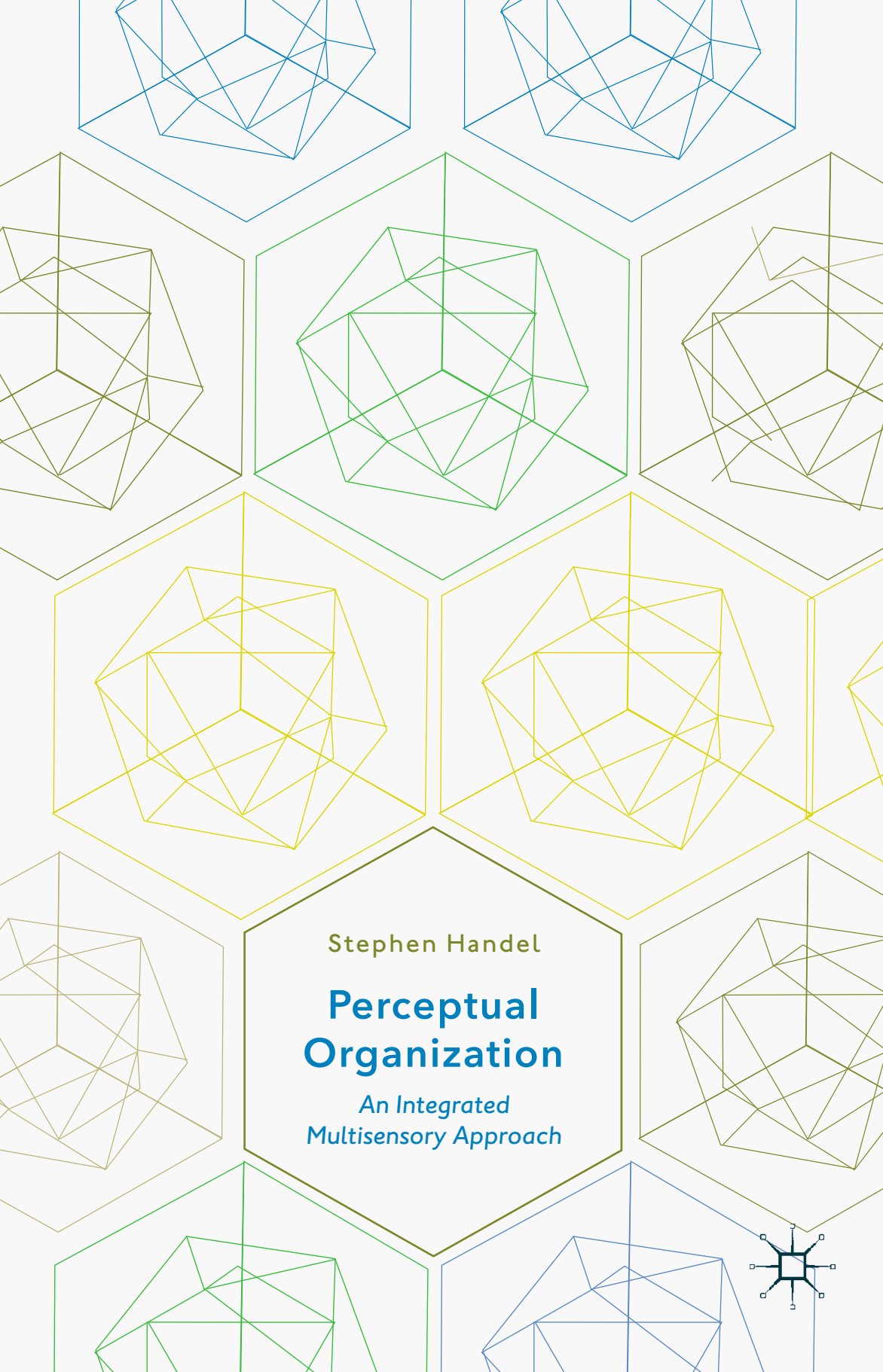


Stephen Handel

Perceptual Organization

An Integrated Multisensory Approach

Perceptual Organization

Stephen Handel

Perceptual Organization

An Integrated Multisensory Approach

palgrave
macmillan

Stephen Handel
Psychology
University of Tennessee, Knoxville
Knoxville, TN, USA

ISBN 978-3-319-96336-5 ISBN 978-3-319-96337-2 (eBook)
https://doi.org/10.1007/978-3-319-96337-2

Library of Congress Control Number: 2018959422

Cover design by Fatima Jamadar

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my Family and Friends

Preface

The goal here is to describe some of the processes underlying how the auditory, tactual and visual modalities and their combinations convert sensory inputs into two- and three-dimensional objects. Those processes are a mixed bag: some depend on cells tuned to various sensory properties, some depend on rhythmic firing in the nervous system, some depend on our knowledge about the environment and some depend on our expectancies about future events. Those processes are interlocked and interactive, and it amazing to me that those objects emerge so quickly and seamlessly. We can take a stab at understanding how these processes create our perceptual world, but a full understanding is in the future.

In writing this book, I have come to appreciate the contribution of Gunnar Johansson and how his notion of vector analysis can be applied to wide variety of perceptual problems. His initial research was based on movies of black-suited dancers with lights on their arm and leg joints. People were able to describe various actions of the dancers using only the trajectories of the lights placed on those joints. This outcome was surprising because the lights were unconnected, and people needed to track the relationships among individual lights to perceive among the dancers actions.

The movies are captivating, but for me, the importance of Johansson's research lies in the conceptualization that there are levels of temporal and spatial structure so that the interpretation of any level depends on interpretation of all the other levels. The motion of faster lights on the hands are not perceived as if they were in isolation, they are perceived in relation to the movement of the slower lights on the body. The vector analysis provides a way to understand how the motion of the faster lights is split into that part common to the slower lights plus that part which is unique.

I have found that the underlying idea of vector analysis, separating the common parts of a scene from the unique parts can be applied to a wide range of perceptual processes. Vector analysis can help understand figure-ground organization in static visual scenes, the formation of a meter in rhythm, embodied

movements in time with metric structures, and the accurate movements of hands and arm in directed actions. I feel strongly that future research will have to confront the relative nature of perception.

I have tried to keep the text non-technical and I hope it will be suitable for a wide audience. There is but one equation, and a smattering of physiology. A course in Sensation and Perception would be helpful to understand the book, but it is not necessary. While Chap. 2 is probably the most important, the other chapters are mostly self-contained so that it is possible to read any single chapter or several chapters in any order.

Every book is a joint effort. I thank Drs. Gough, Hjoetkjaer, McAdams, Toiviainen, and R.M. Warren who kindly allowed me to make use of parts of their research. I am greatly indebted to Dr. Robert Cantwell, Professor Emeritus of American Studied at the University of North Carolina who read and improved the text and who encouraged me to keep keeping on. The staff at the Jackson Laboratory, Bar Harbor helped in several ways. William Barter answered many questions about publishing ethics and pushed me to submit the manuscript to different publishers. Doug McBeth and Ann Jordan, once again, processed my reference and inter-library requests with unfailing good humor and grace.

Bar Harbor, ME, USA Stephen Handel

How to Use This Book

Accessing the Sound Files

All the sound files were produced using the freeware program Audacity (http://audacityteam.org/) using the mp3 format.

Print book readers will have access to the audio files, "electronic supplementary material" for your purposes, on Springer Link. There will be a link printed in the book so print readers know where to access the files. The supplementary material is a set of Sound Files that are linked to the figures and text. The relevant sound files are indicated at the appropriate places in the text. The details of the Sound Files are presented in the supplementary material. To access the files for a chapter, please go to the first page of the chapter and review the footnote "Electronic Supplementary Material."

I suggest using a software program to display the sound files visually so that you can follow the sounds by means of a cursor (remember this is a book about multisensory perception). There are many such programs: These include Audacity, PRAAT, Raven-lite, and Sonic Visualizer. All can be downloaded free from the Web.

The programs will have slightly different control panels but all have similar functionalities. I would try shifting back and forth from the amplitude x time representation to the frequency x time representation; each one provides a differing view of the sound sequences. The latter representation shows how the frequency of each harmonic changes over time. Another function of these programs is to be able repeat sound files continuously (looping). Looping may be necessary to perceive the multistability of some of the sound files.

Readers who purchase the print volume can ask customer service at customerservice@springer-nature.com to access the online book files or just go to the webpage and register to access the online book.

Numbering of the Figures and Sound Files

The numbering of a Figure and its associated sound file is always the same. Thus, Fig. 2.3 and Sound File 2.3 refer to the same stimuli. However, there are figures without an associated sound file and sound files without an associated figure. In those cases, I have kept the figure and sound file numbering in registration. For example, there are no sound files associated with Figs. 2.1 and 2.2. The next figure, Fig. 2.3 does have an associated sound file and that is labeled Sound File 2.3, not Sound File 2.1 even though it is the first sound file. Thus, even while it may seem that there are missing figures and missing sound files, the gaps are due to the numbering system.

YouTube Videos

At several places in the text, I have listed YouTube videos that illustrate different visual phenomena. It is helpful to view these, as they illustrate some of the issues discussed in the text.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Perceiving the world of real objects seems so easy that it is difficult to grasp just how complicated it is. Not only do we need to construct the objects quickly, the objects keep changing even though we think of them as having a consistent, independent existence (Feldman, 2003). Yet, we usually get it right, there are few failures. We can perceive a tree in a blinding snowstorm, a deer bounding across a tree line, dodge a snowball, catch a baseball, detect the crack of a branch breaking in a strong windstorm amidst the rustling of trees, predict the sounds of a dripping faucet, or track a street musician strolling down the road. In all cases, the sensations must be split into that part that gives information about real objects that may change in shape, sound, timing, or location and that part that gives information about the random or non-predictable parts of the background. The object becomes "in front of" the background.

The light energy at the eyes, the sound energy at the ears, and the pressure sensations on hands are neutral. Moreover, the light energy at each eye is two-dimensional due to the "flat screen" structure of the retina, the sound energy at each ear has only a weak spatial component, and the pressure sensations must be integrated to yield the surfaces of objects. For all three senses, the energy must be interpreted to give the properties of the three-dimensional objects and events in the world.

Perception is not passive. Looking at, listening to, sneaking a peek, eavesdropping, rubbing, fingering, shaking, and grasping are all actions that ultimately yield information about objects. But these acts just give us the energy at the receptor, they do not by themselves specify a particular object. Many different objects could produce those same sensations. It is the creative and intentional act of looking or listening that results in the construction of objects (Handel, 2006). Perceptions are not independent of the perceiver. The sensory receptors of all organisms limit what can be known about the real world. It might be too strong to argue that we can never perceive the true physical world, but those limitations act to mark off the consequential features.

S. Handel, *Perceptual Organization*,
https://doi.org/10.1007/978-3-319-96337-2_1

Perceiving is basically focusing on parts of the environment so that depending on their objectives, expectations, and knowledge, people often end up with different outcomes (Felin, Koenderink, & Krueger, 2017).

1.1 The Aperture and Correspondence Problem

Many constraints limit our ability to perceive objects, whether memory or cognitive limitations or environmental obstacles. Auditory, visual, and tactual sensations are constantly changing, and a visual glimpse, auditory snippet, or brief touch must be interpreted in terms of what preceded it in time and space and what will follow it. This has been termed *the aperture problem* to convey the idea that it is like looking through a slit or hole. While this term is couched in visual terms, it obviously also is true for auditory and tactual events that naturally evolve in time and space. The aperture problem is both the cause and a complement to the *correspondence problem*. The sensations at any one instance cannot unambiguously signify objects or events; consequently, the perceptual systems must integrate sensations across time and space to achieve a stable world of objects. This has been termed the *binding* problem—what sensations go with which objects (Burwick, 2014).

The correspondence problem comes in many guises, but the fundamental issue is whether objects or events that occur at different positions, orientations, shapes, intensities, pitches, rhythms, and so on represent the reoccurrence of the same object or represent a different one. I imagine that the observer constructs a trajectory that can link the sensations that occur at different locations and times. Some of these transformations could be predictable (geometrically) particularly for rigid objects. These trajectories or transformations link objects in different orientations, at different pitches, and at different rhythms.

I was driving home one night and noticed two headlights at the same height coming towards me, and naturally assumed that they were mounted on one car, but then the headlights diverged. My first reaction, it was in the 1960s when the TV show *Candid Camera* was extremely popular, that this was a stunt: it was a rubber car or at least a car that was able to shift the position of its headlights. Actually it was two similar motorcycles moving relative to each other. Given my assumption that it was one car, I expected the lights to change position relative to each other based on the geometric properties of rigid bodies, and groped for an explanation when that expectation failed. It is worth noting that I finally realized that they were two motorcycles by the exhaust sounds. Perceiving almost always involves more than one sense. This will be a constant theme in all chapters.

Other transformations would not be as predictable since instruments and singers do not sound alike at widely different pitches and baby pictures do not undergo predictable change as they become adult pictures. Nonetheless, all of these transformations would bind together those pictures or sounds that stem from one object and segregate those that stem from different ones.

The correspondence problem asks whether the "before and after" visual, auditory, and tactual sensations come from the same object. This correspondence might be easy to determine for a single bouncing ball due to its visual trajectory and impact sounds, but it would be much harder to determine which notes correspond to each violinist in an orchestra from the bowing movements.

What this means is that cognitive and physiological constraints along with the uncertainty of the sensory stimulation require the perceptual system to make use of assumptions based on prior experiences with objects in the world in order to perceive accurately. Otherwise, because the sensations could not always correctly make out objects, the world would be ambiguous (Pizlo, 2001). The transformation of the visual, auditory, or tactual sensations into objects depends on many stages and processes at every physiological level. Interactions transform the sensations at the individual receptors in the eye, ear, or hand into information. Cells in the eye, ear, or hand merely respond to energy at different frequencies and pressures. It is the receptive fields in the visual system based on combinations of retinal cells that respond to edges that simultaneously bound objects and separate them from others, it is receptive fields in the auditory system that respond to frequency glides that are characteristic of the sound of many objects (Hubel & Weisel, 1962), and it is interactions of sensors in the skin and muscle that give rise to shape and texture. Further processing isolates the predictable ways objects change over space and time and that predictability shapes our ability to make the world coherent.

The iOS game app Shadowmatic captures this ambiguity by projecting the shadows of familiar objects as those objects undergo various rotations. It is not easy to identify the objects.

In general terms, the organization of the nervous system is a progression from local to global integration (Park & Friston, 2013). Fundamental is that regions of the brain are specialized for a variety of functions such as language, visual recognition, and motor control. To make this work, inputs from different sensory organs are processed in parallel in distinct regions of the cortex. The basic circuitry is similar and hierarchical across modalities: the input ascends to the lower brain centers onto cortical regions, back to the lower regions, and finally back to the cortex (Frangeul et al., 2016). Individual cells are combined into local hubs with dense interconnections that maximize information transfer and processing. These are initially assembled into higher-level global units in development that bring together neural firings from different modalities and further assembled into cortical hubs as illustrated in Fig. 1.1. These higher levels encompass larger spatial regions of the visual, auditory, and tactual fields and longer temporal intervals for all senses; constant feedback at "higher" levels changes the properties of the "lower" levels. Even though information about individual properties (e.g., size, orientation, and location) might be coded in separate "vertical" neural tracts, those tracts need to be integrated because the interpretation of shape is a function of orientation and vice versa. Brincat,

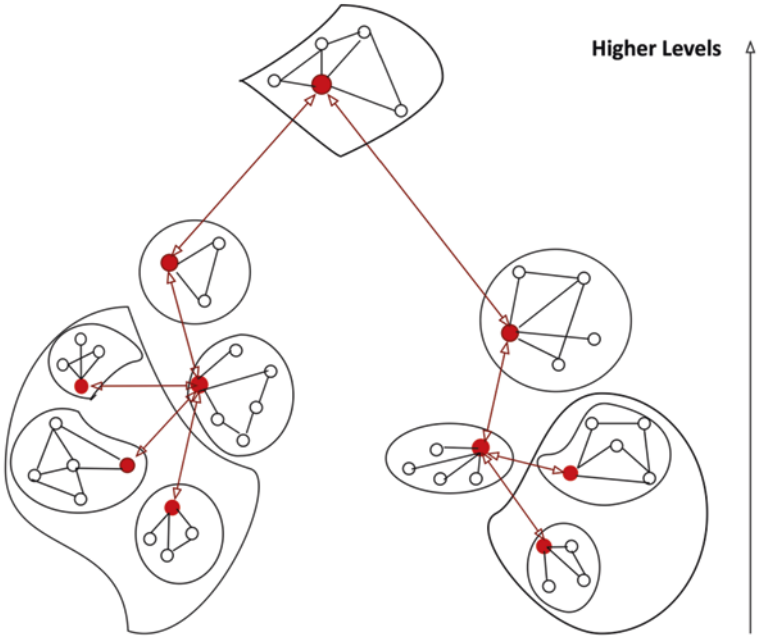


Fig. 1.1 An illustration of the rich club hierarchical modular organization of the brain. Individual nodes (open circles) composed of groups of cells are interconnected anatomically and fire at the same time to individual stimuli. "Rich hubs" (red circles) unify the firings of the individual nodes at a local level. The rich hubs connect at higher levels for specialized functions and ultimately connect at regions of the cortex to create modules for perception and cognition. The connections go from lower to higher modules, but there are feedback connections from the higher levels back to the lower levels depicted by the arrows in both directions. (Adapted from Park and Friston (2013))

Siegel, von Nicolai, and Miller (2018) found that regions in the cortex took on different roles and interacted in unique ways as a function of the task. A hoop changes shape as it rotates vertically. If the levels and tracts were not tightly coupled, it would be impossible to determine if the hoop had rotated or had changed shape (Bassett & Gazzaniga, 2011; van den Heuvel & Sporns, 2013).

To summarize to this point, an accurate construction of the physical world depends on the predictable parts and transformations of objects as well as an integrated nervous system that can abstract those properties as shown in Fig. 1.2. It is this predictability that allows the same object to be identified in different environments. Moreover, visual, auditory, and tactual objects exist in a common extended spatial and temporal framework so that each sense can affect the other. Why you see, hear, or feel what you do is a joint product of the physical stimulation and perceptual processes. In most cases, objects give rise to sensations and information in multiple senses so that cross modal perception is the norm.

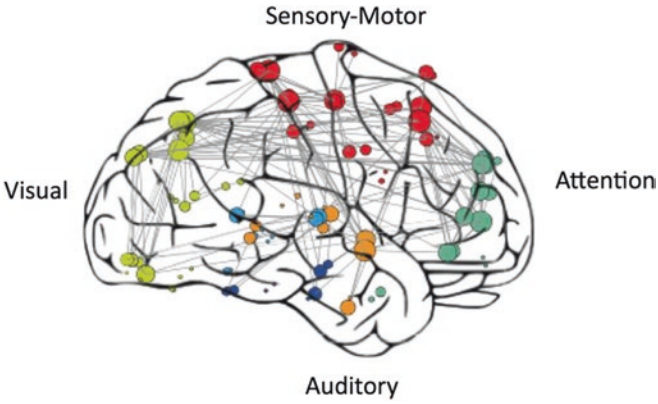


Fig. 1.2 Nodes in localized brain regions underlying specific functions (in different colors) are tightly interconnected; straight lines represent the connections. (Adapted from Bassett and Gazzaniga (2011) and Bullmore and Bassett (2011))

1.2 Similarities Between Perceiving and Business Decision-Making

One way to highlight the assumptions underlying perception is to compare them to assumptions about decision-making in business organizations. While this might seem to be a far-fetched analogy, I think the two have much in common. Both are hierarchically organized with data gathered at the lowest level being successively transformed as it moves up the "ranks." Activities at each level affect all other levels, and purportedly there is a decision-maker at the highest level. Yet there are obvious differences. One is that in organizations the number of workers decreases as one goes up the decision tree, but in perceptual systems the number of neurons in the auditory and visual cortexes is far greater than the number of receptors.

Roberto (2009) lists five myths of business decision-making and the reality underlying each of them and many are comparable to the myths and realities of perceptual decision-making. In business, as in perception, decision-making involves multiple processes and interactions among many levels. The information gathered at the lowest level moves up the system and to a great degree limits the outcomes. Much of the decision-making occurs "off stage," making it difficult to understand how a decision is reached. Current decisions, moreover, are framed by prior outcomes, delimiting which alternatives are considered. As in my headlight example, what you expect to happen can "blind" you to the present.

I have added three more myths (marked by stars) implicit in Roberto's lectures. Businesses in widely divergent fields face the same management problems; success and failure rest on the same solutions. In similar fashion, visual, auditory, and tactual sensations have the same spatial and temporal structures, so that senses must use the same processes to form objects and sources. Moreover, the ability of a business to meet new situations depends on the ability

to enlist different parts of the organization to solve the emerging problems interactively. In similar fashion, perceptual systems solve new problems (e.g., hitting a curve ball, playing the banjo) by combining existing perceptual and motor processes in novel ways. Finally, although complex statistical models may yield accurate decision rules, simpler rules of thumb often equal their success rate. In similar fashion, we think the goal of perceiving is to accurately reflect the nature of the physical world, but really the goal of perceiving may be only a successful action, not a perfect representation.

Business decision-making	*Perceptual decision-making*
1. **MYTH**: The chief executive decides. **Reality**: Simultaneous activity at various levels of organization	1. **MYTH**: Single cells decide ("grandmother" cells: one cell represents one person or object) **Reality**: Perhaps as many as 30 million cells are required to represent one image
2. **MYTH**: Decisions are made in the room **Reality**: Real work occurs "off-line"	2. **MYTH**: Percepts determined only in higher cortical centers **Reality**: Processing starting at eye, ear, and hand determines the ultimate percept
3. **MYTH**: Decisions are largely intellectual decisions **Reality**: Decisions are complex social, emotional, and political processes	3. **MYTH**: Percepts are rational, based on sensory information **Reality**: Most percepts emerge from unreflected processes
4. **MYTH**: Managers analyze and then decide **Reality**: Strategic decisions unfold in a non-linear fashion with solutions often occurring before defining the problem or analyzing alternatives	4. **MYTH**: The visual cortex (back of head) combines all inputs and then decides **Reality**: There are many cortical centers (aural, visual, tactual) that transform and constrain the neural signal along the way
5. **MYTH**: Managers decide and then act **Reality**: Decisions evolve over time and involve iterative process of choices and actions	5. **MYTH**: Percepts emerge and resist change **Reality**: Percepts evolve as additional sensory information occurs
*6. **MYTH**: Each business is unique **Reality**: The success and failure of all businesses are based on the same factors	*6. **MYTH**: Vision, audition, and touch are fundamentally different. Vision is spatial, audition is temporal, touch is spatial and temporal **Reality**: Vision, audition, and touch are all based on similar spatial and temporal changes
*7. **MYTH**: Business success depends on the quality of different departments **Reality**: Success depends on continual interactions among departments	*7. **MYTH**: Skills depend on the quality of different brain regions (e.g., musical center, speech center) **Reality**: Skills depend on the brain's ability to reorganize to meet different situations. Perceiving is based on inputs from all senses at once
*8. **MYTH**: Successful decisions depend on exhaustive analyses, possibly based on big data or artificial intelligence **Reality**: Rules of thumb are often equally successful	*8. **MYTH**: The goal of perceiving is to accurately mirror the environment **Reality**: The goal of perceiving is to take the correct action; a perfect rendering of the environment is not necessary

1.3 Summary

As described previously, abstracting things from sensations involves many interdependent perceptual processes. Edges and boundaries arise when abrupt changes in color, brightness, and/or texture break the visual world into discrete objects, when pitch jumps and rhythmic gaps break sounds in discrete sources, and when discontinuous surface features break hard (and squishy) materials into two-dimensional surfaces or three-dimensional objects. Chapter 2 will consider how those visual, auditory, and tactual contours lead to the perception of "figures" in front of "backgrounds" creating a three-dimensional world. But, there are instances in which the ambiguity of the sensation leads to the alternation of plausible but incompatible figures. Even though the most common examples of reversing figures come from simple visual drawings without context, reversing figures also occur in auditory sequences. It is interesting to note here that it is virtually impossible to stop the alternation even if you try to do so by concentrating on one percept. Chapter 3 will show that this alternation is a joint function of cell fatigue and cortical attentional processes. Rhythms also are figures with strong-weak repetitions in time and Chap. 4 will illustrate how auditory, tactual, and visual rhythms are based on levels of accentuation. Chapter 5 will make use of the source-filter model to understand the production and perception of color, timbre, and the process of echolocation. Chapter 6 will summarize these concepts.

References

Bassett, D. S., & Gazzaniga, M. S. (2011). Understanding complexity in the human brain. *Trends in Cognitive Science, 15*(5), 200–209. https://doi.org/10.1016/j.tics.2011.03.006

Brincat, S. L., Siegel, M., von Nicolai, C., & Miller, E. K. (2018). Gradual progression from sensory to task-related processing in cerebral cortex. *Proceeedings of the National Academy of Sciences, 115*(30), E7202–E7211. https://doi.org/10.1073/pnas.1717075115

Bullmore, E. T., & Bassett, D. S. (2011). Brain graphs: Graphical models of the human brain connectome. *Annual Review of Clinical Psychology, 7*, 113–140. https://doi.org/10.1146/annurev-clinpsy-040510-143934

Burwick, T. (2014). The binding problem. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*, 305–315. https://doi.org/10.2002/wcs.1279

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences, 7*(6), 252–256. https://doi.org/10.1016/S1364-6613(03)00111-6

Felin, T., Koenderink, J., & Krueger, J. I. (2017). Rationality, perception and the all-seeing eye. *Psychonomic Bulletin & Review, 24*, 1040–1059. https://doi.org/10.3758/s13423-016-1198-z

Frangeul, L., Pouchelon, G., Telley, L., Lefort, S., Luscher, C., & Jabaudon, D. (2016). A cross-modal genetic framework for the development and plasticity of sensory pathways. *Nature, 538*, 96–98. https://doi.org/10.1038/nature19770

Handel, S. (2006). *Perceptual coherence*. New York, NY: Oxford University Press.

Hubel, D. H., & Weisel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology (London), 160*, 106–154.

