how well we have characterized the receptive field by simulating the receptive field mathematically, presenting

Table 2.2 Spikes Resulting From the Presentation of Tones Composed of One to Four Frequency Components

	Time														
Frequency	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
F_1		X	X				X	X	X	X					
F_2	X	X	X		X	X			X				X		
F_3	X		X	X	X	X				X	X	X	X		X
F_4			X		X		X				X			X	X
Spike					*						*	*	*		
	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
F_1	X		X	X	X			X	X			X	X	X	
F_2			X		X		X			X	X	X	X		
F_3	X	X	X	X	X							X	X		
F_4		X	X	X	X			X		X			X		
Spike	*	*	*		*										
	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
F_1		X		X	X						X			X	X
F_2		X			X	X	X		X		X				
F_3	X			X			X	X	X	X	X	X		X	
F_4	X		X		X	X		X	X				X	X	
Spike		*			*				*		*		*		*
	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
F_1	X		X	X			X		X	X		X	X		X
F_2						X		X	X	X			X		X
F_3	X	X	X	X		X		X				X			
F_4		X	X		X		X	X		X					X
Spike		*	*	*	*								*		

a realistic stimulus input, and then calculating the output of the simulated receptive field. We then correlate the simulated response to that of the actual neural receptive field using the identical input.

Suppose we manipulate the receptive field, moving the inhibitory region relative to the excitatory region, as shown in figure 2.3 by 20 ms. Assume that only the F_2 and F_3 frequencies are presented, each at 100 units. In the gray region, the probability of response is .25 (resting rate); in the black inhibitory region the probability is 0.1; and in the white excitatory region the probability is .9. Now imagine that we are measuring the output of the cell starting at the onset of the tones. The response rates are shown in table 2.4.

At the tones onset, the cell fires at its base rate to any frequency. Then from 10 ms to 20 ms, F_2 hits the excitation region before F_3 hits the

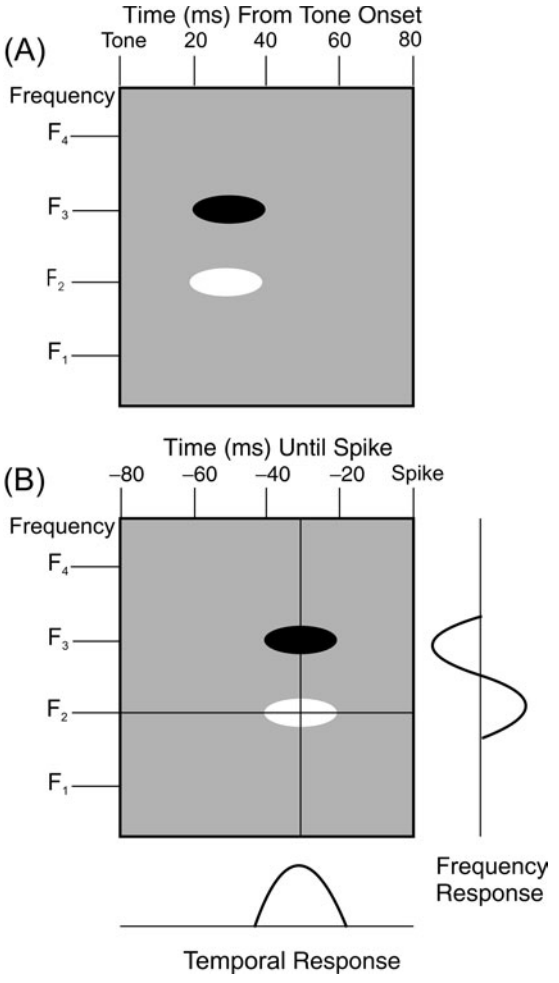

Figure 2.2. The receptive field for a cell. The gray area represents the baseline firing rate; the white area represents the excitation region; and the black area is the inhibition region. In (A), the response is portrayed in terms of the stimulus onset at time 0; in (B), the receptive field is portrayed in terms of the spike. Here, the excitation and inhibition areas can be thought of as features that trigger (or inhibit) a spike.