Park, H.-J., & Friston, K. (2013). Structural and functional brain networks: From connections to cognition. *Science, 342*(6158), 577–587. https://doi.org/10.1126/science.1238411

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research, 41*, 3145–3161.

Roberto, M. (2009). *The art of decision making*, *The great courses.* Chantilly, VA: The Teaching Company.

van den Heuvel, M. P., & Sporns, O. (2013). Network hubs in the human brain. *Trends in Cognitive Sciences, 17*(12), 683–696. https://doi.org/10.1016/j.tics.2013.09.012

CHAPTER 2

Objects and Events

2.1 Introduction

The fundamental issue for all possible perceptual theories is how changing visual stimulation is split into unchanging objects whose appearance may vary over time, how inherently changing auditory stimulation is broken into stable sound events, and how exploratory hand movements are converted into surfaces and solids. All these processes occur so naturally and automatically that we think that the world is split up into stable and discrete things. But, surprisingly, even though it seems that way, spatial separation, silent intervals, and empty spaces often do not mark object boundaries.

Feldman (2003) and Griffiths and Warren (2004) have attempted to define, respectively, visual and auditory objects. Nearly all of their principles correspond (and it is easy to extend them to tactual spatial objects):

(a) Objects are the units of our perceived physical world. They are spatially and temporally coherent bundles of visual or auditory (and material) stuff; the perceptual problem is to isolate information about those bundles from the overlapping information about the rest of the world.
(b) Objects seem to be things; we believe they have independent existence, with relatively unchanging properties and attributes.
(c) Objects are things we think are fixed in a world of changing appearances. Objects will look, sound, and feel differently at every occurrence, and it is our belief in the stability of the objects and their properties that allow us to perceive objects as identical under different conditions.

Electronic Supplementary Material: The online version of this chapter (https://doi.org/10.1007/978-3-319-96337-2_2) contains supplementary material, which is available to authorized users.

S. Handel, *Perceptual Organization*,
https://doi.org/10.1007/978-3-319-96337-2_2

To convey the characteristics of objects is a very difficult problem, often sidestepped, as most theories start with bounded objects and events. One part of the difficulty is that there are many kinds of objects with differing spatial and temporal properties, such as light flashes and drumbeats or sea fog and drone sounds. Another problem is that objects exist at many levels of space and time; a roof shingle becomes part of a roof that becomes part of a house and so on. Similarly, a humpback whale sound unit becomes part of a phrase, which becomes part of a theme, which becomes part of a repeating song. Each level gives overlapping information about the source and event. Rather than attempting to create an overarching theory encompassing all possible objects and events, we will focus on several types of objects and try to derive some general principles.

I think it is a useful simplification to assume that the peripheral nervous system transforms the pixel-like stimulation at the retina, the wave-like stimulation in the inner ear, and the pressures, deformations, and vibrations on the skin into discrete parts: straight and curved lines, dots, colored blobs, angles, vibration frequencies, glides between two pitches, surface roughness, hardness, and compliance, and so on. These are the results of the transformations in the "hubs" discussed in Chap. 1. The central nervous system's task is to group the parts into unified but constantly changing objects and locate those shapes on the surfaces of a three-dimensional world. Again, it is the assembling of the middle-level hubs into the cortical hubs that underlie this process.

There have been two major proposals to explain the emergence of "real" objects, that is, their intrinsic properties. The first is based on perceptual simplicity and is usually associated with the Gestalt psychologists whose general principle was *prägnanz*; what you see, hear, and feel will be the simplest and least complex structure given the stimulation. Thus, given that all *proximal* stimulation at the eye, ear, or hand can be due to many possible *distal* objects or events, the one that is perceived would minimize complexity. One tries out alternative possibilities. Unique to the Gestalt view was that prägnanz was understood to be due to the operation of electric currents in the brain cortex. The end result of the proximal stimulus at the receptors was a set of brain electrical currents that we can imagine correspond to the edges of an object. Those currents determined what we see. They were assumed to be capable of flowing freely in the cortex and ultimately to adopt the simplest, most regular spatial configuration that minimized energy, limited by the stimulation. That last phrase, *limited by the stimulation*, is critical. Prägnanz may lead to one percept as opposed to another one, but it will not "square-up" a lopsided rectangle or fill in a gap in a circle. In Gestalt theory, then, there is neural-perceptual isomorphism: what we see mirrors the flow of the brain currents. Prägnanz seems most to have to do with vision, and less to hearing and touch. The role of learning and past experience is acknowledged, but it is just one of many factors that control grouping.

YouTube Videos

Gestalt Psychology (in 3 parts): Michael Wertheimer and David Peterzell, by DHPPhDPhD, 2010. An interview with the son of Max Wertheimer, one of the originators of Gestalt Psychology.

https://www.youtube.com/watch?v=5_fvAMZh3J8; https://www.youtube.com/watch?v=N-i8AKV0LFk; https://www.youtube.com/watch?v=YnTu8UDWnGY

GESTALT; Tatiartes, 2012. Lots of interesting images, many of which will be discussed in Chap. 3 on multistable images.

The second major proposal is based on the principle that sensations will be perceived as that object most likely (probabilistically) to have occurred in that environment. Originally, the perceptual process involves conscious problem solving using feedback from successful and unsuccessful outcomes, but with experience that process gets telescoped and proceeds without awareness; for obvious reasons this has been termed unconscious inference. The entire process has been called *inverse perception* because the perceiver must work backwards from the sensations at the receptors to the object in the environment. The observer cycles through the possibilities and chooses the one with the highest probability. One of the appeals of this approach is that it focuses on the accuracy of perception and thus on its survival value.

Currently, the emphasis is whether people make use of Bayesian methods to derive their percepts and whether that will maximize the probability of choosing the correct decision. From a Bayesian perspective, the decision process must start with a range of plausible hypotheses based on the context; the *prior* probability of each hypothesis then is constantly updated into *posterior* probabilities as new data (i.e., sensations) are gathered until we have to make a decision. The probabilities are updated by multiplying the prior probability of a hypotheses by the probability of the sensations given that that hypothesis is correct, so that the updated

$$\Pr\left(hypothesis\ A\right) = \Pr\left(sensation \mid hypothesis\ A\right) \times prior\ \Pr\left(hypothesis\ A\right)$$

$$\mathbf{posterior} = \mathbf{likelihood} \times \mathbf{prior}$$

The Pr (sensation | hypothesis A) is the probability of the given sensation if hypothesis A is true and is termed the *likelihood* and the updated probability is termed the *posterior* probability.

Consider the "trick" car incident. The prior probabilities, representing my prior beliefs at that time, might be

Pr (real car) = 0.85
Pr (two motorcycles) = 0.10
Pr (trick car) = 0.05

With the lights at the same height in the distance, which was my first impression, the likelihoods for the real car and trick car are equal; each type of car would create the same visual sensations. But, the likelihood for two motorcycles would be less because two lights at the same height would require two nearly identical motorcycles moving in parallel and that is unlikely. Thus, the posterior probability for the hypothesis of two motorcycles would decrease.

However, as the object(s) moved around a curve, the lights began to diverge and the likelihoods change dramatically. Now:

Pr (diverging lights|real car) = 0
Pr (diverging lights|motorcycle) = 0.95 (my guess)
Pr (diverging lights|trick car) = 0.25 (my guess)

At this point, the posterior probability of a real car dropped to zero, while the posterior probability of two motorcycles jumped up and the posterior probability of a trick car increased slightly. As the object approached further, the characteristic sound of motorcycles became apparent and the posterior probabilities changed further because the likelihoods diverged:

Pr (motorcycle sounds|motorcycle) = 1.0
Pr (motorcycle sounds|trick car) = 0.0

Now, the only remaining hypothesis is that of two motorcycles.

In general, sensations are ambiguous and would support several percepts. We must choose one of these based on expectations from our prior experiences, our present actions (e.g., staying on the right side of the road), the cost of making the wrong decision, and on the present sensations. Even if our prior hypotheses are wildly wrong, the incoming sensations should correct those probabilities and lead to the correct decision. Colloquially, likelihoods swamp priors. In the end, the posterior probabilities represent our *beliefs* in the competing hypotheses. There is a nice parallel between the inverse perception problem and Bayesian inference. Both work backwards from the current sensations to the most probable object or most probable hypothesis, that is, state of nature. Given this, it seems natural to employ Bayesian models.

But is this a realistic perceptual model? How do we enumerate the initial hypotheses and obtain the initial probabilities, and do these probabilities reflect the environmental statistics or are they internal guesses subject to various kinds of errors (Kahneman, 2011)? There is no guarantee that we even have included the correct alternative. Bayesian procedures allow us to compare our beliefs in the alternatives, but do not allow us to know if the most probable alternative is true. How can we tell if each new sensation is even relevant to the hypotheses, and what cognitive processes do we employ to upgrade the initial probabilities? There is not much time to avoid a thrown snowball. Nonetheless, there has been a proliferation of Bayesian models in all aspects of human and animal behavior, even though there is disagreement as to whether these models help explain behavior (Jones & Love, 2011).

We argue in what follows that the way we organize visual, auditory, and tactual sensations into objects and sources is basically the same. Even though the sensations are different, passive and spatial for visual, passive and temporal for hearing, and active and both spatial and temporal for touching, the organizing principles are equivalent. Clearly this is a simplification because all three involve active and passive actions and all include spatial and temporal structure, but it still gives a basis for understanding differences among the sensations. We start by considering the organization of discrete visual and tactual sensations into objects and discrete sounds into sources. But objects and sounds rarely if ever occur in isolation. Spatial objects butt into each other, interlock, pierce one another, sit on top of one another in different orientations and thereby create a mosaic of shapes, edges, colors, and textures. Sound waves that overlap in time merge so that the sound waves from each source become mixed together. Moreover, except in rare situations (like psychology experiments), objects, events, and sources give rise to inputs in more than one sense. Therefore, the perceptual problem is to cleave those jumbled composites into one or more coherent objects and sources. Given the ubiquity of these principles across species and senses, we will show how one of the goals of camouflage is to blur or disrupt the edges, contours, and colorations that make prey visible to predators.

2.2 Grouping Principles

We start by assuming there are discrete elements in the visual perceptual field, that is, straight and curved lines, dots, and colored blobs as givens, and ignore for the time being the ways in which the nervous system constructs those elements. In similar fashion, we will assume the existence of discrete sounds and tactual impressions that occur one after the other. At first we will imagine that these elements exist in isolation, and then consider the organization of overlapping elements in time and space resulting in figure-ground organization.

2.2.1 Gestalt Principles for Non-Overlapping Visual Arrays

The number and types of grouping principles for discrete elements arrayed across space have changed back and forth since the initial descriptions by Wertheimer (1923). When elements are identical and equally spaced as in Fig. 2.1A, there is only a weak tendency to group those elements, possibly into twos or threes. As the elements get more closely packed, then the groups would tend to include more elements (Fig. 2.1B & C). When the elements differ in quality or spacing, the most basic principles seem to be proximity (Fig. 2.1D), similarity (Fig. 2.1E), and common fate (Fig. 2.1I). Elements that are closer together (i.e., proximity), or that are similar in color, shape, or size, or that tend to move in the same way (i.e., common fate) are grouped together. Other grouping principles include connectedness (Fig. 2.1J), continuity of curved lines or white and black dots in a line (Fig. 2.1K), closure (Fig. 2.1L), parallelism (Fig. 2.1M), and symmetry (Fig. 2.1N). Examples of these principles are given in Fig. 2.1.

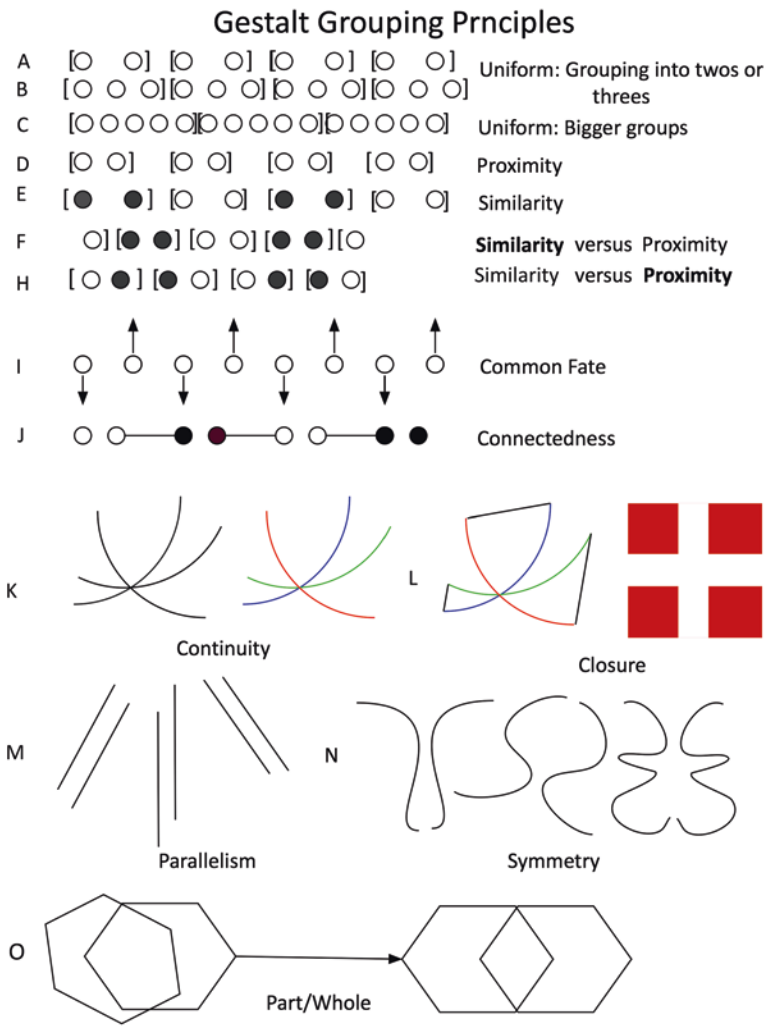


Fig. 2.1 Examples of the classical Gestalt grouping principles. It is easy to see how the groupings change as the proximity or similarity among elements is varied (F & H) or when extra elements are added or subtracted. Connectedness (J) can overcome the principles of similarity and proximity. The three arcs seen in the example of continuity (K) are broken apart when lines are added to create enclosed segments in the example for closure (L). Closure also can bring about the perception of illusionary contours when seeing the white cross in 2.1L. The rearrangement of parts of a figure can bring about a more structured Gestalt (O). The most important principles in the construction of three-dimensional objects from the two-dimensional visual input are probably parallelism and symmetry

It is important to realize that these demonstrations of the organizational principles were designed to be clear-cut and unambiguous. In real-life scenes, it would be rare for all the grouping principles to lead to the same organization. In Fig. 2.2A, we start with a conventional drawing illustrating good continuation using black lines of equal thickness. In Fig. 2.2B, the color of the lines is varied, leading to a stronger tendency to see the drawing as black lines versus red lines (i.e., due to similarity) that violate the continuity principle. But, continuity still dominates. Finally in Fig. 2.2C, the black lines are thickened, dramatically changing the organization.

By varying the strength of each grouping principle, it is possible to balance one against another. Thinning the black line in Fig. 2.2C, for example, would weaken color (similarity) organization. We will find the same trade-off in comparing Sound Files 2.3D and 2.3E; decreasing the difference in the silent interval

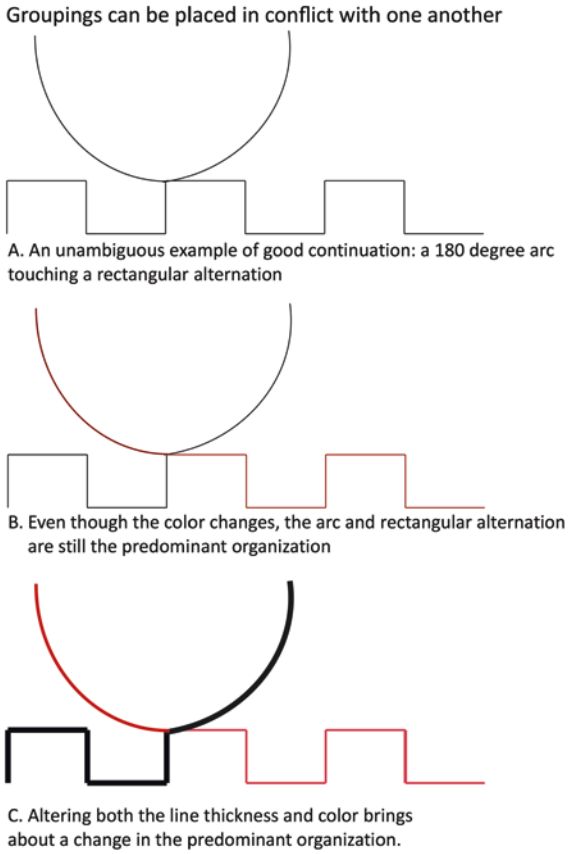


Fig. 2.2 The perceptual grouping is a reflection of the relative strengths of the grouping principles, which can be easily altered. Here, the shift is from continuity to color/line thickness similarity

between the tones that brought about grouping by proximity now brings about grouping by similarity. Moreover, people may be more sensitive to one grouping principle than another. One might be dominated by color, the other by proximity. Grouping is not all or none; people must choose among the alternative organizations and when the cues are contradictory or ambiguous, the percept may alternate as found for the multistable objects we find in the next chapter.

2.2.2 Gestalt Principles for Non-Overlapping Sound Sequences

Similar, if not identical, principles exist for auditory sequences. A series of *isochronous* (i.e., equal intervals between onsets) short beeps are organized into equal-sized groups depending on their rate. When the beeps differ in onset timing and or quality, then the same principles of proximity, similarity, common fate, and so on determine the grouping of the sounds (Fig. 2.3).

2.2.3 Gestalt Principles for Non-overlapping Tactual Objects and Surfaces

The overwhelming majority of research based on the Gestalt grouping principles has been done in vision. Gallace and Spence (2011) found that visual studies outnumbered auditory ones by a ratio of 8:1 and outnumbered tactual ones by a ratio of 16:1. It is easy to identify reasons for these outcomes: (a) Vision seems to be our dominant sense; (b) It was much easier to create precise visual stimuli than auditory or tactual ones although with the advent of computers

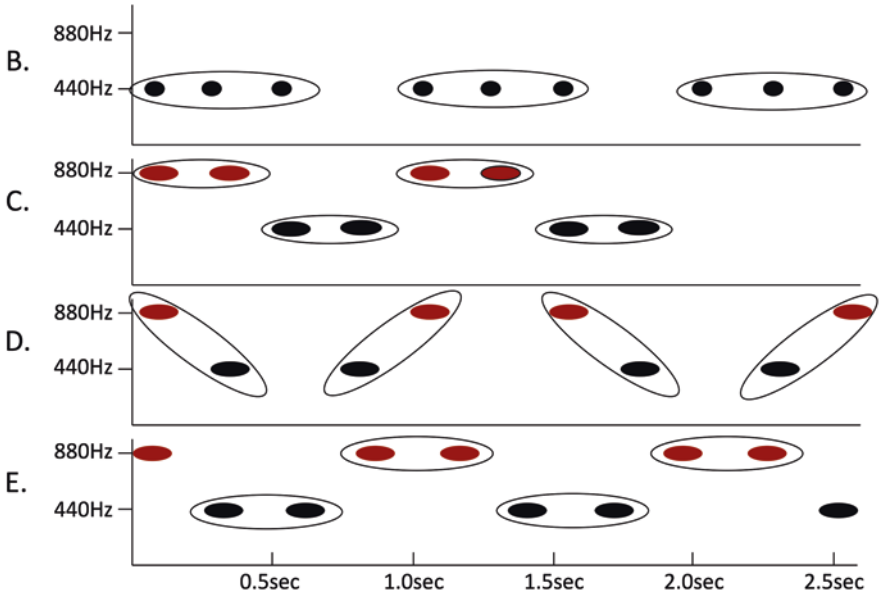


Fig. 2.3 Illustrations of sound files 2.3B–E

Sound Files 2.3: Demonstrations of grouping by frequency similarity and temporal proximity that correspond to Fig. 2.3B–E

that is no longer true; (c) The goal of the grouping principles was to understand how visual scenes were broken into discrete overlapping objects and how auditory sequences were broken into separate sources. But, tactual objects tend to be self-contained, although they may overlap. Thus, the goal of haptic research changed to understand the perception of the properties of those objects, and how those properties underlie the ability to manipulate objects; (d) I think researchers were intrigued by the Gestalt notion of electrical brain currents that resembled the visual perception and by the discovery of "line" detectors in the visual system. It is much harder to imagine brain currents or detectors that resemble auditory or tactual percepts.

Traditionally, vision and hearing have been termed the higher senses and touch was relegated to the lower senses. Nonetheless, David Katz (1925) argued that touch sensations have the most compelling sense of the reality of the external world and cites two quotes from Kant to bolster his contention: "the hand is man's outer brain" (Page 28) and "it is the only sense of direct external perception and for that reason is the most important sense" (Page 240). Eyes and ears are locked into fixed positions in the head, and while each will focus on aspects of sensations at a distance neither can reach out to actively explore objects. In contrast, touch is limited by our reach; in Chap. 4 we will discuss how handheld probes extend that reach. There are several common expressions that illustrate the connection between touch and cognition: to have a grip on the facts, to grasp the concept, to have a handle on it, to have at one's fingertips, to know as well as the back of one's hand.

What makes the study of touch so interesting is that it is so diverse that there is no single subjective characteristic that encompasses it and yet it is one sense. Touching brings forth a particular set of physical properties through the activity of the cutaneous receptors in the skin as well as the kinesthetic receptors in muscles, tendons, and joints. The receptors are so intertwined that it is impossible to make a one-to-one connection between a sensation and a kind of receptor (Hayward, 2018). Movements have a purpose and require the coordination of groups of muscles over time. The diversity of these movements yields complex behaviors. The term "haptic perception" is often used to encompass the role of all such receptors.

The essential characteristic of tactual or haptic perception is that it is serial, a series of hand movements that are used to explore surfaces and objects in order to obtain information about these properties. In this way, touch is similar to audition in which the information arrives temporally, as opposed to vision in which the scene by and large can be understood at once. Across a wide variety of tasks, tactual exploration takes at least twice as long as visual exploration. The purposive hand movements are chosen to maximize the pick-up of those properties and should be thought of as information seeking rather than sensation seeking. The hand motions can scan the surfaces to detect the material and shape of objects by deformations of the skin surface, or encircle objects to identify them. Only motion reveals the emergent properties, stationary sensations quickly disappear.

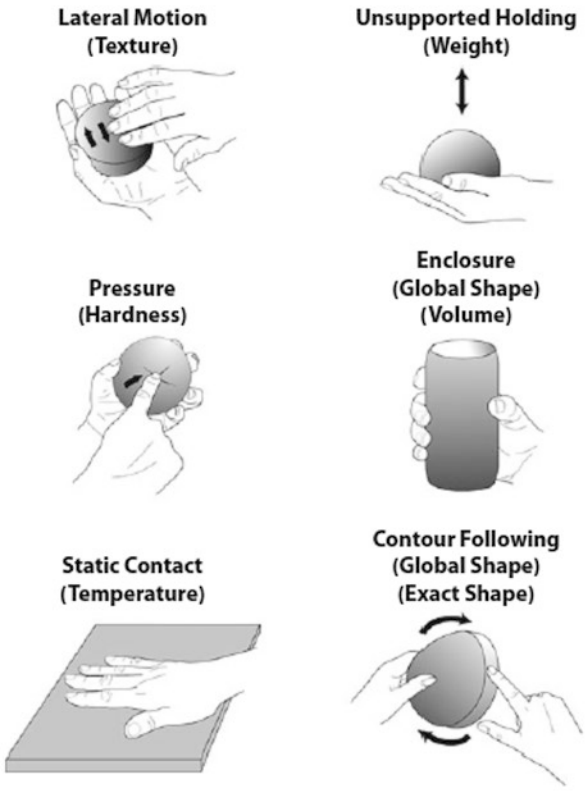


Fig. 2.4 The six exploratory procedures found by Lederman and Klatzky (1987). The "inside" region of the hand is critical. The ridges of the epidermis, which surprisingly act to reduce skin friction due to reduce surface contact, generate oscillations on the skin during sliding motions. Directly below is the "pulp" which allows the skin to conform to external surfaces. Moisture increases the surface friction and softens the external skin to better conform to surfaces. Pacini corpuscles seem mainly tuned to the skin vibrations and Meissner corpuscles seem mainly tuned to the small skin deformations. Yet, as Hayward (2018) points out, all perceptions are the result of a complex and interchangeable set of cues. The weight of an object is perceived to be identical whether it is held by a handle, held overhead, or lifted from a squat position. (Reproduced from Lederman & Klatsky, 1987: Fig. 1. Reprinted with permission, Elsevier)

Lederman and Klatzky (1987) have identified six such haptic exploratory procedures matched to six properties that can be discriminated by touch: texture (especially roughness, slipperiness, elasticity); compliance (softness); temperature; weight and balance; surface contour; and global or volume shape. These motions are illustrated in Fig. 2.4. Each hand motion is optimized to discriminate among values of one surface property. But as Lederman and Klatzky (1987) have shown, each one could be used to discriminate among

values of other properties to a lesser degree. The hand motions optimized to pick up surface roughness could also pick up the shape, compliance, or contour because all require a series of integrated following hand movements.

A wide variety of tasks have been employed to demonstrate that these properties are salient and "pop-out" in haptic perception. What is interesting is that several of these properties are not symmetrical; for example, it is easier to identify a cube based on its edges and vertices among smooth spheres than a sphere among cubes (Plaiser, Bergmann-Tiest, & Klappers, 2009). Moveable objects pop out from anchored ones (Van Polanen, Bergmann-Tiest, & Kappers, 2012). The exploratory movements are optimized for context; to determine compliance, people use higher forces when they expect more rigid surfaces (Bergmann-Tiest, 2010). Another aspect of this research is the demonstration that these exploratory movements often yield more than one property as discussed above. For example, lateral movements can give rise to roughness and hardness but not contour or shape, which are based on following movements. It is likely that when two properties are related, either material or structure, they will be co-processed so that combinations will be easier to perceive. The downside is that if one of the properties is changed, say from rough to smooth, it will be harder to recognize that the shape remained unchanged (Lacey, Lin, & Sathian, 2011).

As stated above, this research was designed to characterize the properties that underlie haptic perception, but were not meant to discover if the Gestalt grouping principles, found for seeing and hearing, were also applicable to touching and grasping. To investigate whether the proximity and similarity among tactual surfaces follows the same Gestalt principles as found for visual surfaces (Chang, Nesbitt, & Wilkins, 2007b) created identical layouts using three different surfaces that differed both in color and roughness (yellow/280 grit sandpaper, red/40 grit, and black/smooth cardboard). Subjects grouped the surfaces by color without hand movements, or by texture using hand movements while blindfolded. Some of the simpler layouts obviously would be grouped by similarity or proximity for both color and texture as illustrated in Fig. 2.5. Other layouts were more ambiguous and were organized differently by 30% of the participants. One such layout also is shown in Fig. 2.5B. The difference in grouping may be due to sweeping arm and hand motions used for the tactual grouping task. However, on the whole these outcomes suggest that the Gestalt principles of similarity and proximity bring about the same visual and tactual organization.

2.3 Figure Ground and Contour Organization

2.3.1 Visual Perception

2.3.1.1 Segregation into Interleaved Figures

We described the groups above in terms of how individual elements are linked together. In some cases, the linked elements form separate but overlapping groups. In Fig. 2.1I, for example, the upward moving dots would form one

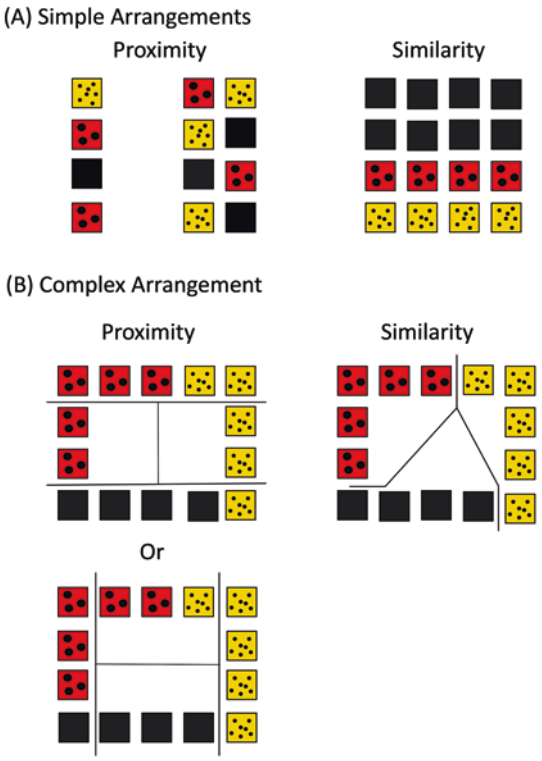


Fig. 2.5 Simple and complex arrangements for visual and tactual grouping. The 240-grit stimuli are represented by the yellow squares/small dots and the 40-grit stimuli by the red squares/larger dots and smooth stimuli by the black squares. These are hypothesized organizations based on similarity and proximity. Simple arrangements are invariably grouped in the same way visually and tactually. Complex arrangements sometimes give rise to different groupings

group, and the downward moving dots would form another group, just as in the Fig. 2.1K the horizontal line of black dots would form one group, while the horizontal line of open dots would form another. Although the elements in each group are interleaved, the groups seem to be at the same depth.

Our environment, however, is one of solid objects that abut against each other and overlap at different depths. The surfaces of these objects fill up the visual field and each is likely to be relatively uniform, rigid, self-contained, and be subject to same transformations (Palmer & Rock, 1994). Palmer and Rock argue that organization into units with uniform properties of color and brightness, termed "*uniform connectedness*" bounded by edges or contours that enclose the surface properties, would therefore be a highly probable way to start to organize the visual field. Even if the elements differ (spots versus lines) within the edges, they are likely to be part of one object. But at this point we

merely have surfaces without any sense of occlusion or depth; edges and contours are broken by other edges with no indication of what are the figures. Such figures could be in front either of other figures or continuous backgrounds. All we have are homogeneous surfaces and lines.

Although it is possible that all such surfaces could be perceived as being at the same depth, the more usual percept is that the surfaces lie at different depths. In the simplest cases, there is a "thing-like" figure region in front of a "shapeless" ground; the occluding figure appears shaped by the ground while the edges and contours that separate the figure from the ground are perceived to belong to the figure. The contours form depth edges. The ground appears to enclose the figure but does not have a shape itself because the border of the figure belongs to the figure itself without determining the shape of the ground.

A simple example is shown in Fig. 2.6A. Here the blue square figure can be seen in front of the grey ground or the blue figure as being a hole or a window

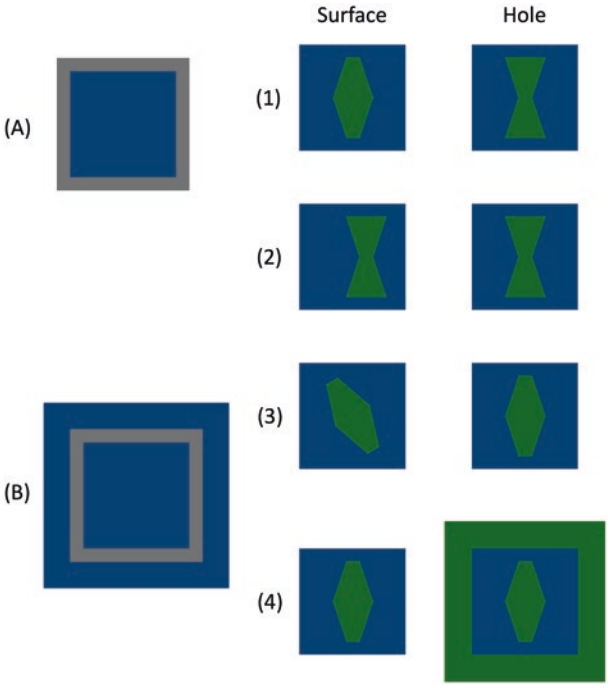


Fig. 2.6 In both (A) and (B), the blue regions can be seen either in front of or in back of the grey cut-out. Moreover, in (B), the blue regions can be seen as one or two surfaces. The perception of a green object sitting on the blue background is more likely if the object is concave (1), offset laterally (2), or at a different orientation (3) than the background. The perception of a hole in the blue surface is more likely if the shape is convex, centered on the background, and if the surrounding background matches the surface seen through the hole (4)

into a grey ground. The important point is that as the percept reverses the boundary never breaks apart; it is a single unit that surrounds the figure. If a blue border is added (Fig. 2.6B), the two blue regions can be seen as one surface or two, resulting in the perception of three levels.

Several factors help determine whether an enclosed shape is perceived as being an object on top of the surround or a hole that allows for perception into the background. As shown in Fig. 2.6, surface objects are convex, holes are concave (Fig. 2.6 (1)); surface objects are offset or turned at an angle to the surround, holes are centered and parallel to the surround (Fig. 2.6 (2) and (3)); and a hole is likely to be perceived when the color (or texture) seen through the shape matches a visible background. Of course, shading would also affect the surface/hole percept (see review by Bertamini & Casati, 2015)

The classic figure-ground principles that predict which surfaces will become figures are akin to those for grouping individual elements. These include:

1. Surroundedness: Any region completely surrounded by another is usually perceived as the figure in front of the surround (Fig. 2.7B & C)
2. Size: The smaller region is usually perceived as the figure (Fig. 2.7A & B)
3. Convexity: Convex regions are usually perceived as the figure (Fig. 2.7C)
4. Symmetry and parallelness: Symmetrical regions with parallel sides are usually perceived as the figure (Fig. 2.7C).

Although there can be confounding effects, for example, surrounded regions are necessarily smaller, most research has shown the convexity and symmetry are the stronger cues for figure-ground organization. Classical Gestalt theory postulates that the emergence of the figure is not based on previous experiences, but is due to the configural properties listed above. (The electrical field theory discussed in Chap. 1 while clearly linked to this view, actually came into prominence many years later). The figure-ground organization would be based solely on the image, and would precede the influence of previous experiences.

The modern view is more nuanced, and while including the configural factors above, also concludes that attention and past experience can affect which regions become figures (Wagemans et al., 2012). I think a useful way to think about figure-ground organization is that it is a competition among different possibilities. In some cases, the several alternative organizations seem equally strong, and can seem to shift back and forth (e.g., Fig. 2.7 and Fig. 3.1B & C). Here, the figure-ground organization can be thought of as being an example of multistable images that will be discussed in Chap. 3. In other cases, one organization is stronger, and that one "sticks." Even so, it is worthwhile to think about the resulting organization as being the result of interacting lower-level sensations and higher-level cognitive processes (see review by Peterson, 2015).

2.3.1.2 Occlusion of Overlapping Figures

The examples above seem to depict shapes in front of featureless backgrounds, but if two surfaces overlap, or if the figure covers part of the background, the

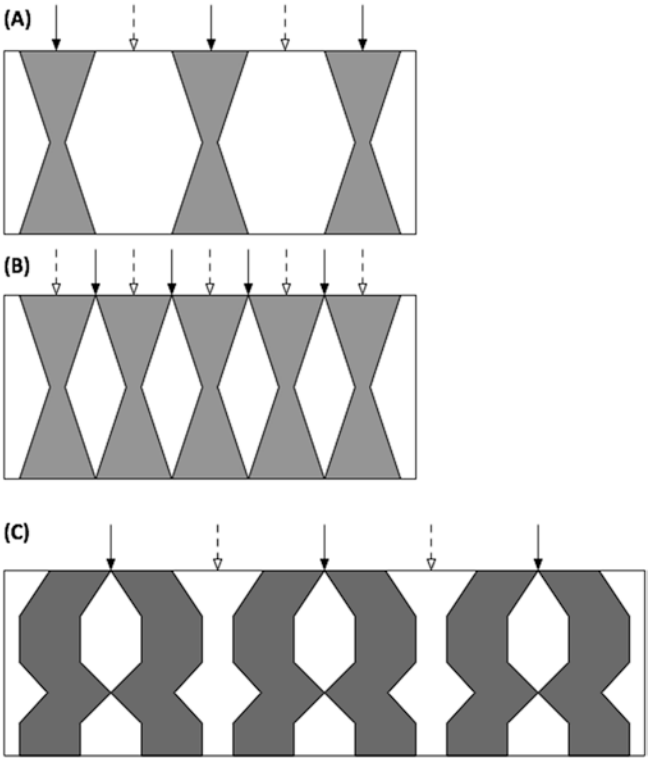


Fig. 2.7 Several factors influence the perception of the in-front figure and the behind ground. Comparison of (A) and (B) show the effect of size (and possibly convexity), and (C) shows the influence of convexity and parallelness. To me, the figure surfaces lie under the solid arrows, although it is easy to reverse the figure and ground and see it the other way

perceptual system must guess at the contour of the occluded part of the ground because the figure is covering it. People need to be able to guess the contour so that they can grasp objects and recognize objects from different perspectives.

If a smaller object partially occludes another, then the first perceptual problem is to determine if the two parts of the occluded object come from a single larger object or come from two different objects. In either case, the second problem is to estimate shape of the hidden contours. For the first problem, Kellman and Shipley (1991) have suggested a heuristic to predict whether the two parts of an occluded object are seen as part of a single continuous object. They termed this heuristic *relatability*, and there are two parts: first, if the edges that connect the sides are extended as straight lines, they should intersect; second, the bend at the intersection should not be greater than 90° (Fig. 2.8A, B, and C). If the bend is less than 90°, they further argue that the split regions will get connected regardless of brightness, color, or texture

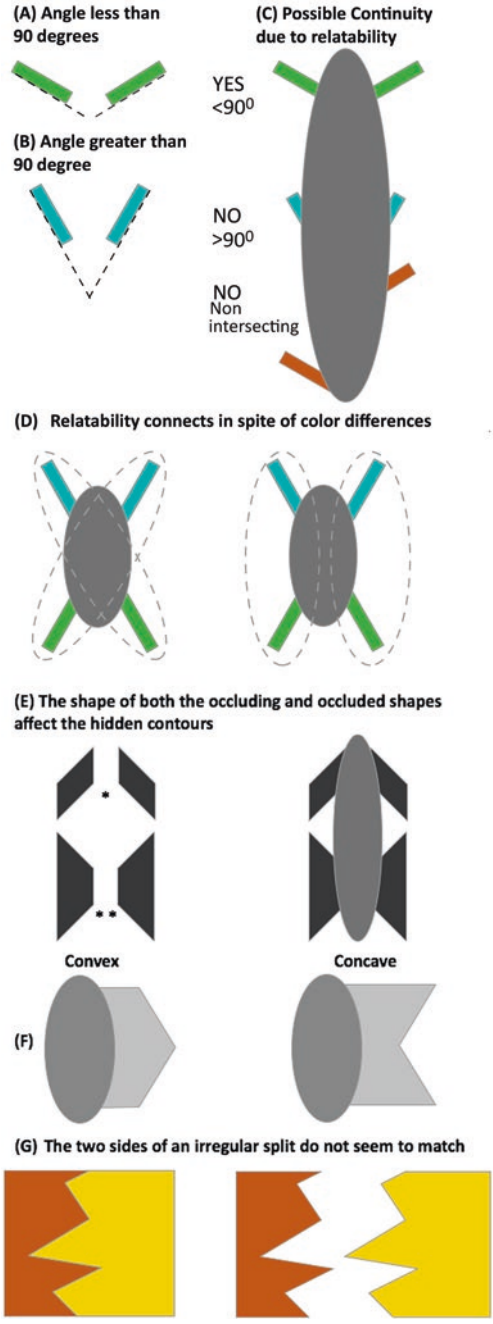


Fig. 2.8 In (A), the "turn" is less than 90° so that occluded parts would appear to be connected. But, in (B) the "turn" is greater than 90° so that the parts would not seem to connect. In (C), the different turn angles create the perception of connectedness only for the top green bars. In (D), relatable segments are connected in spite of color differences. The two blue bars and the two green bars are not connected because they violate the relatability constraint. In (E), the connecting contour seems more rounded in the upper segment (*) than in the lower segment (* *). In (F), the occluded section for the convex object on the left seems to be convex, but the occluded section of the concave object on the right appears concave. In (G), the two sides do not seem to go back together because the points of maximum convexity do not appear to line up

(Fig. 2.8D). The restriction to angles less than 90° may be too restrictive. Fulvio, Singh, and Maloney (2008) found that people would connect the parts if the angle was greater but the judgments were more variable.

If the split regions are seen as connected, then the other perceptual problem is to infer the contours of the occluded sections. Nearly any contour shape is possible, but people perceive only a limited set, demonstrating that our perceptual expectations impose strong constraints. If the surfaces on both sides of the occluding region are relatable, then people draw consistently smooth contours between the sides. This result may reflect the naturally occurring properties of real objects such that the majority of surfaces do lie along smooth contours. If the two sides are not relatable, the responses are inconsistent. The smooth contour follows the orientation and curvature but do not reflect changes in curvature. Given that the hidden contour can be any shape, the geometry of the occluding and occluded objects can also affect the perceived hidden contour as illustrated in Fig. 2.8D & 2.8E.

The feeling that hidden parts of objects are really there has been termed *amodal* perception. This feeling is based on the idea that the visual system extrapolates visible edges and surfaces into two- or three-dimensional objects. (We will suggest later in this chapter that the auditory system also extrapolates sounds into sources). For example, a shaded circle would be assumed to be a complete sphere. Magicians can make use of these unchecked assumptions. For example, the multiplying ball illusion is accomplished by hiding a second ball inside an empty half-shell ball. The viewer assumes the half-shell is really a solid sphere so that when the magician flips out the hidden ball it looks like the original ball (which was really a hollow sphere) has doubled (Ekroll, Sayim, & Wagemans, 2017). The audience is tricked not by the magician's misdirection but by their own misguided perceptual intuitions.

YouTube Video
Ridley's Magic How-to-Multiplying Balls

2.3.2 Auditory Perception

Online Resources: The following web sites have many auditory demonstrations that are related to material discussed below.

(A). http://webpages.mcgill.ca/staff/Group2/abregm1/web/downloadsdl.htm. ***These demonstrations are derived from the following audio compact disk: Bregman, A.S., & Ahad, P. (1996) Demonstrations of auditory scene analysis: The perceptual organization of sound. Auditory Perception Laboratory, McGill University. © Albert S. Bregman, 1995.***

(B). http://www4.uwm.edu/APL/demonstrations.html. **These demonstrations are derived from (Warren, 1999). (Both Dr. Bregman and Dr. Warren have made significant contributions to our understanding of auditory perception)**

2.3.2.1 Segregation into Interleaved Figures

Visual objects, with the rare exception of transparent objects, will block the appearance of one another when they overlap and thereby yield the perception of depth. This is not true for sound events. When two or more sounds occur at the same time, the sound waves from each source simply add and intermingle, scrambled together at the ear so that the acoustic input is inherently ambiguous. To disentangle the acoustic wave into the different sound sources the listener must make use of relationships among parts of the wave at one time point and relationships among parts of the wave across time points even without knowing how many sources there are. This has been termed the "cocktail party problem," trying to identify and track one voice amidst many. As Bregman (1990) puts it "it's like trying to figure out what went on in the harbor from the wave patterns lapping at your feet."

But, before considering the acoustic factors that allow listeners to partition overlapping sounds into those parts originating from each source, we will consider sequences of sounds that do not overlap and that could either be integrated into one figure or segregated into two or more figures composed of interleaved sounds. We start with the simplest case in which each tone is the same duration and the silent intervals between them are equal (termed *isochronous presentation*) while the discrete sounds vary in some way such as frequency or intensity. In this case, the series might be grouped into two or more subsequences, each termed a *stream,* determined by the magnitude of the differences in each property. One stream might consist of the lower frequency or intensity and a second the higher frequency or more intense sounds. Streams could be defined by other properties such as the sound quality, for example, a flute versus a clarinet, or the durations of the sounds.

To investigate stream segregation, a sequence of tones is continuously recycled and listeners indicate whether they hear all the sounds as coming from one integrated stream or some sounds as coming from one stream and the remaining sounds as coming from one or more different interleaved streams. The same principles that affect contour formation in seeing also affect the segregation of sounds into one or multiple streams, namely the similarity between the sounds (e.g., the frequency ratios), proximity in time (e.g., the duration of the silences between adjacent sounds), and good continuation (e.g., the smoothness of the transition between adjacent sounds). All three of these principles contribute to the sense of predictability for the sequence (Winkler, Denham, Mill, Bom, & Bendixen, 2012). To the extent that frequency similarity, timing, and continuity enhance the overall predictability of the sequence, one stream should predominate. To the extent that those principles enhance the predictability of the individual frequencies, two streams should predominate.

In most cases, the initial or default percept is that of a single stream although there is a tendency for streaming to increase as one continues to listen. If the frequency ratio between the tones is small enough, one stream is heard regardless of the presentation rate. As the frequency ratio is increased, there is a trade-off between proximity, the interval between the offset of one sound and the onset of the following sounds, and the frequency separation or any other variable such as

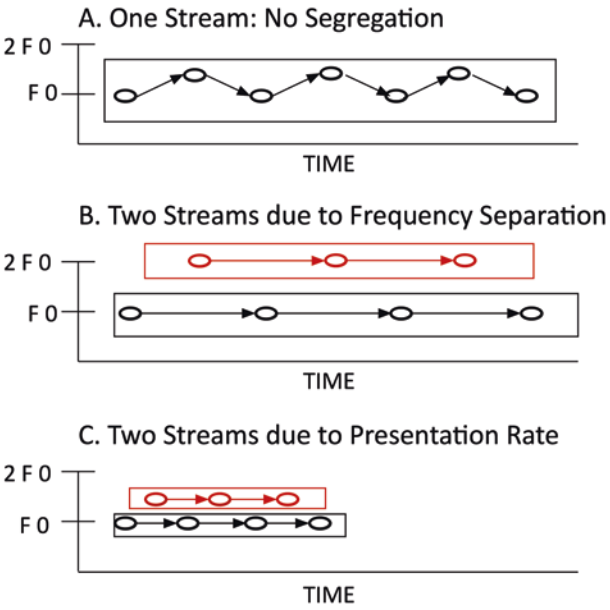


Fig. 2.9 Stream segregation arises if the frequency separation is increased (B) or the presentation rate is increased (C)

Sound Files 2.9: Demonstrations of one integrated stream and two interleaved streams as illustrated in Fig. 2.9A–C

timbre separation. Increasing the presentation rate so that the silent interval between the sounds decreases or increasing the frequency separation (or both) will increase the probability that streaming will occur. Conversely, decreasing either presentation rate or frequency separation (or both) increases the probability that all sounds will seem to come from one stream. It is therefore possible to balance the two outcomes by increasing one variable while decreasing the other (Fig. 2.9).

Bregman (1990) made an important distinction about the difference between one- and two-stream perception. He found that it is possible to attend to one low and high integrated stream as long as the frequency separation between the sounds was less than 10%. However, it is impossible to continue to hear one integrated stream when the presentation rate or frequency separation reaches certain values. There is an obligatory split between the two streams; Van Noorden (1975) suggests that stream segregation occurs prior to focused attention.

One of the consequences of stream segregation is that listeners lose the ability to correctly interleave the streams. Suppose we have a sequence A1B2C3A1B2C3… in which A, B, and C, are three different low-frequency tones and 1,2, and 3 are three different high-frequency tones. Listeners can attend to either stream and attention may shift spontaneously. They can report the order of each stream correctly, ABCABC as opposed to ACBACB and 123123 as opposed to 132132. But, listeners cannot report if the entire sequence was A1B2C3 or A2B3C1 or A3B1C2. The inability to keep the streams in registration is true whether the stream formation was due to

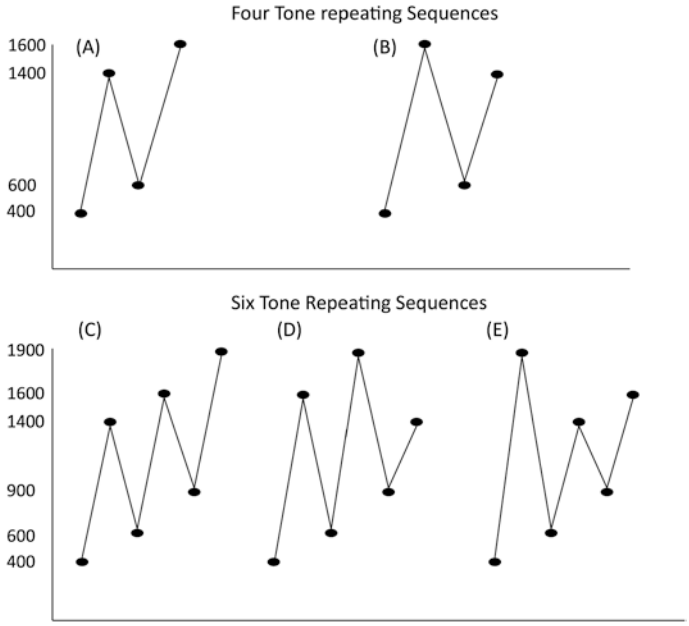


Fig. 2.10 The two versions of a four-tone repeating sequence composed of two low-pitch and two high-pitch tones are shown for two cycles. The order for (A) is 400 Hz, 1400 Hz, 600 Hz, 1600 Hz and the order for (B) is 400 Hz, 1600 Hz, 600 Hz, 1400 Hz. The three versions of a six-tone repeating sequence composed of three low-pitch and three high-pitch tones are (C) 400 Hz, 1400 Hz, 600 Hz, 1600 Hz, 900 Hz, 1900 Hz; (D) 400 Hz, 1600 Hz, 600 Hz, 1900 Hz, 900 Hz, 1400 Hz; (E) 400 Hz, 1900 Hz, 600 Hz, 1400 Hz, 900 Hz, 1600 Hz

Sound Files 2.10: Four and six note interleaved sequences depicted in Fig. 2.10A–E

separation in frequency, intensity, timbre, or spatial position. (By the way, spatial position is only a weak cause of stream separation) (Fig. 2.10).

We can think of the sequence of sounds metaphorically, as creating an imaginary contour connecting one note to the next. As the presentation rate or frequency separation increases, the contour gets sharper and jagged. At some point, the glide between the different frequencies becomes so rapid that the auditory system cannot track it, and the two streams emerge. I imagine this to be similar to relatability discussed above. If the imaginary line connecting the two regions across the occlusion requires too sharp a curve (Fig. 2.11B), the two regions are not perceived as continuous. If the low and high frequency are linked by an actual frequency glide (Fig. 2.11C), then the tendency to form separate streams is radically reduced since the glide is a cue that the tones came from one source (by good continuation). But, if the glide is broken, that weakens the one source percept, and streaming reoccurs (Fig. 2.11D).