Table 2.3 Probability of Firing Based on Table 2.2

	Time Before Spike (ms)				
Frequency	80	60	40	20	0 (Spike)
F_1	.61	.56	.50	.50	.50
F_2	.50	.28	.33	**0**	.44
F_3	.56	.44	.61	**1.00**	**.67**
F_4	.44	.44	.50	.44	.61

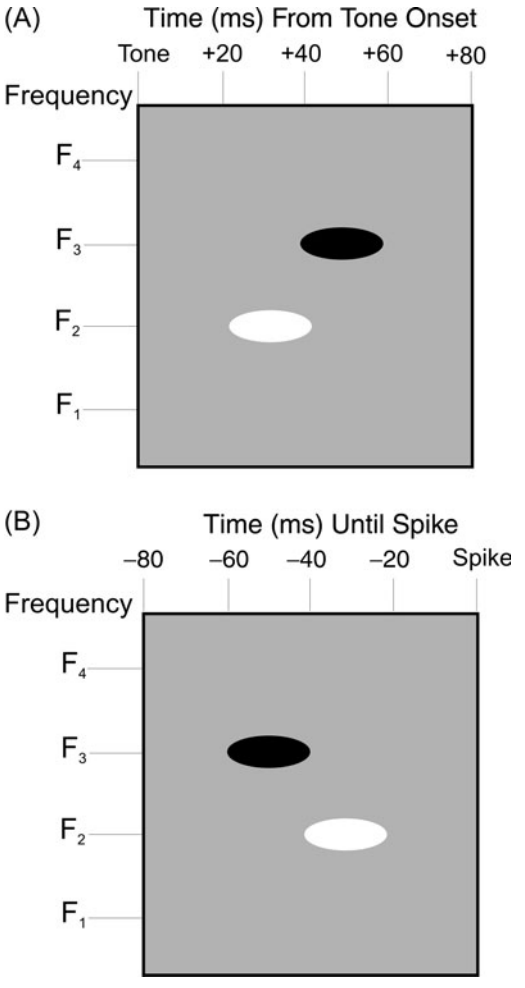

Figure 2.3. The receptive field for the cell in figure 2.2 in which the inhibitory region is offset by 20 ms. The maximum increase in firing rate occurs 20 ms after the tone onset, while the maximum decrease in firing rate occurs 40 ms after tone onset.

inhibition area and the rate increases. As the tones reach the F_3 inhibitory region, the firing rate decreases, particularly beyond 40 ms. Finally the firing rate returns to the resting level.

A more complicated case is shown in figure 2.4 for a cell that is most likely to fire for frequencies around 2000 Hz, but the principle is exactly the same.

This procedure does not completely solve the problem of generating the receptive field for three reasons. First, the choice of the stimuli still limits what you can find out. For example, experiments that use white

Table 2.4 Firing Rate From Time of Onset of Tones

	Time From Onset of Tones (ms)							
Frequency	Onset	+10	+20	+30	+40	+50	+60	+70
F_3	.25 × 100	.25 × 100	.25 × 100	.25 × 100	.10 × 100	.10 × 100	.10 × 100	.25 × 100
F_2	.25 × 100	.25 × 100	.90 × 100	.90 × 100	.90 × 100	.25 × 100	.25 × 100	.25 × 100
Sum	50	50	115	115	100	35	35	50

noise should theoretically be able to induce the features that make up the receptive field of any cell. But such random noise stimuli have not worked out well for neurons in the auditory cortex that are not sensitive to or even inhibited by broadband white noise. Thus, the initial choice of stimuli will affect the ability to identify the receptive field. Second, because the reverse correlation procedure averages the stimuli that create a spike, it would be difficult to distinguish between a neuron that fires only when two different frequencies are simultaneously present and a neuron that fires simply to either of the two frequencies (unless combination stimuli are presented). Third, the majority of real stimuli have internal correlations, so that it is necessary to partial out those correlations to derive the receptive field.

Receptive Fields in Vision

At the Retinal Ganglion Cells and Optic Nerve

The visual system transforms the retinal mosaic into a set of pathways that encode different properties of the visual stimulus. Much of this transformation occurs in the eye itself. The excitation from each retinal point diverges and connects to a set of ganglion cells such that each cell is selective for one property. (I am using the term *property* simply to mean a particular spatial configuration of brightness.) Every retinal point becomes represented by a set of equivalent ganglion cells. Thus, combining the analogous ganglion cells across the retinal points creates a retinal map of that property, and the convergence of all the ganglion cells in the optic nerve creates a parallel set of retinal property maps. The single-excitation map is transformed into multiple-property maps.

Briefly, the eye can be conceptualized as being composed of three layers. Light entering the retina first passes through the ganglion cells, then through the inner and outer plexiform layers that contain the amacrine cells, the bipolar cells, and the horizontal cells, and finally reaches the rod and cone receptors. The light energy always causes an increase in firing