We can generalize the concept of a sound contour to other situations. While a visual contour connects different spatial parts of an object together, the sound contour connects different parts of the temporal sequence together to create one

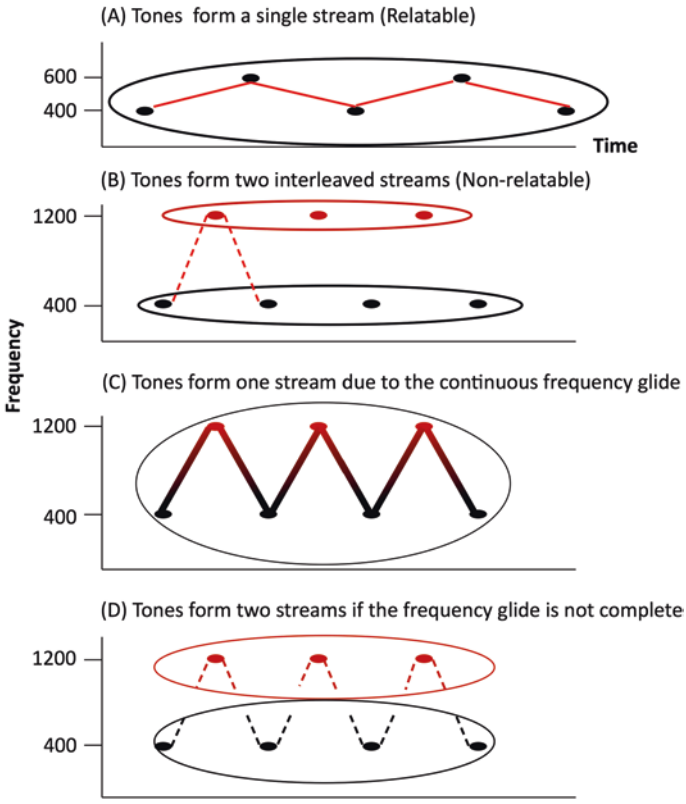


Fig. 2.11 (A) If the contour connecting the alternating tones is flat (i.e., relatable depicted by the solid red lines), the tones form one stream. (B) If the contour is sharp (i.e., non-relatable), the tones form two independent streams. (C) A frequency glide connecting the tones brings about one stream, but if the glide is interrupted, two streams reoccur (D)

Sound Files 2.11: The effect of complete and incomplete frequency glides on streaming as illustrated in Fig. 2.11A–D

source. Due to the Doppler Effect (which arises from the motion of the source relative to the outgoing sound waves), as a sound moves toward or away from the listener both frequency and intensity change. As the sound source moves directly toward the listener, the source moves with the sound wave and compresses it so that the frequency increases (i.e., the wavelength gets shorter). As the source moves away, the source moves away from the sound wave that is travelling back to the listener so that the frequency decreases (i.e., the wavelength gets longer).

This effect is more complicated if the source passes in front because the degree of frequency shift is a function of the change in distance between the source and listener. If the sound is approaching the listener, the frequency increase is greatest when the source is furthest away but diminishes as the source passes in front of the listener because the rate of change of distance is reduced. As the source passes directly in front, the frequency will equal its true value, and then the frequency will begin to decrease in progressive degrees as it moves further away.

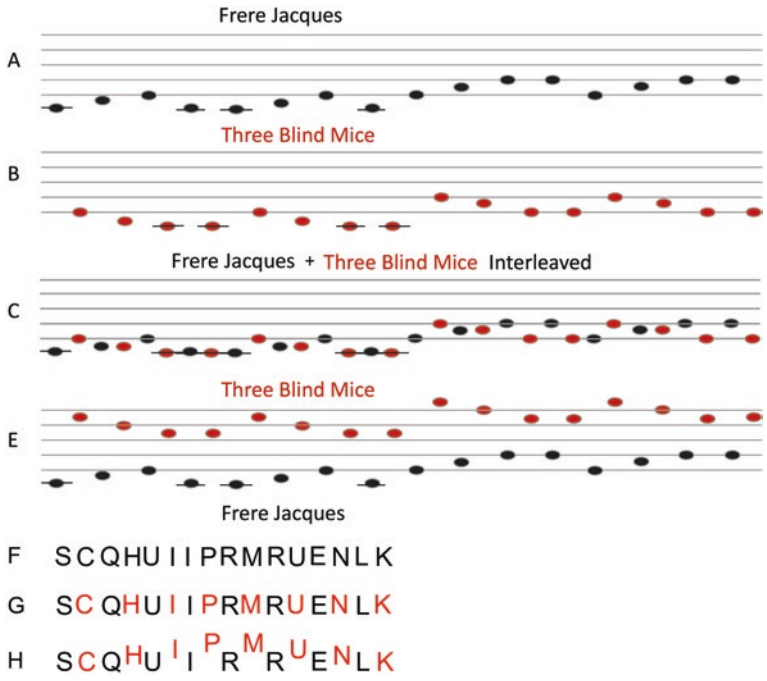


Fig. 2.12 "Frère Jacques" and "Three Blind Mice" are illustrated in (A) and (B). In (C), they are interleaved so that the contour has many simple repetitions and it is nearly impossible to pick out the two tunes. In (E), the notes of one tune ("Three Blind Mice" in red) are shifted by an octave; the two melodies split apart and both are easy to recognize. If two words are interleaved, it is also quite difficult to recognize each word (F). The identical color and shape and linear arrangement of the letters (e.g., proximity) inhibits isolating each word. Coloring one word, analogous to changing pitch, makes recognition easier due to Gestalt similarity (G), and changing the contour makes the two words pop out (H)

Sound Files 2.12: Interleaved and octave separated tunes corresponding in Fig. 2.12A–E

The smooth contour of the frequency change therefore is a clue to the coherence of a moving sound. Discontinuous shifts in frequency would suggest two different sources. A somewhat parallel visual effect is that of *looming*. If an object gets larger over time, the perception is that of approach, not that the object changes in size. An object is assumed to be rigid so that a smooth change in the size contour appears to be approach. It is very difficult to see it as a size change.

A second example of the conflict between the formation of a single stream (i.e., one overall contour) and the formation of two streams occurs for interleaved melodies. Consider the situation in which listeners try to identify two familiar tunes "Frère Jacques" and "Three Blind Mice" when the notes of each tune are played alternately. When the notes of each song are in the same pitch range, it is very difficult to do so; the notes form one continuous but incomprehensible contour (Fig. 2.12C). Only after the notes in one tune are shifted

in pitch so that there is little or no overlap can the two tunes be identified. When notes are in the same pitch range, the tunes can be identified if each tune is played in a different timbre, say one song by a clarinet and the other song by a violin, an example of grouping by similarity.

Contours also may signify musical structure and linguistic meanings, though there is a fundamental difference between the contours of music and speech. Music is built around a stable set of pitch intervals; these vary from culture to culture, but every musical system is based on such a structure. Such tonal sequences are sets of notes that have an internal cognitive structure and with one central note, the tonic, at the center to which other notes seem to lead back. Notes and intervals create an aesthetically and emotionally pleasing pattern that is integrated with the rhythm of the piece.

Speech, in contrast is built up from a continuous set of pitches that vary throughout the utterance. In fact, most speech sounds, with the exception of vowels, are made up of upward or downward frequency glides, or contain noisy parts like the beginning of the fricative "f" sound. Speech intonations, for the most part, do not have an aesthetic purpose (although the way words are said can surely affect their meaning to the listener). They exist to emphasize the meaning of the utterance, to describe "who did what to whom."

We will start with speech intonation. In a typical utterance, the pitch contour wends its way up and down in a roughly smooth shape ensuring that to the listener there is but one speaker. There are many different pitch contours, and it is often unclear how many of them have significance and how many merely reflect individual differences. However, there are some consistencies. For example, the pitch, the pitch variation (i.e., range between high and low pitches), and intensity of the sound decrease at the end, probably on account of the physiological consequences of running out of air so that less air is forced through the vocal cords. It seems that listeners expect this drop and therefore judge the beginning and end of a sentence as being equal in frequency and loudness.

Intonation contours can serve several functions. A rise in pitch, creating an accent, can be used to identify the subject in an ambiguous phrase ("THEY are flying planes" versus "they are FLYING PLANES"), emphasize a particular word in a phrase (the WHITE house as opposed to the white HOUSE), mark the end of a phrase in a sentence, particularly in French, and indicate a question by a pitch rise at the end of an utterance (who CALLED?).

As described above, music is constructed out a set of notes with fixed frequencies that create a hierarchic structure with privileged notes and intervals. Note sequences tend to follow simple expectations: (a) adjacent notes tend to be close in pitch so that the last note is a good predictor of the following note, another example of the Gestalt good continuation principle; (b) after a series of small steps in one direction, the next step is likely to be in the same direction; (c) but after a large interval either up or down, the next interval will reverse the contour direction and will be roughly equal in size), analogous to the symmetry principle (Schellenberg, Adachi, Purdy, & McKinnon, 2002). These three

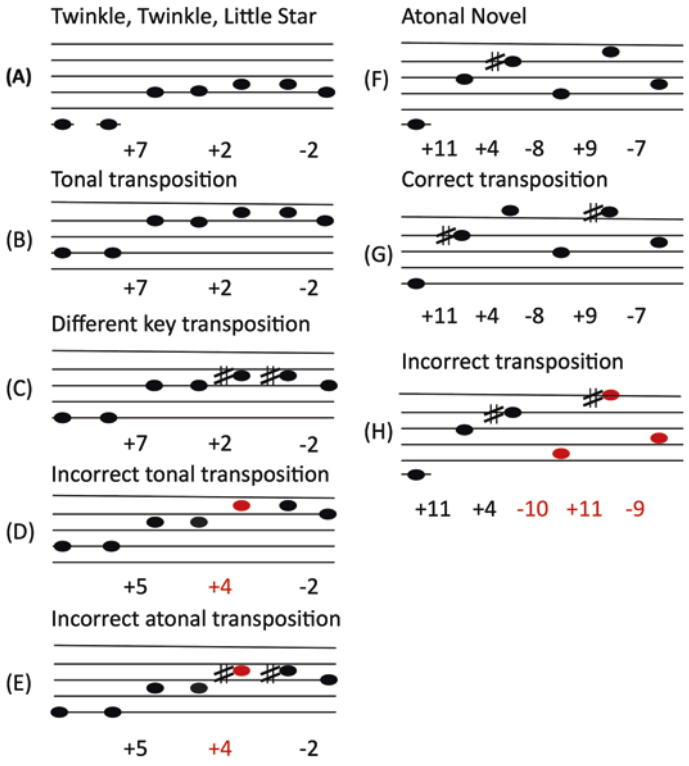


Fig. 2.13 The "target" melodies are (A) and (F). Listed beneath these short melodies are the numbers of semitone steps between the two surrounding notes. For "Twinkle, Twinkle, Little Star" (A), the correct transpositions (B) and (C) have the identical number of steps between notes. The incorrect transpositions (D) and (E), although maintaining the same contour, have different-sized steps between notes. The same is true for the atonal melody (F). The correct transposition (G) maintains the step sizes, but the incorrect transposition (H) does not

Sound Files 2.13: Tonal and atonal transpositions corresponding to Fig. 2.13A–H

principles may be understood as a way of maintaining the continuity of one melodic line, avoiding a split into competing streams.

Typical of classical and folk music is the repetition of melodic themes starting at different notes, different interval sizes, and different key signatures. In these cases the pitch contours are identical, and the use of the different versions of the same theme acts to tie the piece of music into a whole. A body of research has asked whether listeners can distinguish transpositions in which the intervals are exactly alike (Fig. 2.13B, C & G), from transpositions in which the contour is identical although the interval sizes are changed, or from transpositions in which the contour has changed say from *up, up, up, down, up* to *up, up, down, up, up* (Fig. 2.13D, E, & H).

In the typical experiment, a series of notes is first presented as the standard, then by a short delay, followed by the comparison series of notes. In some

experiments, the task is to decide whether the standard and its comparison are identical; but in the more relevant cases, the listener must decide if the comparison is an exact transposition of the standard. The results are complicated because the outcomes depend on the melodic sequence, the presentation rate of the sequences, and the age and musical experience of the listener. Halpern and Bartlett (2010) provide an extensive review of this work.

If the standard and comparison start at the same note, regardless of whether it is a familiar or novel sequence, it is relatively easy to judge if the comparison is a true transposition. If the sequences are played with different notes, the judgments are more difficult. For atonal novel sequences composed of notes from more than one scale, when the comparison is transposed and the one note changed does not affect the shape of the contour, then both experienced and inexperienced listeners cannot distinguish true transpositions from the comparison lures. If the changed note does alter the contour, then the task is easy. For novel tonal sequences, if the changed note in the comparison does not alter the contour, but does not occur in the original key (a black piano key when all other notes are white), then the judgment is relatively easy. The black key just does not fit. For familiar tonal sequences (e.g., "Twinkle, Twinkle Little Star"), the discrimination is easy for all participants.

It has been suggested recently that dolphins use the characteristic frequency contour of each individual dolphin's song for recognition. The contours range from simple increasing frequency glides, to double U-shaped ("W's") ups and downs. These contours become unique to each one as the dolphins mature (Kershenbaum, Sayigh, & Janik, 2013).

To sum up, the basic mode of organization for tunes is simply the shape of the up-down contour. Only direction counts so that there are several correct possible notes at each point in the tune. It is the type of organization first discriminated by infants and small children. Musical experience can lead to another level of organization based on the sizes of the intervals, but only if the notes fit into an existing musical scale. For such tunes, the interval size organization restricts the next note to a single one. Otherwise, contour organization is still dominant.

2.3.2.2 Occlusion of Overlapping Sounds

We can speculate how the auditory figure/ground principles may be analogous to the proposed progression for the visual system. In seeing, the fundamental step is to identify clumps of connected points that have the same motion, and those points consequently come to represent solid objects. By analogy, while visual objects have connected surfaces, sequences of tones with related frequency components act like visual surfaces. Most sounds are harmonic, so there is a fundamental component at the lowest frequency (F_0) and the frequency of the other harmonic components would be integer multiples. Usually, the pitch of a complex tone is based on the frequency of the fundamental, and the quality of the sound, termed *timbre*, is based on the number and relative strengths of the harmonics. Technically, timbre is defined as the sound quality at one frequency and intensity, but it seems more natural to think of timbre as

belonging to one object (i.e., a clarinet) regardless of frequency or intensity. (This will be discussed further in Chap. 5). Therefore, based on harmonicity, if the components (in Hz) of an ambiguous sound reaching the ear were 100, 130, 200, 260, 300,390, 400, 520, and so on, it would be likely that there was one source with an F_0 = 100 Hz and harmonics of 200, 300, 400, and so on, and a second source with an F_0 = 130 Hz and harmonics of 260, 390, and 520.

Again, by analogy, for the components of a sound to have the "same motion," they would have to start and end at the same time or have the same oscillation in amplitude and/or frequency. Temporal synchrony has been found to be the property that most fundamentally signals which frequency components go with each source, since frequency components from one source invariably have an identical temporal pattern; all the components start at the same time and decay at the same time. It would be highly improbable that sounds from two different sources would start at the same time, so that in a complex sound, grouping those components with the identical onsets would be a useful heuristic to isolate each different source. Work by (Elhilali, Micheyl, Oxenham, & Shamma, 2009) suggests that temporal synchrony is such a dominant cue that two widely separated synchronous frequencies are heard as a single sound even though each frequency may generate nerve impulses along separate neural tracks. It is the resulting synchronous, correlated firing of those neural tracts that limit the perception to one sound. If those tracks fire at differing times or rhythms, multiple sounds will be heard (Fig. 2.14).

To review, the acoustic wave that reaches the ear is the sum of the sound waves from each source that in turn is simultaneously the sum of a set of individual frequency components. The origin of any part of the wave is lost at the ear in the composite. A fundamental step would be to split apart the frequency components coming from each source. This division would be based on the core concepts of temporal synchrony and harmonicity common to speech, music, and environmental sounds like wind, impacts, or air conditioning.

Online Resources: YouTube: The section "Musical Acoustics and Sound Perception has many videos pertaining to sound. I would suggest starting with these two:

"Musical acoustics and sound perception" by Tiku Majunder at Williams College

"UNSW Physics Public Talk: The physics of music and voice" with Prof. Joe Wolfe

We can consider the organization of songs at two time spans. For a single sound, temporal synchrony and harmonic relationships tie frequencies together into one sound source, while asynchronous onsets in particular, and non-harmonic relationships among the component frequencies, tend to create the perception of two or more sources. For sequences, all the sounds can be heard as coming from one source organized into one contour or can be heard as coming from different sources and perceptually segregated into two or more

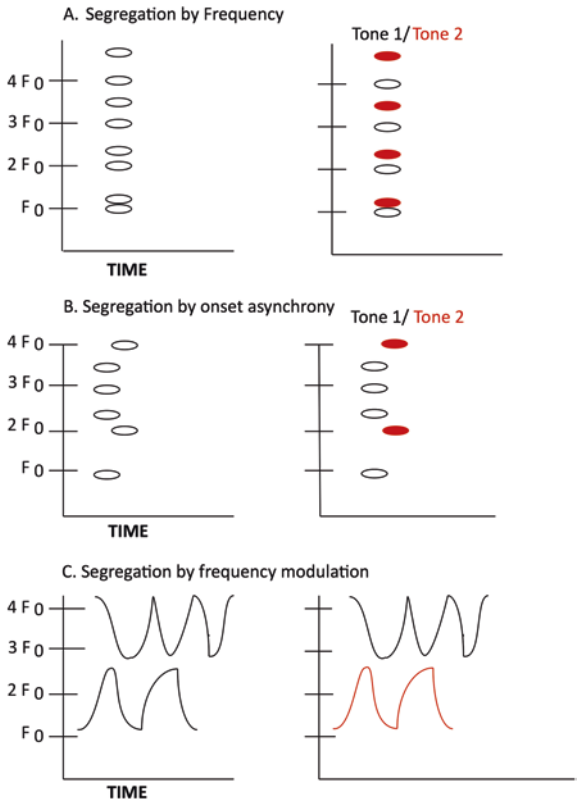


Fig. 2.14 In all three panels, the left side presents the auditory sound as a function of time along the horizontal axis, and frequency along the vertical axis. The right side represents the segregation into two parts. In (A), segregation is due to the harmonic relationships among the frequency components. In each sound, the frequency components are simple multiples of the fundamental. In (B), the segregation is due to onset asynchrony. Here, the asynchrony dominates so that the harmonic relationships are violated. In (C), although not discussed in the text, segregation is due to a different pattern of frequency modulation and amplitude modulation that may be caused musically by deliberate use of vibrato or by inadvertent changes in bowing or breathing

Sound Files 2.14: Segregation due to harmonic differences, synchrony, and temporal modulation shown in Fig. 2.14A–C

disjoint but overlapping contours. For both individual sounds and sequences, there is a constant competition between fusion into one source and segregation into distinct sources. The auditory system works to maintain a consistent representation of the environment when no new information suggests that reanalyzing the scene is necessary.

In the same way that one rigid object can occlude another, one sound may mask, occlude, or "drown out" another. In most experiments, a segment of noise is used to mask a single tone or part of a musical passage. In general, the term noise refers to a non-periodic sound so that at any time point each

frequency has an equal probability of occurring. There are several noise variants, termed white, pink, and brown, in which the power at different frequencies creates sounds that will vary from the hissing sound of white noise to the deeper waterfall sound of brown noise. Here we will simply term the masker "noise" without concern for the particular type used.

We start with the simplest case, a tone composed of only a single frequency and a noise burst whose frequency range includes that of the tone (e.g., a tone of 400 Hz and a noise spanning 200–2000 Hz). The noise needs to be louder than the tone. Eight possibilities are shown in Fig. 2.15. In (A), two tones are presented with a short silent gap between the tones so that two separate tones are heard. In (B), a noise burst is placed in the gap with the result that the tone appears to continue through the noise, the continuity illusion. The illusion is strongest if the gap if less than 300 msec; for longer gaps, the illusion is weaker and fades so that it appears as if the tone is partially on. The tone captures its frequency in the noise burst so that the noise seems to change its timbre slightly. For continuous presentation, the noise seems to form one stream and the tone a second stream. In (C), the tone ends before the onset of the noise and resumes after the end of the noise. Whatever the frequency of the noise, the tone does not seem to continue through it. If the frequency range of the noise burst does not overlap the frequency of the tone (D), then the tone is correctly perceived to consist of two segments. Suppose the tone is a single continuous glide or two glide segments. In (E), the glide is separated by a silent gap, and two separate segments are heard. In (F), the glide begins before the noise, stops at the noise, and appears to follow the same trajectory after the noise ends; the perception here is that the glide continues through the noise. In (G), the glide appears to be continuous even if it reverses direction. In (H), however, because the glide going into the noise cannot easily match the glide starting at the offset of the burst, it does not seem to continue through the noise. While it is perilous to make use of visual depictions of auditory phenomenon, these outcomes match those predicted by the concept of visual relatability used above to explain if two visual segments separated by an occlusion are perceived to connect. Initially, the auditory and visual sensations are detected, then auditory and visual contours act to create coherent objects, and finally those objects are matched across time and space.

If the tone ends before the onset of the noise and restarts after its offset, the tone is not perceived to continue through the noise (Fig. 2.15C); listeners hear two tones. The identical outcome occurs visually. If a motion display shows an object moving toward but stopping before an occluding barrier, reappearing beyond the barrier and continuing to move at the same speed, the object is not seen to move behind the barrier; there are two objects (analogous to Sound File 2.15C). But if the object moves against the barrier, disappears, and then reappears next to the barrier and continues to move at the same speed, it does appear to move behind the barrier (analogous to Sound File 2.15B). The gap in the former case breaks the contour and the perception of a single tone or object does not occur.

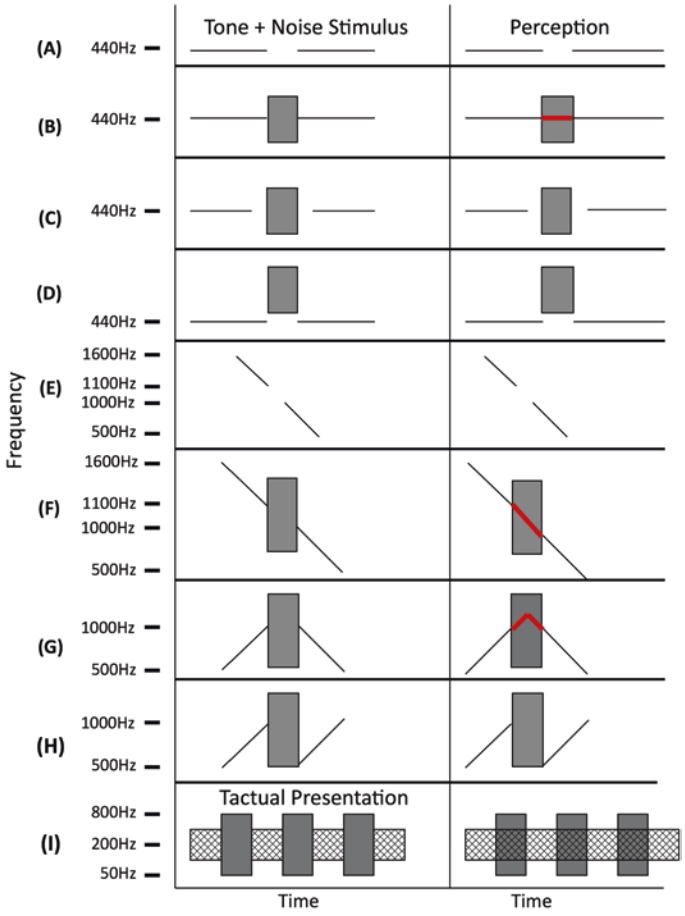


Fig. 2.15 Eight configurations are shown to illustrate the masking of a tone by a noise burst. A split tone is presented in (A) and a split glide is presented in (E). In both cases, the perception is veridical; two separate tones or glides are heard. In (B) and (F), a noise burst is inserted in the gap and in both cases the sound is perceived to continue through the noise burst. The "illusionary" segments are shown in red on the right. In (C), if there is a silent interval between the offset of the tone and the onset of the noise, the tone does not seem to continue through the noise. In (D), the noise burst is high-pass filtered so that it does match the frequency of the tone and the tone does not appear continuous. In (G), the continuity of a glide in noise occurs even if the glide reverses in direction if the two glides are "relatable." But in (H), if the glides would not connect (not relatable), the noise has no effect. Two separate glides are heard separated by the noise burst. In (I), if a 200 Hz tactual vibration is alternated with a 50–800 Hz vibration, the 200 Hz vibration is perceived to continue through the alternation (discussed below). The tactual outcome matches that for a tone and noise masker (B)

Sound Files 2.15: Continuity across silences within tones and glides due to interpolated noise (Fig. 2.15A–H)

In everyday life, sounds compete and overlap. Spoken conversations are masked or overwritten by a variety of louder sounds and yet we rarely realize that the utterance was interrupted and parts were not heard. If the surrounding segments are coherent and appear to come from the same source, listeners pierce the noise and perceive parts of the utterance that were masked by the louder sound or even did not occur at all. This has been termed *phonemic restoration*.

Phonemic restoration, occurring for complex sounds, is another example of the continuity illusion found for simple tones as in Fig. 2.15B, E and F in which the restored sound connects the segments, or "turns" to connect opposing glides. Furthermore, phonemic restoration can illustrate how context and the listener's prior knowledge affect how the masked sound is understood. Warren (1999) and colleagues first constructed an ambiguous phrase in which the initial sound of the last word was masked: #eel. They then created three sentences beginning with that phrase in which only the final word differed:

It was found that the #eel 1. was on the orange
or 2. was on the table
or 3. was on the axle

Listeners reported hearing peel, meal, or wheel depending on the final word. The context determined the illusionary percept, and eliminating the meaning of the sentence by playing it backward reduced the degree of restoration.

Sound Files 2.16: Phonemic restoration due to interpolated noise

2.3.3 Haptic Perception

2.3.3.1 Segregation into Overlapping Surfaces

If we place extended fingers on a surface or object, we perceive a continuous surface in spite of the point-like stimulation from each finger. Katz uses the example of a hairbrush; the sensations are the bristles, yet the perception is that of a continuous surface. If the surface is curved, then we perceive a concave or convex surface, not the position (e.g., height) of each finger (Kappers & Bergmann-Tiest, 2015). In similar fashion, as we manipulate an object using one or two hands we perceive a solid in spite of the serial sensations from spatially discrete regions. This is similar to the filling in of the visual scene across the blind spot of the eye. (At the point at the back of the eye where the optic nerve exits, the eye is functionally blind because there are no cells sensitive to light at that point. Normally we do not see a hole in the visual scene because the visual system fills in the hole based on the sensations on either side of the blind spot). Why does filling-in occur? As described above, the initial assumption of visual and auditory organization is that there is one object or source, and the filling-in of the tactual surfaces seems to be based on the same assumption. The concept of uniform connectedness was suggested by Palmer and Rock (1994) to explain visual grouping, but it seems plausible to reverse the concept and argue that perceptual systems assume uniform connectedness in the environment.

Our expectation is that the figure-ground organization for touch should resemble that for seeing and hearing. As is true for visual objects and auditory sources, physical surfaces and material objects rarely occur in isolation. They abut and lie above one another spatially, and sensations can overlap in time. We have used the concept of relatability to understand the perception of continuity when visual objects occlude and overlap, and a similar principle would seem to apply to tactual surfaces and objects. Chang, Nesbitt, and Wilkins (2007a) created simple visual stimuli such that parts occluded each other and simple tactual surfaces in which two pieces of different roughnesses lay on top of each other that resemble the examples for relatability shown in Fig. 2.5. The results showed that participants connected the occluded parts in both the visual and tactual stimuli in the same way.

2.3.3.2 Perception of Surface Contour

To investigate the perception of contour in haptic displays, Overvliet, Krampe, and Wageman (2013) created raised dot patterns. In half the patterns, a subset of the dots formed a circular array amidst a background of interspersed randomly placed dots, while in the other half of the patterns all the dots were placed randomly. The participant's task was simply to determine if a circular array was present and they did this by scanning the display by zigzagging across it with either one finger or one hand (this difference did not affect discrimination). The participants could locate the target because the raised dots forming the circle were always closer to each other than the interspersed random dots.

A visual representation of the experimental conditions is shown in Fig. 2.17. For eight of the 10 conditions, participants were nearly perfect in detecting the circle. Performance was slightly poorer in the 5.5/11 mm conditions (5.5 mm refers to the spacing of the dots in the circular array, and 11 mm refers to the spacing among the masking dots) while performance was below chance (50%) for the 5.5/7 mm conditions. Although a distance of 5 mm is the maximum separation that can be perceived by the sensitive pad of a finger, discrimination was still excellent for the 5.5/11 mm conditions. The below chance performance for the 5.5/7 mm condition suggests that it is the similarity in the spacing between the target and interspersed dots that interferes with the detection. As described above, visual and auditory grouping according to the Gestalt principles is often in conflict. This is also true for tactile grouping. If the color or textures in the Chang et al. (2007b) experiments were made more similar, then proximity organization should gain prominence and vice versa. In similar fashion, if the spacing among the dots in the background becomes more equal to the spacing among the dots in the circular target, discrimination will decrease and vice versa. This outcome seems likely given the simple examples in Fig. 2.17A, B, and C.

In another experiment, subjects were required to search for a patch of rougher or smoother paper randomly placed along a fixed strip (Van Aarsen and Overvliet, 2016). Participants were faster and made fewer errors when the difference in roughness between the target patch and the background strip was

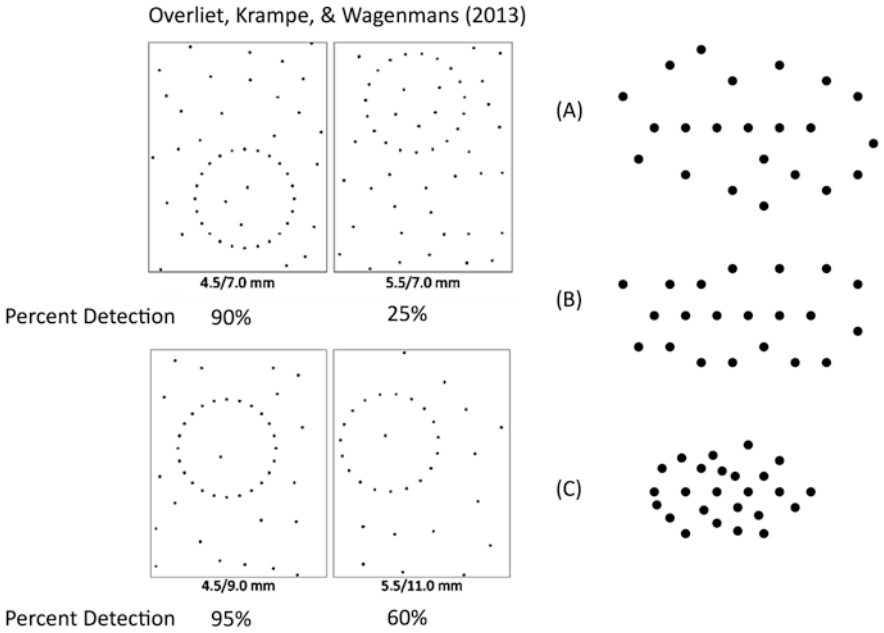


Fig. 2.17 Detection of raised contours. If the target dots are spaced closely, the circle target is easy to detect regardless of the spacing of the masking dots. But, if the target dots are spaced further apart (5.5 mm), then the spacing of the masking dots can completely obscure the target. It is relatively easy to hide a straight-line target by matching the spacing of the target dots with extra dots (A, B, & C). (Adapted from Overvliet et al., 2013)

greater. Although this research still does not demonstrate balancing among tactile properties, it does show that bigger differences in magnitudes affect the ease of tactual discriminations.

Embedded figures provide another way to investigate figure-ground organization. In the original embedded figure test, participants were presented a simple visual figure made up of straight lines, and then had to trace that figure hidden in a more complex figure made up of the original plus extra background lines that act to change the overall shape and organization of the combination. Heller, Wilson, Steffen, Yoneyama, and Brackett (2003) adapted the test to tactile perception by the use of raised lines and compared the ability to detect the embedded figures among congenitally blind, late-blind, very low vision, and blindfolded sighted subjects. The target figures were embedded either in simple and complex backgrounds. Two examples of target figures and their simple and complex backgrounds are shown in Fig. 2.18.

The number of correct responses and mean response times for the simple and complex backgrounds for each group also is shown in Fig. 2.18. There was no difference in the performance of the late-blind and the very low-vision subjects. Although the results were complex, one clear outcome was that the congenitally

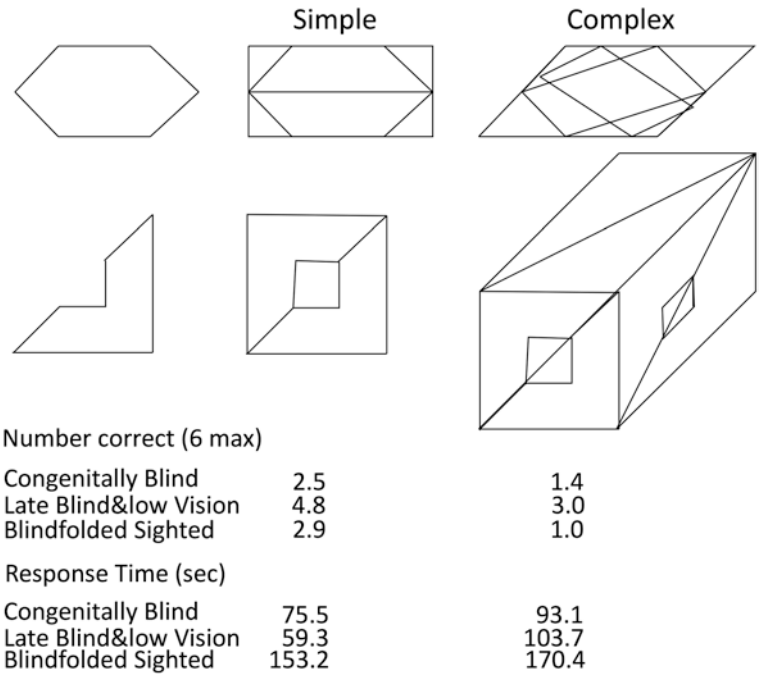


Fig. 2.18 Examples of simple and complex embedding of two geometric figures from Heller et al. (2003). The data are the averages across all the figures

blind subjects' performance equaled that of the blindfolded subjects. In fact, the congenitally blind subjects were equally accurate and twice as fast as the blindfolded sighted subjects. This is a clear suggestion that visual experience is not necessary for figure-ground organization. There has been a long-standing controversy whether haptic perception needs prior visual experience to perceive shape and this work shows that such experience is not necessary.

The congenitally blind subjects had more experience using touch to perceive patterns. In contrast to blindfolded sighted subjects who used one hand to explore both the target figure and embedded figures, the visually impaired subjects explored the target and embedded figures with two hands, a more rapid strategy. Moreover, the blindfolded subjects usually explored the perimeter of the raised figure and that made it extremely difficult to discriminate the target in A. The low-level-vision groups explored the interior of the embedded figures more extensively and that led to improved discrimination.

We normally explore solid objects sitting or embedded on surfaces, for example, a pencil on top of a cluttered desk surface, so that the tactual exploration of raised dots or line figures does not represent normal tactual perception. To compensate, Pawluk, Kitada, Abramowicz, Hamilton, and Lederman (2011) investigated the figure-ground segmentation of three-dimensional solid objects located of top of flat or indented surfaces. Three properties of the

shapes were varied: (a) size (small or large width and height) and shape (vertically or horizontally oriented); (b) movability (rigid or wobbly), and (c) texture (object and supporting surface have the same surface roughness or the object is rough and surface smooth). The participant's task was to briefly lower their open hand toward the surface and detect whether they had touched one of the solid objects or the supporting surface. The brief touch could act to create slight movements if the object was able to wobble, but in general gives only a coarse estimate of the object's properties.

All three variables affected the judgment of whether the participants had touched the object or the surface. Taller, wobbly, and rough sensations were more likely to be judged as representing objects sitting on top of the surface. Given the experimental procedure, it was impossible to judge the relative importance of the three properties. But, given that perceptual judgments occur in a context, it seems that the importance of any of the properties could be varied by changing the values of each.

2.3.3.3 Occlusion of Overlapping Vibrations

All of the above research involved static stimuli, but parallels between seeing, hearing, and touching also occur for temporal sensations. The temporal continuity illusion illustrated in Fig. 2.15 for sounds (also found for visual displays) also occurs for vibrations on the skin, illustrated in Fig. 2.15I. Kitagawa, Igarashi, and Kashino (2009) created a 200 Hz target vibration and a 50 Hz–800 Hz masking vibration that were both placed on the pad of the forefinger. To test for the perception of continuity, the target and mask vibrations were alternated five times. The participants judged whether the target appeared to be continuous or not. The target vibration seemed continuous as long as it was weaker than the mask vibration, and the illusion persisted even if the mask was 500 msec long. To put it another way, the target seemed to continue so long as the mask could have actually masked (or occluded) the target vibration. The participants could easily differentiate between the target + mask and the mask so that the continuity was not due to the inability to discriminate between the target + mask and the mask. It is difficult to compare the continuity illusion in touch with that in hearing because the stimulus conditions were so different. Auditory experiments usually present only one noise masker, not the five maskers used here. Even so, the outcomes are comparable.

2.3.4 Temporal/Spatial Coherence

Most of the discussion up to this point has concerned simple, relatively static visual and auditory objects. While auditory events constantly change over time, even for waterfalls, vacuum cleaners, and wind, we tend to think of the visual world as stationary, even though objects move, new objects come into view to block old ones, and "limbed" objects move in coordinated ways (this will be discussed later in this chapter). But even these examples involve the motion of rigid solid objects. In this section we want to consider visual organization based

on purely temporal coherence. The flashing of fireflies is an example. Imagine a species in which the flashes occur randomly. Is it possible to identify a set of flies that move in the same direction while the remaining flies move in random directions? This problem would be even more difficult if the flies did not flash consistently; effectively each one goes dark at different times.

Below we will consider three cases. In the first, a block of texture moves, in the second individual elements move much like the fireflies, and in the third the individual elements do not move but switch direction in corresponding ways.

2.3.4.1 Rigid Arrays

Typically, a random pattern composed of half black and half white squares is constructed in a square array, say 24 × 24. A second pattern modified in some way is then alternated with the original pattern remaining at the same spatial position. There are several modifications possible, but in each one the change occurs in a small block of connected squares within the array.

a) The cells in a small region of the large array (e.g., 16 squares in a 4 × 4 region) are reversed by chance at each alternation. The probability that the black and white cells reverse would be 0.50, so any cell in the small region would have an equal probability of switching or remaining constant. In this case the small region is perceived to flicker or glitter (See Fig. 2.19A)
b) To create the second array, a group of squares in a small region in the first array is shifted by a certain amount to overwrite the original cells. This leaves a hole in the second array since the original cells have now been shifted and that hole is randomly filled with black and white cells. If we compare the two arrays, all of the cells are identical except for that small block that occurs at different positions in the two arrays. The block seems to shift back and forth as the two arrays alternate or to float back and forth on top of the larger array. The visual system must be comparing the arrays globally since the small shifted region blends into and disappears in either of the two arrays when presented alone. But any global comparison will likely come up with several possible but incorrect matches between the two arrays. Remember that the alternating arrays are nearly identical; only a small block of cells changes. The key to perceiving those segments as a single moving region is to isolate connected cells that undergo the identical change. The spatial organization of the elements strongly affects the detection of the temporal grouping (Nishida, 2011) (See Fig. 2.19B)

That a rigid block of random texture appears to move explains the collapse of camouflage due to movement. If a hunter wearing camouflage clothing remains still, the clothing will enable the hunter to blend into the foliage. However, any twitch or shake will make the rigid pattern of the clothing move relative to the background and make the hunter visible. This is sometimes termed “shearing” the texture. Invariably, when a hunter is inadvertently shot by a partner, it occurs when the hunter is still and the partner does not remember the hunter’s position.

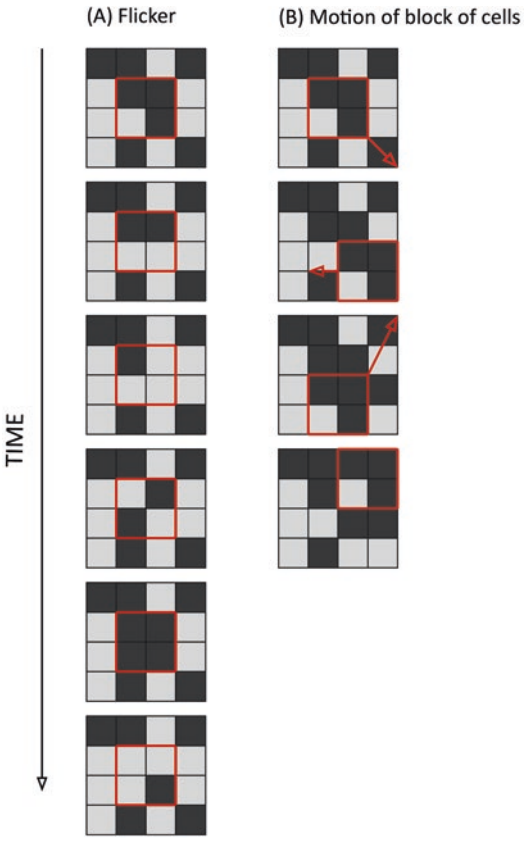


Fig. 2.19 (A) If the cells in a fixed region randomly shift brightness, that region appears to flicker. (B) If the cells in a fixed region shift position, those cells appear to move and float above the background cells

2.3.4.2 Nonrigid Arrays

Here, a random set of dots in the initial frame move to new positions in successive frames. Each dot could move in a random fashion, or a subset of the dots could move in a constrained direction or distance in successive frames. For example, 5% of the dots could move either downward to the right or upward to the right while the remaining dots move randomly; the subject's task would be to identify the direction of the motion (See Fig. 2.20).

This is a much more complex task than that in the rigid arrays. In most instances, the dots that move consistently are scattered across the array instead of being clustered into a single region. Each dot can move a different distance on each step, different dots can be used to portray the movement on successive steps, and any single dot can move in one direction only for a limited number of steps. All of these restrictions prevent an observer from tracking a single dot to determine its direction. To solve the direction problem, the observer must

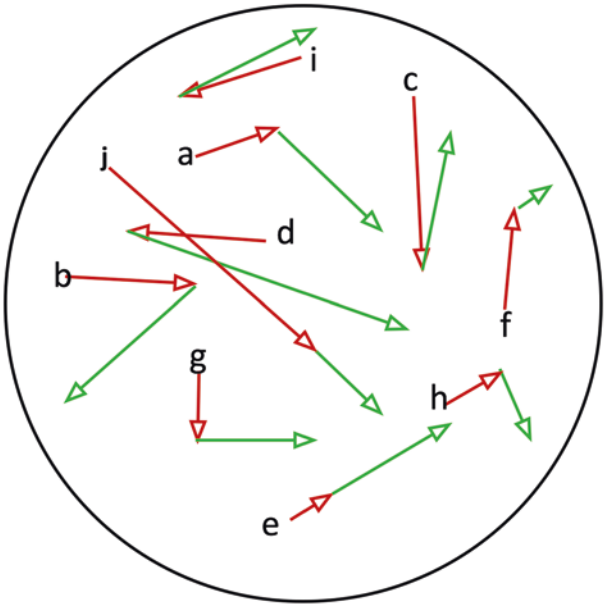


Fig. 2.20 The motions of 10 points are shown. The first step is drawn in red and the second in green. Only two points (e & j) move in the same direction on both steps. The common motion, although carried by different points (c,e,f,i), is up to the right

integrate the changing motions of the dots across the entire perceptual field. In spite of these difficulties, the perception of coherent motion occurs even if only a small percentage, roughly 3% to 5%, of the dots are moving in one direction (Braddick, 1995). The motions in local regions are integrated first, and then integrated into global regions (Watanabe & Kikuchi, 2006). The dots moving in one direction create the perception of a surface, which allows us to perceive the direction.

2.3.4.3 Temporal Synchrony

2.3.4.3.1 Visual Scenes

For both rigid and nonrigid arrays described above, it is the apparent or real movement of parts of the array that creates the perception of a surface upon which the movement occurs. Additional research has demonstrated that it is possible to create the perception of surfaces purely by means of temporal synchrony, without the lateral movement that occurs in the cases above. For example, Sekuler and Bennett (2001) made use of a checkerboard pattern in which the squares were of different brightness. A small set of the squares of different brightness underwent simultaneous increases in brightness that contrasted from the simultaneous decreases in brightness in the remaining squares. This difference led to the grouping of the subset, without movement.

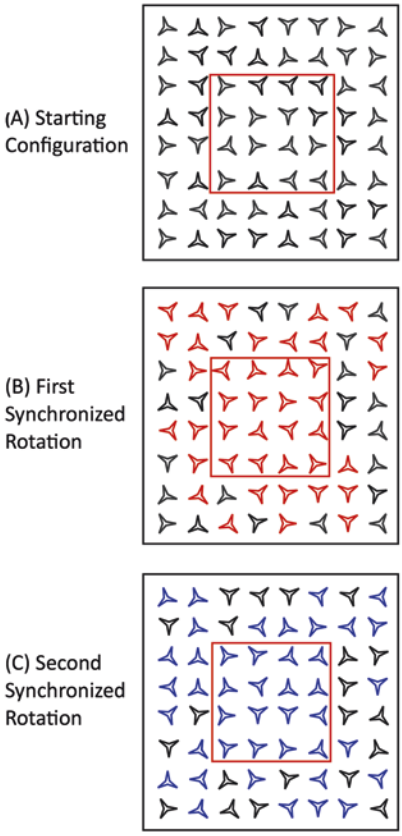


Fig. 2.21 In the starting configuration (A), the "windmills" are oriented randomly. In the first rotation (B), the windmills within the figure region rotate randomly in both direction and number of degrees (in red). Some of the windmills outside of the figure region also rotate, shown in red. In the second rotation (C), the windmills within the figure region continue to rotate, but a different group of windmills outside of the figure also rotate (in blue). Only the windmills within the figure region rotate (or do not rotate) in a correlated fashion

Lee and Blake (1999) demonstrated that the correlated timing of local motions can bring about the perception of segregated figural regions. The stimuli consisted of many little "windmills" at set positions that could rotate either clockwise, counterclockwise, or remain stationary as shown in Fig. 2.21. In the figural region, at every time point all the windmills either rotated or remained still. The direction of the rotations were independent; two adjacent windmills could both rotate in the same direction or could rotate in opposite directions. It was only the timing synchrony of the changes, not the rotation direction that defined the figure. This has been termed a *point process*, because only the timing of the changes matters, not the direction of the rotation. In the non-figural regions, at every

time point, the rotations occurred randomly so that some windmills would remain still while surrounding ones could rotate in either direction. By chance, there will be windmills in the non-figural region that will have the identical timing pattern as those in the figural region. The visual system probably disregards those windmills by restricting figural regions to connected elements only.

Rotation itself, however, does not lead to figure-ground organization. If we construct an array of needles that rotate around their central points so that the needles in one area rotate in one direction and the needles in the other areas rotate in the opposite direction, the two regions do not segregate. In contrast, if the needles in the two areas rotate at different speeds then segregation is easy (Watson and Humphreys, 1999).

2.3.4.3.2 Auditory Scenes

In our natural environment, sound sources start, overlap, change position, and undergo frequency shifts, and yet we are usually able to track each source successfully. While all of the ways we can do this are unknown, one critical property is the temporal synchrony and coherence of the frequency components of each source. In the work described above (Lee & Blake, 1999), the temporal coherence of the movement in a small region of a larger array leads to that region being perceived as a figure against the random movements of the rest of the array. In an analogous auditory demonstration, Teki, Chait, Kumar, Shamma, and Griffiths (2013) created what they term a "stochastic figure ground." A stable set of synchronous frequency components is embedded in a longer sequence of randomly varying frequency components. The synchronous set is understood to be a coherent object in a noisy environment. A simple example from their research is shown in Fig. 2.22. In (A), the figure consists of four frequency components and each component occurs for the middle three of the five sounds shown. Some of the other frequency components occur for each sound, but are inconsistent and found only for a minority of the four sounds. The figural components form a sound that "sticks out" from the randomly occurring components. In (B) by contrast, there is no consistent temporal pattern of the components so that the figure does not appear. It is as if each sound is made up of a random set of frequencies as shown in (C). The consistent frequency components play the same role as the rotating windmills; in both, synchrony, whether of frequencies or rotations, creates a figure.

All of these examples of grouping by synchrony can be thought of as demonstrations of the Gestalt principle of common fate. What this all shows is that the perceptual "rules" we use to understand the visual, auditory, and tactual world are quite similar even given the large differences in the kinds of stimuli. It is unlikely that two separate objects or sources would undergo the same changes in step. The physical properties of things create the same perceptual information whether they are visual objects, sound sources, or tactual vibrations or surfaces, but that information is understand in terms of the spatial or temporal context in which it occurred.

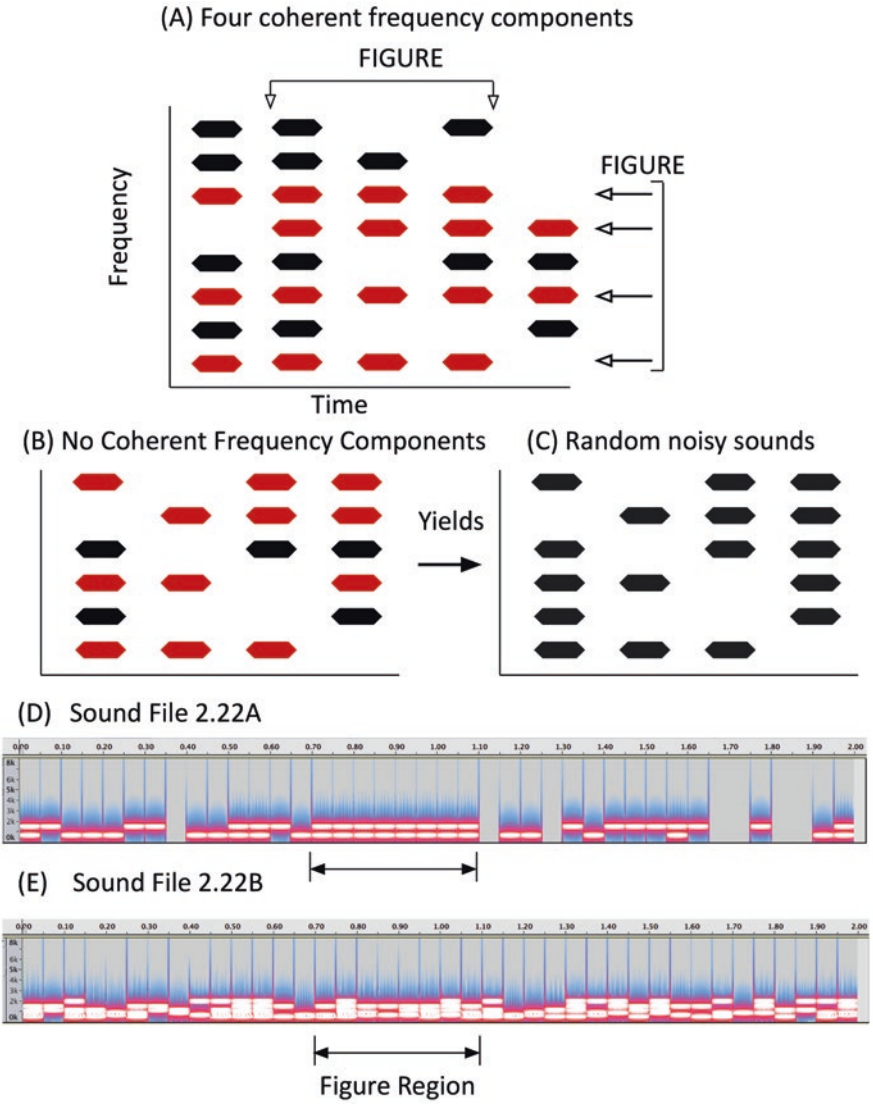


Fig. 2.22 In (A), the four coherent frequency components (in red) in the middle three sounds form a figure, that is, a sound that seems to occur in each of the three sounds in spite of the other overlapping components. In (B), even though each (red) component occurs in three of the four sounds, none are completely coherent and no figure is heard. In (C), it sounds like a series of random frequency components. In (D), two figure components are presented alone in a sequence of eight elements. In (E), the two figure components are presented along with four non-figure components. The identical figure region occurs

Sound Files 2.22: Stochastic figure ground stimuli pictured in Fig. 2.22D & E

2.3.5 Multisensory Integration and Organization

All of the above concerns binding or belongingness within a single modality. The sensations can be bound into a single object, surface, or source, or they can be segmented into different ones. We have argued that a broad interpretation of the Gestalt principles of organization can be applied in the same way to each sense. It seems intuitive that the same sort of principles will come up again when we consider how the sensations in more than one modality affect each other in creating one or more percepts. Our basic perceptual assumption is that visual and tactual sensations come from one object or surface and that sounds come from one source. Given those assumptions within each modality, then a similar assumption would be that sounds, lights, and tactual pressures or vibrations that occur together also come from one event or source. This has been termed the *unity* assumption. If sensations from different senses are perceived as coming from the same source (the roughness and scraping sound when rubbing sandpaper, or the tactual hardness, sound, and visually perceived indentation when tapping an object), then we might expect that the sensations will interact with each other. Given that the sensations from each sense are noisy and hard to discriminate, then combining the sensations from the different senses is likely to result in a better estimate of the object or source. In this case, we might expect that people would judge which source is more reliable or accurate, and more strongly attend to that source. It is important to note that the reliability of the information from each modality depends on the specific context. No modality is inherently more reliable. If the sensations are not perceived as coming from the same object for any reason, different spatial positions, onset timings, and so on, then it is unlikely that the sensations will affect each other and the estimates will remain the same. This is simply another instance of how binding works.

The organization of auditory, tactual, and visual sensations further illustrates the balance among the organizational principles. If we start with an auditory sequence of two tones that seems to stream into high and low sequences, we can reintegrate the overall sequence by reducing the presentation rate, and then split that into two streams by increasing the frequency ratio between the high and low tones, reintegrate it again by further slowing the sequence, and so on. We can expect the same balancing when combining stimuli from two or three modalities. Making the onsets synchronous or making the stimuli seem to occur in the same physical location will increase the binding of the cross-modality stimuli and the probability that those stimuli will affect each other. Conversely offsetting the onsets or making the stimuli appear to arise from different locations will reduce the probability that the sensations will interact.

Another aspect of cross-modality integration is the individual differences among people. At one extreme, individuals can understand all combinations of sensations as being bound to one object, and at the other extreme attend to only one of the sensory inputs and treat the other as irrelevant or even noise. It is important to keep in mind that integration is almost always a perceptual and

cognitive decision about how to treat the information about the world from each modality. We have to make these decisions constantly because events and sources nearly always result in sensations across modalities.

Cross-modal integration has been studied from many perspectives. We will start with research under the rubric of cross-modal correspondences, and then consider how within-modality and between-modality organization affect each other, and finish with examples of cross-modal integration in which the sensations in one modality affect the perception of the attributes of a second modality.

2.3.5.1 Cross-Modal Correspondences

2.3.5.1.1 Compatible and Incompatible Associations

Parise (2016), in an instructive review, defines such correspondences as systematic associations found across seemingly unrelated sensory features in different sensory modalities. For example, higher-pitch tones are thought as coming from small vibrating objects located at higher locations. In fact, as will be discussed in Chap. 5, for strings of equal density and under the identical tension, shorter strings will vibrate at higher frequencies, and the resonances of smaller hollow objects will be higher than larger hollow objects of roughly the same shape, for example, violins versus violas, versus cellos, versus double basses. There are many other common correspondences; bigger objects tend to be heavier and louder than smaller ones, and shiny objects tend to be slipperier than matte ones.

The degree of association lies on a continuum. Size and pitch are tightly coupled, size and weight less so, while size and color or pitch and color would not be associated to any degree. I think that the beliefs in any of these associations stems from experience and do not believe that these sorts of associations are "hardwired." In fact, it is relatively easy to bring about an association. Ernst (2007) paired the brightness of a green LED light with haptic stiffness and within 500 trials was able to induce a weak association.

To study the cross-modality correspondences, researchers have compared response times for congruent and non-congruent stimuli. For example, suppose the stimuli for the auditory modality was a low- or high-pitch tone, and the stimuli for the visual modality was a light above or below a fixation point. The two congruent stimuli would be low-pitch/below-fixation light and high-pitch/above-fixation light; the incongruent stimuli would be low-pitch/above-fixation light and high-pitch/below-fixation light. On the cross-modality trials, one of the four stimuli would be presented and the subject would be required to judge either the auditory or visual stimulus (not both). On the baseline trials, only the auditory or the visual stimulus would be presented.

Two comparisons are possible. Compared to the baseline trials, if there is a *congruent* correspondence, then it should be easier to identify the low-pitch tone if it was paired with a light below fixation and the high-pitch tone if it was paired with a light above fixation. Moreover, it should be easier to identify a light below fixation if it is paired with a low-pitch tone and a light above fixation if it is paired with a high-pitch tone. Conversely, if there is an incongruent cor-

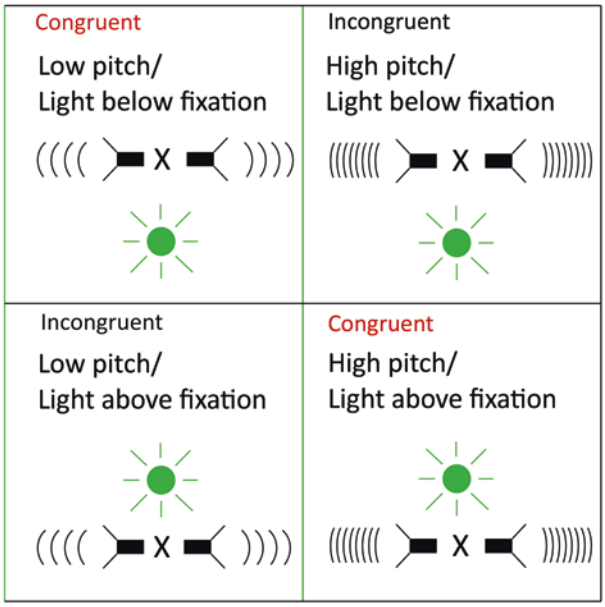


Fig. 2.23 Congruent and incongruent stimuli used to study cross-modal correspondence. The "x" is the fixation point, and the sound is often presented by two speakers equilateral from the fixation point placed behind a screen

respondence, low-pitch/above-fixation light or high-pitch/below-fixation light, then it should be harder to identify either the pitch or position. These tasks are really about the interference due to the irrelevant modality; remember the participant is judging only the stimulus of one modality. Although Evans and Treisman (2010) and Parise and Spence (2009) used slightly different experimental procedures, both found better performance for congruent than non-congruent stimuli. The strongest effect was found for combinations of pitch and position, and pitch showed a greater change in reaction times than position. Not every pairing of attributes did lead to a congruency gain; pitch and visual brightness did not (Fig. 2.23).

To sum up, in this task, the effect of cross-modal congruency seems to occur across a variety of auditory and visual features with the greatest effect occurring for auditory pitch. There are many auditory properties of sounds and many properties of visual stimuli. The correspondence is between specific aspects of each. There are others aspects that will not support any correspondence. As Parise (2016) notes, a better understanding of how these congruencies come about will require a better understanding of environmental correlations, e.g., do higher pitches actually occur more frequently at higher elevations.

Another kind of cross-modal congruency is sound symbolism, the association between the acoustic and articulatory qualities of consonants and vowels and the perceptual qualities of stimuli (see Sidhu & Pexman, 2018 for an

extensive review and analysis). If asked to assign the nonwords *mil* and *mal* to objects of the different sizes, people call the smaller object *mil* and the larger one *mal*. In similar fashion, if people are asked to assign the nonwords *maluma* and *takete* (or *bouba* and *kiki*) to rounded or sharp angular objects, people assign *maluma* or *bouba* to the round smooth object and takete or *kiki* to the sharp edged object.

In general, vowels produced with the tongue placed against the roof of the mouth and toward the front of the mouth (/i/ as in heed) that yield a higher pitch are associated with small objects while low and back vowels that yield a lower pitch (/u/ as in who'd) are associated with large objects. This connection may reflect the fact that smaller objects do tend to vibrate at higher frequencies. In similar fashion, consonants that do not involve stopping the airflow, /m/ as in mac, /b/ as in barn are associated with round objects, while consonants that do involve a blockage, /p/ as in pat, /t/ as in take, are associated with sharp, angular objects. This blockage followed by a sound burst may convey the abrupt directional changes in angular objects.

Sidhu and Pexman (2018) offer several possible explanations for these seemingly arbitrary associations. The associations are arbitrary because aspects of the words form cannot be used to infer its meaning. The articulatory movements generate intertwined acoustical properties so that there are multiple possible associations and it will be difficult to determine the importance of any property. There could be specific acoustical characteristics of the sounds or articulatory motions used to produce the sounds that lead to the connection of */mil/mal/* to small/large and /maluma/takete/ to round/angular. This is analogous to the cross-modal congruency between a pure high pitch tone and location or size. Usually, though, the symbolism is due to many acoustic and contextual factors.

2.3.5.1.2 Bimodal Judgments of Physical Properties

One physical property that has been extensively studied is texture, particularly the roughness of surfaces. Roughness can be perceived through vision, in terms of grain size, density regularity, or reflectance; through touch, in terms of sharpness, stickiness, friction or hardness; and through hearing, in terms of the sounds produced by rubbing or tapping the surface. The visual perception of roughness is mainly based on the spatial variation in the brightness of the elements illustrating depth due to shadowing. The tactual perception of roughness seems to occur in two regimes. If the "bumps" on the surface are somewhat widely separated (greater than 0.2 mm), the perceived magnitude of roughness is based on the area of the finger that is contact with the bumps, that is, indented by the surface texture. Here, roughness is spatial, and the speed of the hand motion across the surface does not affect the perceived roughness. If the bumps are tightly packed, the perception of roughness becomes based on the vibratory pattern on the skin. The intensity of the vibrations is more important than the frequency. The auditory perception of roughness would be based on the sounds produced by the exploration of the surface. Movements of bare

fingers across the surface yield only low-amplitude sounds and do not seem to affect roughness judgments. Movements of a probe such as a dowel yield louder sounds, but they seem relatively unimportant when judging roughness. Auditory cues can and do affect perceived roughness in other instances. If the amplitude of higher frequency (greater than 2000 Hz) sounds produced when rubbing hands together is increased, that brings about the perception that the hands are more paper-like. The perceived roughness/moisture of the skin decreases and smoothness/dryness increases. If the sound was delayed by even 100 msec, the illusion was weakened; the sensory integration required temporal coincidence. Jousmaki and Hari (1998) have termed this the "parchment-skin illusion."

On the whole, tactual and vision judgments of roughness are equal but based on different properties. Bergmann-Tiest and Kappers (2007) asked participants to judge about 100 different objects, including plastics, glass, metals, abrasives, papers, foams, and so on, in terms of their roughness. In the visual condition, the participants could not touch the stimuli, but could examine the objects from all directions. In the touch condition the participants were blindfolded, but could feel and grasp the stimuli as often as they wished. The results indicated that the visual and tactual roughness judgments were highly correlated for each individual. But, participants often judged roughness differently, demonstrating that roughness is not a single property. The accepted physical measure for roughness is the variance of the heights along the surface, but participants also mentioned judging roughness visually in terms of the indentations, shininess, as well as dull spots and tactually in terms of softness, fine or coarse bumps, and the friction along the surface.

These results suggest that visual and tactual representations of roughness are different. Vision seems attuned to the structural properties of shape, size, and element density that can be identified in one glance, while touch seems attuned to surface properties that must be identified with slower hand motions. Thus, it would be unlikely that bimodal presentation of visual and tactual stimuli would give rise to better discrimination or identification. In fact, it is rare that bimodal presentation does produce better outcomes. When faced with conflicting visual and tactual stimulation, for example, a smooth visual surface along with rough sandpaper, participants will often average their judgments 50:50. But, if given instructions to emphasize one sort of property, participants can readily bias their judgments toward visual or tactual properties so that it is unlikely that the visual and tactual single-modality judgments are lost when making a combined judgment. Both Whitaker, Simões-Franklin, and Newell (2008) and Klatzky and Lederman (2010) come to the similar conclusion that visual and tactual roughness perception are independent but complementary.

2.3.5.2 Conflict Between Cross- and Within-Modality Organization

As described previously, auditory stream segregation of interleaved low- and high-pitch tones is determined by the frequency ratio among the sounds and

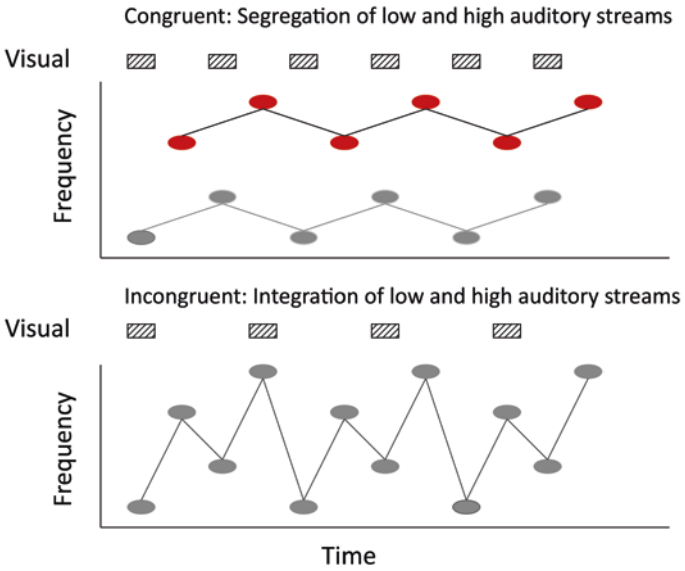


Fig. 2.24 Visual stimuli were presented with the low-pitch tones in the congruent presentation condition and that promoted segregation. In contrast, the visual stimuli were presented with every fourth tone in the incongruent condition. Here, the light occurs equally often with each high and low tone and that interfered with segregation leading to the integration of the low- and high-pitch tones

the rate of presentation. These two factors interact so that it is possible to shift the perception from one stream to two streams by either increasing the frequency separation or increasing the presentation rate. Rahne et al. (2007) investigated whether visual presentations of geometric forms synchronous with the sounds can also affect the organization of the interleaved sounds. Their experiments were complex, and in Fig. 2.24 I have created a simplified version of the human streaming part of the research. The simplified auditory sequence in all cases is L1, H1, L2, H2. Rahne et al. (2007) used three different frequency separations (although only one is shown in Fig. 2.24). The critical variable was the timing of the synchronous visual squares. In the top panel, the squares are synchronous with every low-pitch tone, thereby making the low pitches more distinctive. This presumably would increase the probability of separate streams for the low- and high-pitch tones. In the lower panel, the visual squares occur every third sound so that it bounces back and forth among the four different tones. In a 12-note sequence, the squares are synchronous with each pitch one time, and this presumably would tend to knit the low and high tones, and yield one stream.

The results confirm these predictions. Except for the narrowest frequency separation where the tones rarely, if ever, form separate streams, the congruent presentation of the visual squares led to a higher percent of separate streams than did the incongruent presentation or presentation of the tones

without any visual stimuli at all. Surprisingly, there was no difference between the congruent segregating presentation of the squares and the no visual presentation conditions.

Instead of creating the congruent conditions using a second modality, we can imagine this experiment using tones that vary in frequency and timbre. In the congruent conditions, the lower-pitch and higher-pitch tones would have different timbres (a violin versus a clarinet) and this should increase the probability of two streams. In the incongruent condition, the timbres would oscillate among the tones and that should lead to a higher probability of a single stream. What these results would show is that the relationship between the stimuli in different modalities can affect the organization of each one in the same way that variations in the properties of stimuli in one modality can do.

In the bouncing ball configuration, two circular stimuli starting at the opposite sides of a screen move toward each other. Two different percepts can occur. In the first, each stimulus seems to pass through the other and continues to the other side of the screen (termed *streaming* by the authors). This can be understood as an example of the Gestalt principle of good continuation. In the second, the stimuli seem to bounce off each and reverse direction (termed *bouncing* here).

Watanabe and Shimojo (2001) investigated whether brief auditory sounds affected the probability of those two percepts. As pictured in Fig. 2.25, the simple visual presentation leads to a preponderance of streaming response. If a sound occurs when the two circles overlap (B), the percept shifts and now the vast majority of responses indicated that the circles seemed to bounce and reverse direction. The sound suggests a physical collision. However, if the identical sound is presented several times during the visual motion, the percept reverts back to streaming (C). The multiple sounds are grouped together, and thereby disrupt the perception that the coincident sound with the overlapped circles indicated a collision. Remember, a sensation is affixed **only** to one event or source. The bounced percept can be recovered if the coincident tone is either a different frequency or different loudness (D). The unique coincident tone is not integrated with the other tones and thus is interpreted as signaling a bouncing event (Fig. 2.25).

The research of Rahne et al. (2007) and Watanabe and Shimojo (2001) demonstrate that stimuli in one modality can affect the probability of alternative organizations in a second modality. Moreover, Watanabe and Shimojo (2001) show the trade-off between within-modality organization and between-modality organization. To the extent that the sensations in one modality are perceived as belonging to each other, that is, grouped together, they will have little effect on the other modality. We will see this again in Chap. 3 for multistable percepts.

Another kind of cross modality organization has been termed "*intersensory gestalten*." The underlying question is whether it is possible to bind elements from different modalities into a multisensory percept that differs from the percepts that occur within each modality if presented separately. It is actually easier

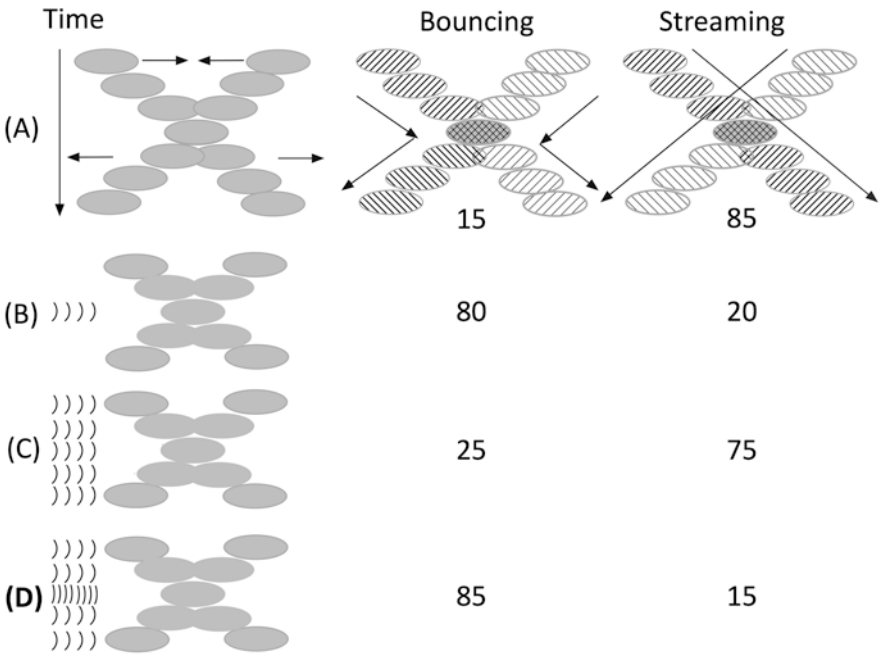


Fig. 2.25 When two lighted circles approach each other and then separate, two percepts are possible. The circles could appear to cross each other and continue on their way or they could appear to bounce off each other and return to their original locations. The normal perception is streaming (A). If a tone occurs as the circles merge, the circles now seem to bounce off each other (B). However, if the tones occur throughout the movement, the tones are perceived to be independent of the visual motion, and the perception reverts to streaming (C). Finally, if the tone synchronous with the merge is changed in frequency or increased in loudness, that tone loses its connection with the other tones so that bouncing becomes the dominant perception again (D). (The hatching in the top row is merely to illustrate the bounce and streaming percepts)

to give an example than provide a simple definition. Huddleston, Lewis, Phinney Jr, and DeYoe (2008) first placed four lights or four speakers at the three, six, nine, and twelve o'clock positions. The lights or sounds were presented in the clockwise order or counterclockwise order: the 3, 6, 9, and 12 positions or the 12, 9, 6, and 3 positions, respectively. Participants were able to correctly judge the direction for either the light or sound arrays. For the critical test, the authors placed just two lights and two speakers on a horizontal board in front of the participants. The lights were placed at twelve and six o'clock and the speakers were placed at three and nine o'clock. The stimuli were presented either clockwise or counterclockwise. If clockwise, the order was the 12light, the 3sound, the 6light, and the 9sound and so on. The basic question was whether the sequence was perceived as a unified circling on the board or as a sequence of two lights moving back and forth vertically along with a sequence of sounds moving

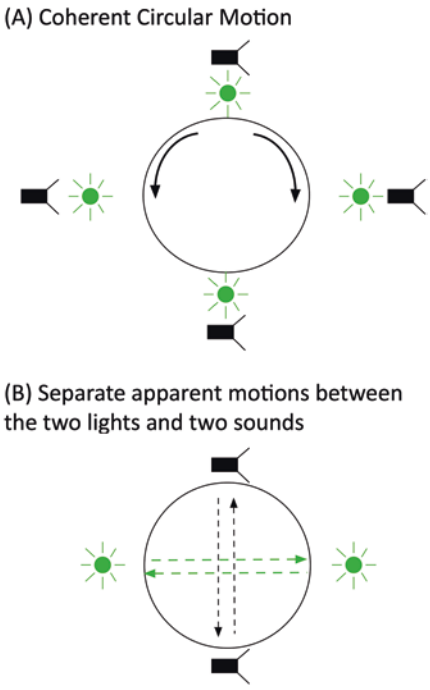


Fig. 2.26 Coherent circular motion is perceived using either the four lights or four tones, and participants can judge the direction of rotation (A). However, the perception of rotation disappears if the participants have to integrate the positions of two lights and two tones. Instead, the tones are heard to move back and forth horizontally and the lights to move back and forth vertically (B)

back and forth horizontally. Participants always saw the lights and sounds move independently; they never saw an integrated circular motion (Fig. 2.26). (The independence of visual and auditory motion will be discussed in Chap. 3).

It is risky to draw a conclusion from a negative outcome given the many variables that could affect the result. Possibly blinking lights are poor stimuli here. Alternately, a different procedure in which there are four lights and four tones, one at each clock position. Initially all the stimuli are presented to create a strong circular motion percept, and then two of the lights and two of the tones are faded out. Possibly the circular motion would be maintained. At this time, I would be hesitant to draw any conclusion.

2.3.5.3 Perceptual Shifts Due to Cross-Modality Interactions

To summarize at this point, cross-modal correspondences can lead to faster detection of congruent stimuli, although there is little effect on accuracy. In addition, the stimuli in a second modality can affect the organization of another modality if those stimuli are perceived as connected to the first. But, in both cases, the sensations in the second modality do not change the perceptions in

the first. The second modality may change the probability of alternatives in the first modality that would occur in their absence, but new percepts do not occur. Here we consider three instances in which the percept is altered: the double flash illusion, spatial and temporal ventriloquism, and the McGurk effect.

2.3.5.3.1 Double Flash Illusion

Shams, Kamitani, and Shimojo (2002) found that if more than a single auditory beep accompanies a single flash, the percept is that of two or more flashes. The light seems to oscillate on and off. A single flash accompanied by two beeps is perceived to be two flashes, but the effect tapers so that three or four beeps do not give rise to significantly more flashes. To determine the limits of the timing of the auditory beeps, the authors made one beep simultaneous with the flash and then varied the timing of a second beep so that it either preceded the simultaneous pair or followed it. The timing window was basically symmetrical. As long as the second flash was within 100 msec of the paired flash/beep, two flashes were perceived. In later research, Mishra, Martinez, and Hillyard (2013) found that the illusionary flash was the same color as the actual flash. If the initial flash was red (or green), then the illusionary flash was also red (or green). The color was identified before the tone brought about the illusionary flash.

Roseboom, Kawabe, and Nishida (2013), noting that the inducers were two auditory beeps in the same modality, varied the modalities of the two inducing stimuli in later studies. In some conditions they used the same inducers, two auditory noise transients or two tactual pulses (presented successively on one finger). The double-flash illusion was identical for both modalities, demonstrating that the illusion was not restricted to auditory inducers. In three other conditions the two inducers differed: (a) a noise transient and a tactile pulse; (b) a noise transient and a sine wave tone; (c) a low-pitch sine wave tone and a significantly higher-pitch sine wave tone that would not stream together. (In all conditions, the two inducers were in the first and second position equally often). What is critical is that the double-flash illusion did not occur in any of these conditions. Thus the two inducers have to be of the same sort (i.e., to be grouped together) in order to bring about the double-flash illusion. One possible explanation is that the two sounds or touches recruit a second light flash so that the number of inducers and lights become equal thereby creating a stronger correspondence between the sounds (or touches) and lights.

2.3.5.3.2 Temporal Ventriloquism

Temporal ventriloquism occurs when the onsets of a sound and visual stimulus differ slightly, and the onset of the visual stimulus is incorrectly perceived as being closer in time to the abrupt onset of the sound. The auditory onset captures the visual onset.

Previous work has used sequences of auditory and visual stimuli and investigated the influence of one on the other. For example, Recanzone (2003) presented an auditory sequence and a visual sequence at the same time. In the

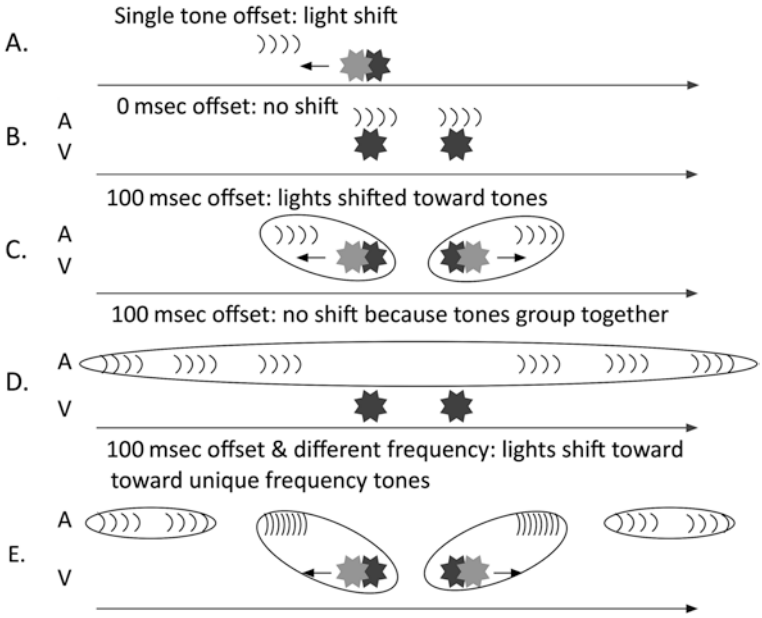


Fig. 2.27 Temporal ventriloquism: Presentation of tones can affect the perceived timing of visual stimuli. The darker stars represent the original timing of the lights; the gray stars and arrows indicate the change in temporal position due to the tone presentation

experimental conditions, the visual rate differed from the auditory one. If the participants were asked to judge the auditory rate and ignore the visual one, they were able to do so. But they could not ignore the auditory rate, and their judgments of the visual rate were strongly influenced by the auditory rate.

A simpler example of temporal ventriloquism arises if onset of a single tone that precedes (or follows) the onset of a visual stimulus within a temporal window of about 100 msec. In these cases, the visual stimulus is perceived as being shifted in time toward the onset of the sound (see (A) in Fig. 2.27). More typically, two sounds either bound two visual stimuli, AVVA, or are presented between the two visual stimuli, VAAV. In the former case, AVVA, the interval between the two visual stimuli is perceived as being longer. In the latter case, VAAV, the interval is perceived to be shorter. Somewhat surprisingly, temporal ventriloquism will occur regardless of the position of the tones. Even though the flashes are in front, both tones can be on the same side of the head with no effect on the visual displacement.

But, as illustrated throughout this chapter, perceiving is a compromise between within-modality and cross-modality organization and this is also true for temporal ventriloquism. Using a similar procedure as Watanabe and Shimojo (2001), Keetels, Stekelenburg, and Vroomen (2007) embedded the auditory tones that typical attract the visual stimuli onsets in a sequence of tones. If the onsets of the tones and lights were synchronous as shown in (B), temporal

ventriloquism in terms of a change in the perceived interval between two lights did not occur. But, if the first tone preceded the second light by 100 msec and the second tone followed the second light by 100 msec, the interval between the two light flashes seemed to increase. This latter outcome (C) is the expected outcome of temporal ventriloquism. However, when the two bounding tones are part of a series of equal-timed preceding and following tones (D), temporal ventriloquism did not occur. All the identical tones were perceived as being part of one auditory stream unrelated to the light flashes and therefore did not capture the onsets of the lights. (Another possible reason for the lack of effect is the difference in the number of identical auditory sounds and the two lights). But if the two bounding tones are made different, either by changing the frequency or intensity, temporal ventriloquism reoccurs. The bounding sounds are not perceived as part of the ongoing sequence, so they capture the onsets of the lights. These outcomes are identical to those for the bouncing/streaming experiment discussed previously.

One common finding is that temporal ventriloquism is asymmetric. The auditory stimulus captures the visual one, but the reverse does not occur. The auditory-visual offset does not change the timing of the auditory sound. The explanation is that the perceiver "goes with" the more reliable signal with the least variability. Timing discrimination is much better for on/off beep auditory signals than on/off flash visual signals and therefore from a Bayesian perspective the optimal strategy is to focus on the auditory signal. If this explanation is true, then "weakening" the auditory signal so that it reliability equals that of the visual one should eliminate the asymmetry. Following this strategy, Vidal (2017) masked the auditory signal by white noise and found that the difference in capture between audition and vision disappeared. Furthermore, it also should be possible to equate auditory and visual capture by finding visual stimuli that are more rhythmically salient. For example, Iverson, Patel, Nicodemus, and Emmorey (2015) found that both normal and hearing-impaired participants could more accurately synchronize to a bouncing ball visual stimulus than a flashing light one.

Two points are important here. First, based on these results, we might question whether any reported modality asymmetries could be enhanced, neutralized, or even reversed by the choice of the auditory and visual stimuli. Second, we would still expect that perceivers to place greater weight on the more reliable cue, whether that is auditory, visual, or tactual.

2.3.5.3.3 Spatial Ventriloquism

Spatial ventriloquism occurs when a sound and visual stimulus occur at the same time and the perceived location of the sound is misplaced toward the perceived location of the visual stimulus. The obvious example occurs when the ventriloquist's voice is located at the moving mouth of a dummy or when the sound from a television loudspeaker is heard as coming from the visual source on the screen. Another kind of spatial ventriloquism has been termed sensory saltation (to be discussed in Chap. 4). If one stimulus is presented at one location, and a

second is presented within roughly 80 msec at a different location, the first stimulus is perceived to be closer to the second stimulus. Tactually, it seems to hop.

A bare-bones set-up to demonstrate spatial ventriloquism consists of a tone presented from one position with a light flash presented from a second location. In contrast to temporal ventriloquism, the onsets of the tone and light are synchronous. If the participant is asked to point toward the tone and disregard the flash, the response nonetheless is angled toward the light flash. The "pull" of the light flash is strong, but the flash does not capture the tone, rather the perceived location is a compromise. Rarely is the tone located at the exact position of the flash. The ventriloquism effect occurs even if the participant merely imagines the light (Berger & Ehrsson, 2018). In these configurations, there is little or no effect of the auditory beep on the location of the visual stimulus (see review by Chen & Vroomen, 2013).

From the perspective of the unity assumption, we would expect spatial ventriloquism to happen as long as the light and tone are perceived to come from the same event. This would suggest that there are both spatial and temporal disparities beyond which the sound and flash are assumed to come from different sources. In these cases, there is no reason to integrate or recalibrate the two sources and the displacement of the tone should not occur. The temporal window of integration extends from -100 msec (sound before light) to +300 msec (sound follows light), with the degree of displacement decreasing at the extremes. The spatial window of integration is roughly +/− 15°. For temporal ventriloquism, the temporal window is slightly shorter, but in contrast, there is really no spatial window.

Spatial ventriloquism can occur between sounds and tactile pulses. Caclin, Soto-Faraco, Kingstone, and Spence (2002) had participants place fingers on a centrally located vibrator and presented tones to the right or left. The vibrators influenced the perceived location of the sound so that the sounds were perceived closer to the center than when the same sounds were presented alone. The influence of the vibrations occurred only if the onsets of the sound and vibration were synchronous, emphasizing the link between the two.

Why does such ventriloquism occur? There is always going to be the need to recalibrate our senses. The speed of light is far greater than the speed of sound, but the speed of auditory neural processing is faster than the speed of visual neural processing. This means that at most distances the visual and auditory sensations will not occur at the same time. It has been estimated that the sensations will be synchronous at distances around 10 m. Sounds will appear before the visual stimulus if the source is closer than 10 m and the reverse if the distance is beyond 10 m. The rather long temporal windows alleviate some of these problems. The sensations from two senses do not have to match perfectly to yield a fused percept. Observers are far more likely to judge stimuli from different modalities as being simultaneous than to judge two stimuli within the same modality as being simultaneous. The difference can be up to five times greater, 200 msec between senses versus 40 msec within a modality (Vroomen & Keetels, 2010).

In addition, there are aftereffects of the spatial and temporal discrepancies that can lead to long-term recalibrations that might be necessary to accommodate changes in body size or environment. If a sound-flash or flash-sound sequence is presented repeatedly, on subsequent presentations the interval between the sound and flash would seem shorter and the sound and flash may even appear synchronous. If a flash follows a finger press at a fixed interval, it is possible to create an illusion that the flash preceded the finger press if a subsequent flash occurs at an unexpectedly short interval. Similar aftereffects occur for spatial ventriloquism. After the synchronous presentation of displaced sounds and lights, sounds presented alone are shifted toward the position of the light. Usually the aftereffect occurs only after multiple synchronous presentations and that shift is a fraction of the original discrepancy. The aftereffect is restricted to the actual sound; the aftereffect does not occur if the frequency of the adapting tone is changed, say from 750 Hz to 300 Hz (Recanzone, 1998). Aftereffects of temporal asynchrony occur between vision and touch (Takahashi, Saiki, & Watanabe, 2008). Vision, the less reliable modality, becomes aligned to the more accurate touch modality and the degree of the aftereffect is roughly the same as the aftereffect between vision and audition.

Since the publication of "Hearing Lips and Seeing Voices" (McGurk & McDonald, 1976), the McGurk effect has become the focus of research on the integration of auditory and visual phonemic information, and as described by Alsuis, Paré, and Munhall (2017), a proxy measure for audiovisual integration. An auditory phoneme presented by a hidden loudspeaker occurs at the same time that a visible face silently mouths a different phoneme, and the participants are asked to report the auditory phoneme and disregard the "silent talking face." Critically, the auditory phoneme would have been perfectly identified if presented alone. Surprisingly, participants were unaware of the conflicting information and the classic McGurk illusion response is the incorrect fusion of the auditory and visual phonemes; if an Auditory [ga] was paired with a Visual [ga], the fused response would be [da], or if an Auditory [ga] was paired with a Visual [ba], the response would be either [ba] or [b'ga], a combination of the two phonemes. In general, the strength of the illusion is measured by the ability of the irrelevant visual phoneme to disrupt the identification of the auditory phoneme either by phonemic fusion, combining the two phonemes, or by simply identifying the visual phoneme as the auditory one.

Speaking yields both auditory and visual information about the articulation of speech, and people normally assume that the auditory and visual sensations come from the same event. Thus, we would expect that the simultaneous auditory and visual information would signal the same phoneme. When there is conflict, listeners must decide if the auditory and visual sensations do or do not come from the same event. Many studies have attempted to probe the limits of this integration. Since the McGurk stimulus configuration is analogous to those found for both for temporal and spatial ventriloquism, we would expect the same limits. Although the auditory and visual stimuli do not need to be synchronous due to the temporal window, the effect of the visual input decreases as the asynchronous reaches +100 msec when the auditory leads, or reaches +480 msec

if the visual leads. The quality of the visual face also affects the strength of the illusion; the orientation and size of the face, whether the voice and face are the same gender, clarity of the image, and the familiarity of the face all affect the illusion. Moreover, the illusion can be affected by cognitive factors such attention, awareness, expectation, and suggestion. Moreover, there are rather large differences in participant's susceptibility to the illusion (see Alsuis et al., 2017 for an extensive review). It is interesting to note that although inverting the face reduced the McGurk illusion, it did not affect the spatial localization of the face.

Another aspect of audiovisual integration is the balance between the perceived within coherence of the auditory and visual stimuli yielding auditory and visual streams against the perceived coherence between the auditory and visual stimuli yielding fused or combination stimuli. Research discussed above shows that due to the unity assumption, at least initially within-modality coherence takes precedence. It is possible therefore to minimize temporal and spatial ventriloquism by embedding one of the stimuli in a unified stream or by disrupting any linkage between the auditory and visual stimuli. Nahoma, Berthommier, and Schwartz (2012), following the latter strategy, initially presented a series of auditory phonemes accompanied by an unrelated motion picture so that the phonemes and visual gestures on the film were unrelated. Following this incoherent context that presumably led to the unbinding of the auditory and visual sensations, the usual McGurk incongruent auditory/visual stimulus was presented. The preliminary stage led to a large reduction in the percent of McGurk responses (i.e., the fusion or combination of the auditory and visual stimuli). Based on this outcome, and the fact that the McGurk illusion could be reinstated by a short series of congruent auditory and visual stimuli, Nahoma et al. (2012) argue that the first step in multisensory interactions is the binding of the stimuli in the component modalities followed by the interpretation of that binding.

2.3.6 Visual Event Perception

Up to this point, we have discussed the perception of static visual stimuli or the movement of rigid objects. Here we will consider the perception of objects composed of different parts that undergo common and relative motion. Starting with simple geometric stimuli, Gunnar Johansson evolved a powerful theory based on vector analysis and then applied this theory to the recognition of dancers from lights at different skeletal joints.

YouTube Videos

Gunnar Johansson: Motion Perception, Parts 1 &2, Biomotionlab. Narrated by James Maas, Cornell University. (Old but classic). This is an important video.

Gunnar Johansson Experiment, Brain Hackers Association

Issey Miyake A-POC INSIDE, www.dv-reclame.ru. Fun, but slightly weird

Biomotionlab.ca (Nikolaus Troje Research). Demonstrations

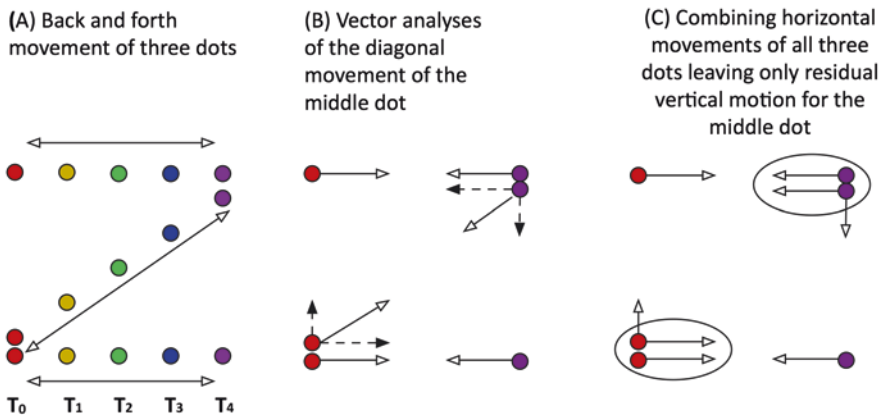


Fig. 2.28 In (A), the movement of each dot is shown at five time points (each colored differently). In (B), the diagonal movement of the middle dot (open arrow) is split into its horizontal and vertical motion components (closed arrows). In (C), the horizontal components are bound to the horizontal motion of the outer dots leading to the perception of up and down motion only. A similar example is shown in the YouTube video "Motion Perception, Part 1"

The basic idea is that the percept often does not correspond to an accurate physical description of the movement of each part, but is based on an abstraction of how the parts of an object move relative to one another. Points in motion are always perceived in relation to each other, and those points undergoing simultaneous motions are automatically perceived as rigid objects moving in three-dimensional space. Points with similar motion paths are seen in the same plane as found for the common motion in non-rigid arrays discussed in Sect. 2.3.4.2.

The overall motion is broken into two parts. The first is the common movement of all of the parts of the rigid perceptual object that becomes the reference frame and seen as undergoing one motion. The common parts are spatially invariant and usually undergo the slowest movements. The second is the relative movements among the parts, seen as units attached to the common motion. Johansson (1973) argues that vector analyses ensure that the percept of the object is maximally rigid so that it maintains constant size and form.

In Fig. 2.28, one dot moves diagonally in phase with two dots that move horizontally so that at the rightmost horizontal position (purple dot) the diagonal one is adjacent to the upper dot, and at the leftmost horizontal position (red dot) the diagonal dot is adjacent to the lower dot. The "accurate" perception is to see the inner dot as moving diagonally; instead, the perception is that of the inner dot moving vertically up and down in phase with the horizontal movement of the other dots. (B) The diagonal movement of the inner dot can be broken into two vectors at right angles. The horizontal vector is integrated with the horizontal movements of the upper and lower dots and no longer seen as part of the inner dot's motion; (C) What is left is the relative up and down vector, which is seen as vertical movement only.

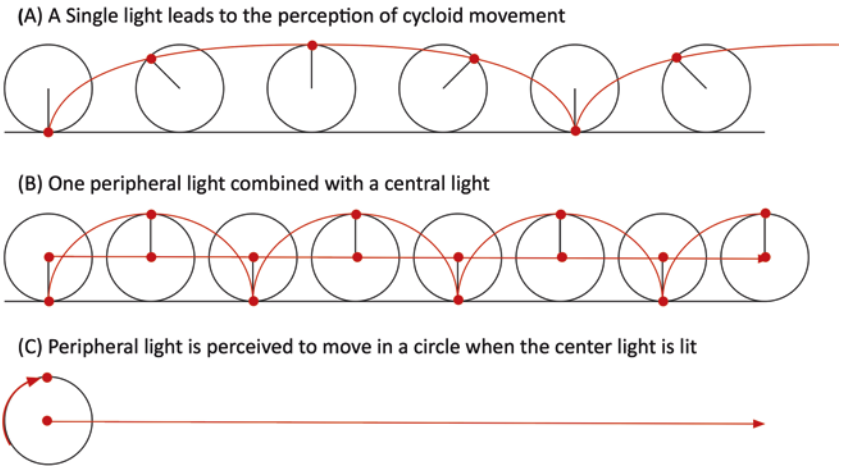


Fig. 2.29 (A) A single light mounted on the periphery of a rolling wheel generates the perception of cycloid motion. (B) If a single central light is added to the peripheral light, viewers report seeing a single light circling around a moving wheel (C). A demonstration is shown in the YouTube video "Motion Perception, Part 1"

A second configuration to demonstrate vector partitioning comes from a rotating wheel with a lighted point on its periphery and a second lighted point at the center axis, also illustrated in Fig. 2.29. If the peripheral light is shown alone (A), the light will follow a cycloid, a loopy path composed of the rotation around the center point and the translation of the wheel. If the center light is added, the peripheral light now seems to rotate around the central light and both move at the speed of the central light. The central light strips away the translational component of the cycloid, leaving only the rotational motion. Johansson (1975) argued that the abstraction of the common vector components is obligatory and occurs early in the visual pathways.

These simple examples give us insights into how we see complex motions composed of many moving parts and paths. For walking movements, the visual system creates a hierarchical arrangement of the different motions, each understood in terms of the common motions of the higher and slower hierarchical levels. The motion of a person's ankle rotation is relative to the knee, the motion of the knee is relative to the hip, and so on. Here the relative motions are pendulums due to the circular motion of the joints, but that is not necessary for vector analyses. Because the motion of the ankle rotation must follow that of the leg due to knee rotation, it must be faster than the leg.

Johansson made use of point-light displays to investigate the perception of biological motion. The light spots are attached to the joints of moving actors against a dark background. It is very difficult, if not impossible, to perceive the form of the actor from one frame. Moreover, the trajectories of each light created by the movement of various joints oscillate up and down and are meaningless taken one by one. Johansson was interested in how individuals used the patterns of moving lights to derive a coherent form depicted by the motions. His pro-

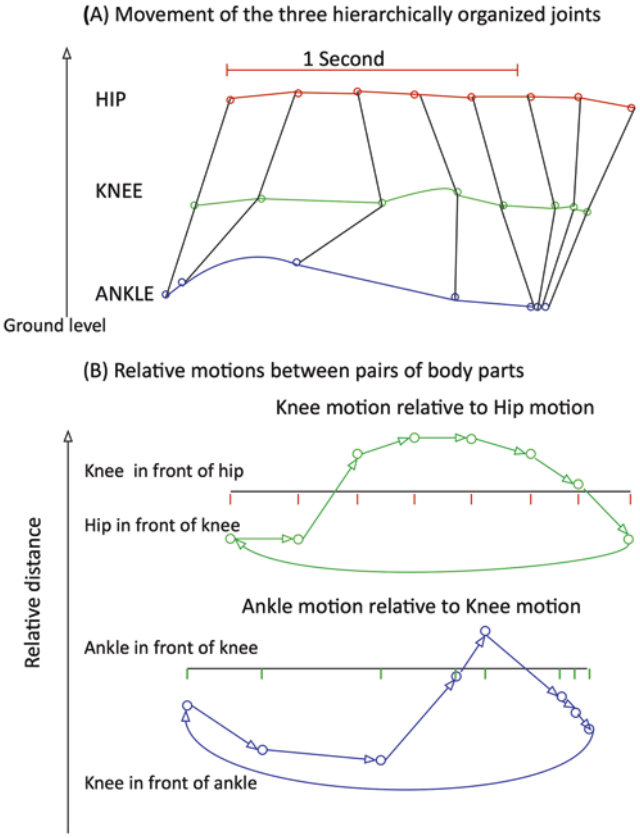


Fig. 2.30 (A) The vertical movement of the hip, knee, and ankle is shown for slightly more than 1 sec. Each motion is an entirely ambiguous, slow vertical oscillation. (B) For each pair, hip to knee and knee to ankle, the relative motion between the lower faster body part (knee and ankle) and the connected slower part (Hip and knee) is shown at each time point. The relative motions illustrate the pendulum-like motion of the lower parts. In both cases, the lower body part is first behind and then in front of the slower bigger part. (Adapted from Johansson, 1973)

posed solution was based on the vector analyses; the motion of each dot is perceived relative to a connected dot moving more slowly over a longer distance as shown in Fig. 2.30. It is these intrinsic non-common motions across time rather than the common motions in space that give rise to the structure of the object.

The movements of dots that maintain the same separation signify parts of rigid object and that makes the identification of form easier. There are only about a dozen disconnected dots, but identification is rapid taking less than 200 msec. Observers can recognize the gender of walkers, whether they are running or walking, and even recognize friends from the "style" of the movement seen in the patterning of the motions of the joints. As discussed

above, when people are asked to assign the nonsense syllables *maluma* and *takete* to rounded and angular static visual figures they invariable match *maluma* to the rounded figure and *takete* to the angular figure, an example of crossmodal correspondence and sound symbolism (Kohler, 1929). In similar fashion, jerky, rapid movements are labeled *takete* and smoother, slower motions are labeled *maluma* (Koppensteiner, Stephan, & Jäschke, 2016). There can be multisensory enhancement. Thomas and Shiffrar (2013) added the sounds of footsteps that were either synchronized or out of phase to the point light foot movement. Surprisingly, either timing made identification easier. Yet, presenting a meaningless tone did not affect identification using either timing.

The perception of action is quite resistant to various sorts of degradation, including randomly varying the contrast of the lights, running the motion backwards, or presenting only a subset of lights on the body. Inverting the display made the body motion far more difficult to perceive (Pavlova & Sokolov, 2000). As the upright figure is rotated more than 60°, the percept changes from a walker to "swinging dots back and forth" or "rotation of a stick or a hand." The upright orientation was necessary for participants to make use of prior knowledge to identify the target. In addition, Cutting (1981) found that placing the light spots off the joints so that the distance between the dots was not constant made the perception of form slightly more difficult while changing the timing of the lights made the perception quite difficult (Hiris, Humphrey, & Stout, 2005). All this suggests that the perception of the dots as representing walking may depend on the expectations and visual frame of reference of the observer. The motions of the lights are constrained by the hierarchical construction of the body; the trajectory and speed of different lights will vary with body position. Making a shoulder light move like a wrist light impairs recognition. The observer makes use of a pre-existing framework to search for particular kinds of within- and between-modality correlations.

It was been often argued that the ability to detect biological motion is privileged, better than other forms of movement. This does not seem to be the case, however. Hiris (2007) found no difference between the detection of biological motion and motion of other types of structured forms. Differences in detection will depend on the specifics of the forms, presentation methods, and the disrupting stimuli used to mask them.

As will be discussed in Chap. 4, there are many parallels between Johansson's vector model that yields rigid objects in vision, rhythmic beats in sound that yield coherent musical passages, and the timing of body movements when dancing. There is still another parallel when reaching to make one of the exploratory hand movements. The shoulder and arm movements must be disregarded to isolate the hand and finger movements. In all, there are layers of structure (and this has been a constant theme) that form frameworks that interlock the movements in space and time, and the notes in time.

2.3.7 Camouflage

The appearance of an animal can aid survival in many ways. Neutralizing the grouping principles used to isolate objects and events can make those animals difficult to detect. Alternatively, bright striped coloration that makes the animal more visible may signal that the prey is noxious, having secondary defenses such as spines, poisonous chemicals, or a nasty taste. In the same way, making human danger signs red and caution signs yellow increases their effectiveness. Other kinds of skin patterning could be used to make animals appear bigger and stronger to bluff predators and yield better outcomes in pursuit of sexual partners.

Nearly all the research on camouflage involves visual detection. I can think of several reasons for this. First, many auditory and olfactory cues to animal identity and location may not be perceivable to human observers. For example, many vocalizations of small mammals are in the ultrasonic range (greater than 20,000 Hz) and inaudible to humans. Second, while visual detection reveals both the identity of the animal and its location, auditory detection, while possibly reliably identifying the animal, is unlikely to reliably locate the animal. There is some evidence that animals choose high-frequency calls that quickly dissipate in the environment to minimize location information (Ruxton, 2009). But, except for movies in which humans try to disguise their voices using handkerchiefs over phones, or birds that imitate the calls of other birds, there is little evidence that animals attempt to disguise their vocalizations. For these reasons, we will concentrate on visual camouflage.

The mechanisms and tricks animals use to avoid visual detection are quite diverse. In general, *crypsis*, defined as initially preventing detection, aims to break down figure-ground organization. Its mechanisms include (a) matching of the background color or texture, that is, masquerading; (b) obliterative shading, which minimizes three-dimensional form; and (c) disruptive coloration in which a set of markings creates the perception of a false set of boundaries and edges. A different type of crypsis is self-shadow concealment, in which shadows caused by directional light are cancelled by countershading that will be discussed in Chap. 5.

Other kinds of camouflage seek to mislead predators by masquerading as a different animal or object (e.g., an insect masquerading as a twig), or by patterning that makes the calculation of speed and direction of movement difficult (e.g., a zebra's stripes, although this is controversial). It must be kept in mind that our understanding of detection camouflage is based on our own perceptual systems' sensitivities, and may not match those of the predators that the animal is trying to outwit (Stevens & Merilaita, 2009). For example, insects and birds see ultraviolet light and other animals may possess receptors tuned to specific regions in the visual field. Moreover, there is little information about how animals integrate motion, color, and/or contrast.

We start with the two basic forms of cryptic coloration. Background markings attempt to match the background in terms of color, luminance, or texture, while disruptive colorations attempt to create false edges and boundaries and/or obscure the real ones. In terms of the Gestalt grouping principles, both interfere or overrule contour formation based on proximity and good continuation (Kelley & Kelley, 2014).

Background camouflage is understood in terms of how well the animals' visible surfaces match the texture and coloration of the background. In some cases, the animal cannot change its coloration even if the environment varies. Nonetheless, the coloration may match two or more aspects of the environment. Fishes that swim near the surface have bright, shiny underbodies to match the bright water surface above them, but have dull backs to match the darker deep water below them. Both serve to hide the fishes from predators (McPhedran & Parker, 2015). In other cases, the animal coloring does change to match variations in the background. The simplest cases are small Arctic species that change appearance due to the shedding of feathers and furs. Here, the animal's change in color from brown to white and white to brown goes along with the season, that is, temperature and light periods. But, this change is not caused by the appearance or disappearance of snow in the background.

Finally, there are animals that can change color almost instantly as a function of changes in the background. Cuttlefish are magicians at this, being able to match continuously varying backgrounds. Cuttlefish have one of the highest ratios of brain mass to body mass among invertebrates and it seems that nearly all that mass is dedicated to camouflaging. There are three characteristic camouflage patterns: (a) a uniform surface when the background consists of large contrasting surfaces; (b) a mottled surface when the background consists of small contrasting units; and (c) a disruptive pattern when the background surfaces roughly match the size of the cuttlefish as illustrated in Fig. 2.31 (Barbosa et al., 2007). Cuttlefish can also match the texture (smooth versus rocky) of the sea bottom and the transition can take as little as 0.5 sec. What is common in all cases is that the effectiveness of the camouflage depends on the match of color and luminosity and texture. We can imagine that a striped moth would be easily detected against small flower petals, or that a spotted moth would be easily detected perched on a tree with bark that has long, straight fissures. Background camouflage seems best if the animal lives in a stable environment.

Nova DVD

Cuttlefish: Kings of Camouflage (WG41899)

Disruptive coloration is understood in terms of how well the markings on the animal's visible surfaces break up edges and mask its overall shape. Disruptive coloration is strengthened when some of the body parts match the background, while others differ strongly. Sharp changes in color or luminosity

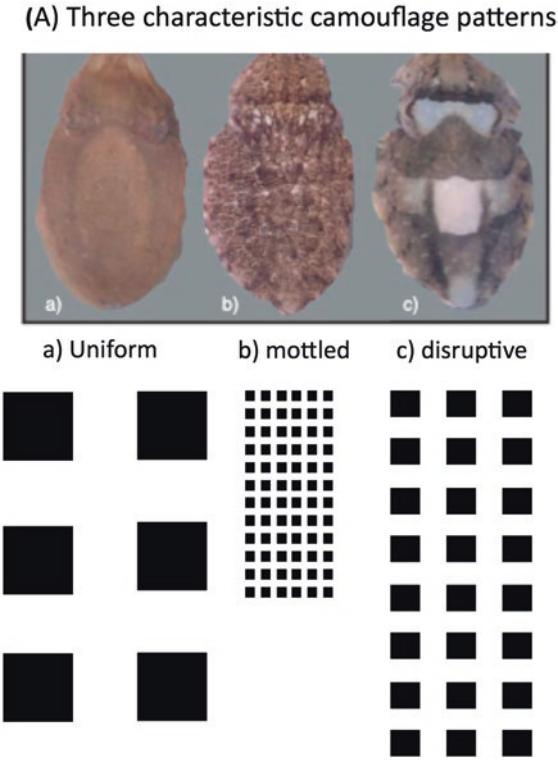


Fig. 2.31 In (A), the three characteristic kinds of camouflage of the cuttlefish are shown. The use of each kind depends on the size of the checker squares of the background. (Adapted from Barbosa et al., 2007)

at the boundaries can be used to create the perception of false edges and disruptive coloration is enhanced by adding highly contrasting color patches within the animal's body that can make its continuous surface look like a set of discontinuous ones (Cott, 1940). These distracting internal patches (not too many, according to Cott) attract attention away from the animal's form. The internal patches dominate the picture, destroying form by "leveling out the contrasts between the animals themselves and their broken backgrounds" (Cott, Page 52). Predators focus on the internal distractor, not the form of the animal, as sketched in Fig. 2.32. Paradoxically, disruptive coloration can act to counter the effect of background coloration so that the best strategy is some combination of the two.

Cuthill et al. (2005) demonstrated that random patterns that butted against the edges of an object were more concealing than the same patterns entirely within the object. Edged patterns broke up the contours of the object; edged patterns that were the same color as the background maximized concealment

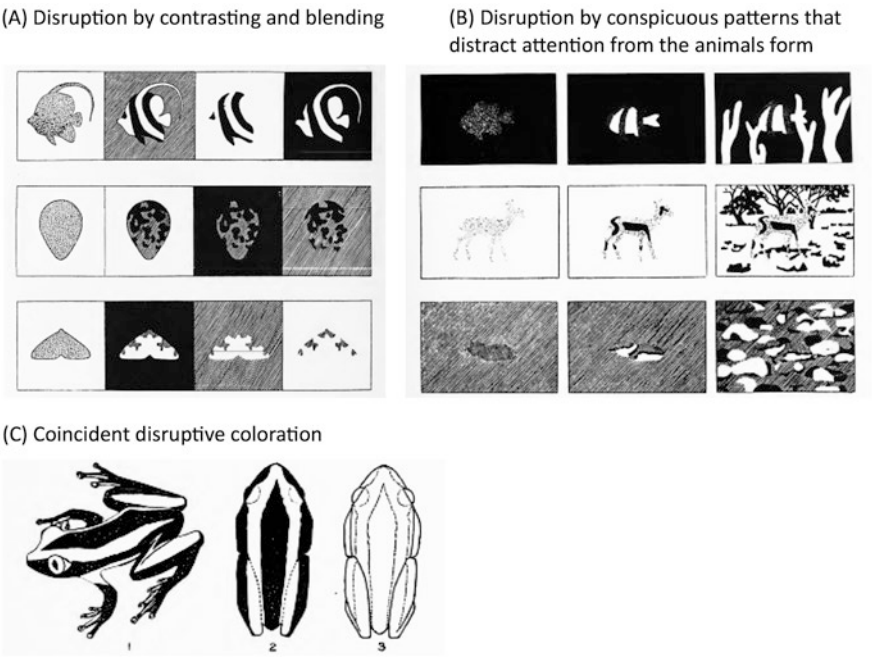


Fig. 2.32 Illustrations from Cott (1940) showing disruptive markings that lead predators to focus on distracting internal patterns and not on body shape (A and B). Coincident disruptive coloration split the frog's body into three separate parts (C). The rightmost drawing shows the frog colorless; the leftmost shows the frog in a jumping position; while the middle shows the frog in a resting position where the coloration conceals the body shape. Cott (1940) emphasizes that any coloration will work only in a limited set of environments, and we should not expect it to be always effective. Cott, H. B. (1940). Adaptive coloration in animals. © London: Methuen & Co. (PDF available from egranth.ac.in)

by giving the impression of scalloped edges (Webster, Hassall, Herdman, Godin, & Sherratt, 2013). Cuthill et al. (2005) further demonstrated that greater contrast between the random edged patterns and the object increased the degree of concealment shown in Fig. 2.33.

Another kind of high contrast disruptive coloration has been termed "dazzle" marking as illustrated in Fig. 2.34. Dazzle markings are thought to work by drawing the eye away from the outline of the object to "destroy the continuity of the surface" (Thayer, 1909). Thayer suggested that dazzle marking worked best when they did not match the background. Imagine a zebra; its dazzle markings would be the white vertical stripes between the black stripes. This alternation is extremely similar to patterns used by Gestalt psychologists to illustrate figure-ground perception (see Fig. 2.4) and therefore might conceal the shape of the entire animal. Thayer's insight was used during World War I to camouflage warships, as shown in Fig. 2.34.

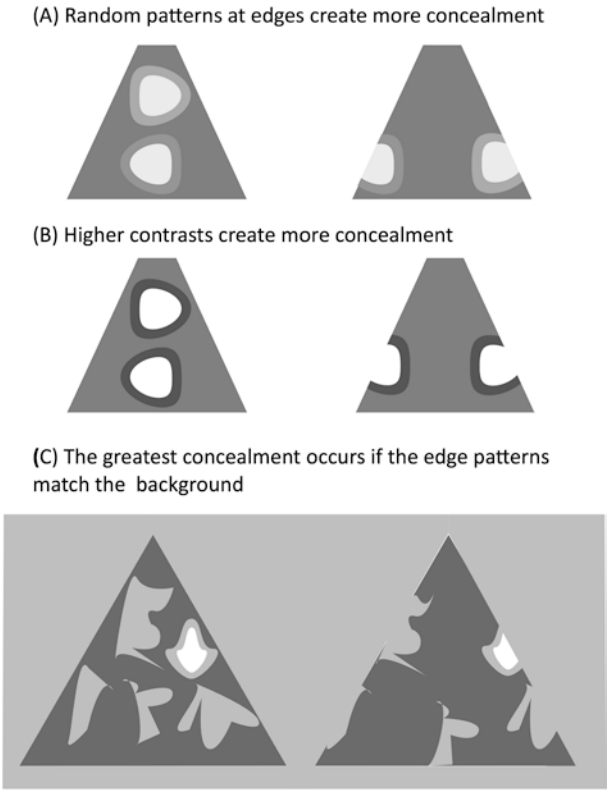


Fig. 2.33 (A) Random patterns at the edges of objects conceal the shape of those objects and bring about a higher survival rate. (B) This advantage is increased when the random patterns are higher contrast such as the "eyespots" on the wings of moths. (C) Concealment is maximized if the edge spots match the coloration of the background

2.4 Perceptual Development

Getting the world right is harder than it seems; being accurate depends on many kinds of grouping principles. The first question, therefore, is whether these principles reflect the properties of the visual, auditory, and tactual fields. Are elements in the visual field that lie close to one another, or share the same coloration, more likely to be parts of a single object? In similar fashion, are sounds that are continuous, close in time, and share the same timbre more likely to come from one source, and are surfaces with the same texture likely to be the same object? If so, then these principles are clearly useful heuristics to perceive the environment accurately. In essence, these questions are equivalent to asking whether an analysis of the sensory data based on previous experience is the best strategy.

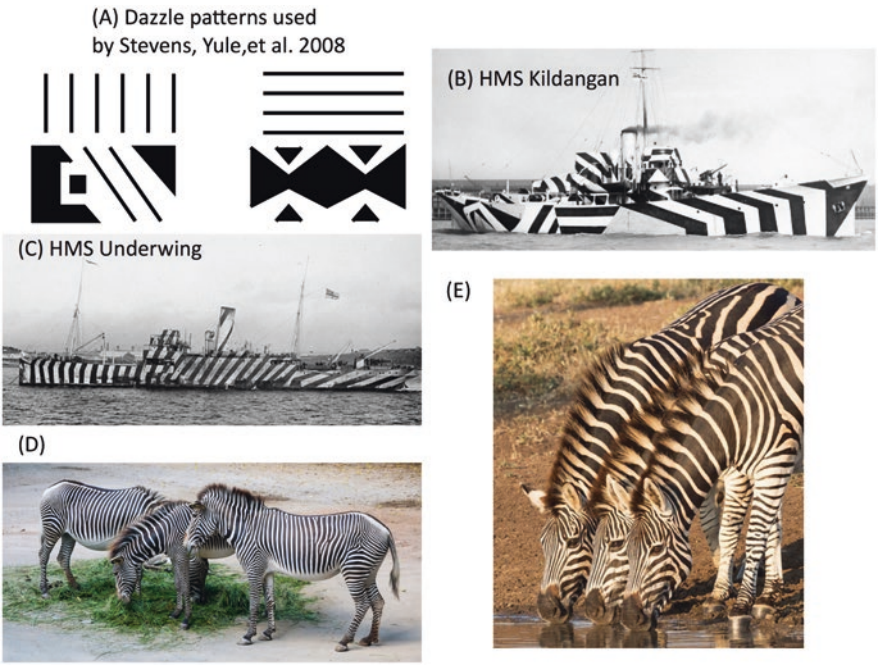


Fig. 2.34 (A) Four different kinds of dazzle markings that resulted in reduced predation. (B & C) Dazzle markings were used extensively during World War I to conceal ships. (B) HMS Kildangan. Photograph Q 43387 in collections of Imperial War Museum (collection no 2500-06); (C) HMS Underwing by Surgeon Oscar Parkes. Photograph SP 142 in Imperial War Museum. Collection no 1900-01). (D & E) The striping on the zebras masks the body outline of each zebra. It is difficult to count the number. (D) pixabay.com; (E) unsplash.com, Vincent van Zalinje, photographer (Creative Commons CCO license)

The only way to answer these questions is to do a detailed analysis of the environment, which means measuring the properties of visual points (e.g., lightness and color using a photometer), and the distance between them. Several observers simultaneously indicate the coherent objects in the scene. Then you calculate statistically whether these points, which have the same physical characteristics or which are in close proximity, are more likely to be part of the same object. In similar fashion, it is possible to determine if sounds with similar timbre close in time tend to come from the same source or if surfaces with the same roughness belong to the same object?

On the whole, the Gestalt grouping principles do accurately reflect the environmental properties (Geisler, 2008) so that they are useful in grouping elements into objects. The probability that two points or sounds that share those physical properties are more likely to be the same object or source is higher than the probability that two randomly chosen points come from the same object or source.

Given that these grouping heuristics reflect properties of objects and sources in the world, the question then arises whether these principles are innate or are learned and constructed early in life. The more general question is whether the human mind is a collection of specific function mechanisms (innate processes) or a single general learning device used by all senses, also innate. This is difficult to decide.

Assume that a baby's vision, while not fully developed, can register differences in brightness and color. A bounded rigid object then could be recognized by the common movement of contiguous (touching) points. For such an object, these points would not move relative to one another, would not intermix with different sorts of points, but would move relative to other bounded objects. These core constraints in the two-dimension "pixel" array at the eyes enable babies to split the external three-dimensional world into surface objects and might suggest that infants possess innate physiological mechanisms that pick up these relationships (Spelke & Kinsler, 2007).

But an infant does not perceive a world of objects like an adult. With experience, as infants handle and manipulate objects they would learn the validity of such principles without needing innate geometric models, that is, augmenting visual experience with tactile experience. Visual, auditory, or tactual experiences with such units enable the infant to abstract properties of objects and then construct solid objects from the somewhat fragmented images. This limited foundation prepares the infant to be able to learn what needs to be learned, namely properties that are correlated and redundant, and that specify three-dimensional objects (see Newcombe, 2011 for an excellent summary of this position). It is important to realize that we are not concerned with basic sensitivity or discrimination abilities that are weak for very young infants although color sensitivity reaches adult levels by two months of age, and by four months infants prefer and categorize colors (Werner, 2012). It takes about three years for the ability to discriminate fine details to reach adult levels.

This process of partitioning the visual surface into objects provides a starting point for abstracting and learning which properties characterize those objects. The learning process can occur by association, as images move out of sight and back again, or by actual manual exploration to reveal the sides and backs of solid objects (Johnson, 2010). There is no need for specific knowledge to be innate. As babies explore, they can discover that the objects are likely to be symmetrical, have smooth convex contours, similar elements, similar textures, and parallel edges. This brings about the expectation that stimulus arrays that have these properties are in all probability rigid objects (with hidden back sides as mentioned before).

In sum, this model assumes a two-stage process. The first splits the retinal images into coherent images at the eye. We would attribute that outcome to the operation of innate processes of connectedness and common fate assuming that these processes are the result of evolutionary forces. The second would be discovering what organizational properties are likely to be true of those objects. This exploration process gives rise to the classical Gestalt grouping principles, which

can be learned at different times and continue to evolve over a lifetime. Proximity, lightness similarity, common motion, and good continuation arise first, followed by form similarity that may require multiple examples (Quinn & Bhatt, 2015). One of the nice features of this approach is that later perceptual skills evolve from previous ones, and do not require discarding older ones (Bhatt & Quinn, 2011).

By analogy, we would hypothesize that the innate perceptual processes that allow the infant to break the evolving sequence of sounds into sources are founded on the duration, timing (e.g., temporal synchrony and inter-element intervals), harmonic relationships among frequencies, and similarity (e.g., pitch, loudness, timbre) among the sounds. For example, temporal groupings would be a useful heuristic to segregate the sounds and there is strong evidence that even two-month-old infants perceive the relative size of the intervals between groups and discriminate among rhythms with different orderings of the intervals (e.g., xxx---xx-x versus xxx-xx---x).

Given this initial split, over time the infant can learn other aspects of sound generation and sequences of individual sources in both musical and speech contexts that yield other grouping principles. In music and speech the timing among sounds originally would just break the sequence into clumps; but with further listening those clumps are differentiated into strong and weak sounds that result in beat and meter grouping. Moreover, originally frequency similarity may just break sounds into different frequency regions. Further listening can yield melodic contour and intonation grouping. Even more experience can result in tonal versus atonal grouping. Recently, Plantinga and Trehub (2014) found that the preference for consonant tone combinations was not innate; six-month old infants did not listen longer to consonant melodies than to dissonant melodies. The preference for "smoother" consonant intervals (e.g., those with frequency ratios such as 3:2, or 4:3) than for "rougher beating" dissonant intervals (frequency rations such as 16:15) depends on extensive musical listening and does not appear until ages 9–12 years. McDermott, Schultz, Undurraga, and Godoy (2016) found that adults living in the Amazon region who had little or no experience with Western music also showed no preference for consonant intervals.

There is little information about the development of tactile perception. There is, however, a steady progression in the ability to make use of tactual features. Infants can probably distinguish size before shape or texture and show some memory for objects by two months. By six months infants can distinguish coarse differences in surface roughness and distinguish sharp angles from smooth curves, but it takes another nine months before young children can identify shape on the basis of overall spatial configuration. The ability to identify geometric shapes occurs at about five years, although children still have difficulty identifying objects that are either larger or smaller than their actual size. There is continual improvement in spatial acuity for up to 10 years before reaching adult levels (Bleyenheuft & Thonnard, 2009). It is tempting to connect the progression of these improvements to the skills involved in the exploratory motions. However, even 10-year-olds did not adjust their exploratory movements according to the task requirements.

2.5 Summary

In this chapter we have described many of the phenomenal characteristics of the visual, auditory, and tactual percepts and argued (or insisted) that the perception of a visual object, auditory event, or tactual surface or object is the result of interacting processes. As discussed in Chap. 1, the myth is that percepts are constructed only in the higher cortical centers. In reality, there are grouping, contour, figure-ground, motion, and temporal cues to the location and identity of objects and events that emerge at the eye, ear, and hand at intermediate cortical centers. The perception of objects is the end result of many perceptual processes and brain regions.

The current view is that neural firings at the lower centers are transformed and elaborated as they travel to the higher centers and are isolated into two neural tracts that convey "what" and "where" information to different parts of the visual and auditory cortex. These tracts are hierarchically organized so that basic features such as color and orientation are transformed into more complex features such as faces at higher levels. Ultimately, specific regions become specialized for particular kinds of stimuli, for example, faces versus text, music versus speech. Moreover, many descending pathways act to "tune" the transformations at the lower cortical centers to the overall properties of the visual and auditory stimuli. There are actually more descending tracts than ascending ones. In sum, cortical regions are constantly changing to match and interpret the sensations.

Visual and auditory objects are immensely complicated so that it would be very difficult to create neural circuits to calculate all these properties at one place. Hence, it makes sense that the brain would calculate each feature separately, allowing it to attend to one feature at a time and to evolve in a changing environment. The visual and auditory regions in the cortex are not uniform as imagined by the Gestalt psychologists where electrical fields automatically yielded the simplest possible organization. But, at the same time the cortical regions are neither encapsulated nor autonomous. Anatomically there are many interconnections so that each tract must influence the other. It seems most reasonable to me that the various tracts and brain regions form fluid coalitions to interpret the perceptual information and to prepare movements. Perceptually it is impossible to determine the "where and motion" without determining the "what and form." Processing at the intermediate and higher cortical regions along with the surrounding context, the particulars of color, timbre, motion, brightness, loudness, and so on, yield the percepts. Those percepts are "layered," formed at different spatial distances and temporal intervals and we might expect the cortical processes to reflect those interactions. I think the ideas advanced by Gunnar Johansson are crucial here. I do not think that distinctions between lower-level physiological processes or higher-level cognitive processes will help understand how people form frames of reference that allow for the abstraction of different motions and different illuminations as discussed in Chaps. 4 and 5.

The outcomes from the multisensory experiments support these concepts. The sensations from the different modalities form fluid bonds that are extremely sensitive to the spatial and temporal synchronies. It is clear that within-modality organization takes precedence over between-modality organization. It is easy to break the bonds between modalities and difficult to institute them. The inability to construct intersensory Gestalten highlights these limitations. The cortical regions act in a collaborate fashion, and at this point we do not understand the significance of activation of the same region by sensations from multiple senses.

There are alternative theories about cortical organization. de Hann and Cowey (2011) suggest that cortical regions are networked, so that regions associated with properties such as color or shape are active when such properties are being analyzed as illustrated in Chap. 1. Deficits for so-called lower-level properties like color are no different from those for higher-level properties such as faces. This suggests that the differences between the two levels may be slight, reducing the importance of a hierarchical organization. A second argument is based on evolution; de Haan and Cowey believe that evolving additional centers devoted to environmental properties are more likely than a massive reorganization of the cortex yielding the what and where tracts. Ongoing research will address these issues.

References

Alsuis, A., Paré, M., & Munhall, K. G. (2017). Forty years after "Hearing Lips and Seeing Voices": The McGurk effect revisited. *Multisensory Research*, 111–144. https://doi.org/10.1163/22134808-00002565

Barbosa, A., Mathger, L. M., Chubb, C., Florio, C., Chiao, C.-C., & Hanlon, R. T. (2007). Disruptive coloration in cuttlefish: A visual perception mechanism that regulates ontogenetic adjustment of skin patterning. *Journal of Experimental Biology, 210*, 1139–1147. https://doi.org/10.1242/jeb.02741

Berger, C. C., & Ehrsson, H. H. (2018). Mental imagery induces cross-modal sensory plasticity and changes future auditory perception. *Psychological Science*, 1–10. https://doi.org/10.1177/0956797617748959

Bergmann-Tiest, W. M. (2010). Tactual perception of material properties. *Vision Research, 50*. https://doi.org/10.1016/j.visres.2010.10.005

Bergmann-Tiest, W. M., & Kappers, A. M. L. (2007). Haptic and visual perception of roughness. *Acta Psychologica, 124*, 177–189.

Bertamini, M., & Casati, R. (2015). Figures and holes. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 281–293). Oxford, UK: Oxford University Press.

Bhatt, R. S., & Quinn, P. C. (2011). How does learning impactdevelopment in Infancy? The case of perceptual organization. *Infancy, 16*, 2–38. https://doi.org/10.1111/j.1532-7078.2010.00048.x

Bleyenheuft, Y., & Thonnard, J. L. (2009). Development of touch. *Scholarpedia, 4*(11), 7958. https://doi.org/10.4249/scholarpedia.7958

Braddick, O. (1995). Seeing motion signals in noise. *Current Biology, 5*, 7–9.

Bregman, A. S. (1990). *Auditory scene analysis: The organization of sound*. Cambridge, MA: Bradford/MIT Press.

Caclin, A., Soto-Faraco, S., Kingstone, A., & Spence, C. (2002). Tactile "capture" of audition. *Perception & Psychophysics, 64*, 616–630.

Chang, D., Nesbitt, K. V., & Wilkins, K. (2007a). *The Gestalt principle of continuation applies to both the haptic and visual grouping of elements.* Paper presented at the Procedings of the second Joint EuroHaptics Conference and Symposium on Human Interfaces for Virtual environments and Telecomputing Systems, Tsukuba, Japan.

Chang, D., Nesbitt, K. V., & Wilkins, K. (2007b). *The Gestalt Principles of Similarity and Proximity apply to both the Haptic and Visual Groupings of Elements.* Paper presented at the Conferences in Research and Practice in Information Technology Ballarat, Australia.

Chen, L., & Vroomen, J. (2013). Intersensory binding across space and time: A tutorial review. *Attention, Perception, & Psychophysics, 75*, 790–811. https://doi.org/10.3758/s13414-013-0475-4

Cott, H. B. (1940). *Adaptive coloration in animals.* London, UK: Methuen & Co.

Cuthill, I. C., Stevens, M., Sheppard, J., Maddocks, T., Parraga, C. A., & Troscianko, T. S. (2005). Disruptive coloration and background pattern matching. *Nature, 434*, 72–74. https://doi.org/10.1038/nature03312

Cutting, J. E. (1981). Coding theory adapted to gait perception. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 71–87.

de Hann, E. H. F., & Cowey, A. (2011). On the usefulness of 'what' and 'where' pathways in vision. *Trends in Cognitive Science, 15*, 460–466. https://doi.org/10.1016/j.tics.2011.08.005

Ekroll, V., Sayim, B., & Wagemans, J. (2017). The other side of magic: The psychology of perceiving hidden things. *Perspectives on Psychological Science, 1745*, 91–106. https://doi.org/10.1177/174569161665467601/11/2017

Elhilali, M., Micheyl, C., Oxenham, A. J., & Shamma, S. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron, 61*, 317–329. https://doi.org/10.1016/j.neuron.2008.12.005

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision, 7*, 1–14. https://doi.org/10.1167/7.5.7

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*, 1–12. https://doi.org/10.1167/10.1.6

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Science, 7*, 252–256. https://doi.org/10.1016/S1364-6613(03)00111-6

Fulvio, J., Singh, M., & Maloney, L. T. (2008). Precision and consistency of contour interpolation. *Vision Research, 48*, 831–849.

Gallace, A., & Spence, C. (2011). To what extent do Gestalt grouping principles influence tactile perception. *Psychological Bulletin, 137*, 538–561.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology, 59*, 167–192. https://doi.org/10.1146/annurev.psych.58.110405.085632

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience, 5*, 887–892.

Halpern, A. R., & Bartlett, J. C. (2010). Memory for melodies. In M. R. Jones, R. R. Fay, & S. E. Palmer (Eds.), *Music Perception* (Vol. 36, 1st ed., pp. 233–258). New York, NY: Springer.

Hayward, V. (2018). A brief overview of the human somatosensory system. In S. Papetti & C. Saitis (Eds.), *Musical haptics: Springer series on touch and haptic systems* (pp. 29–48). Cham, Switzerland: Springer.

Heller, M. A., Wilson, K., Steffen, H., Yoneyama, K., & Brackett, D. D. (2003). Superior haptic perceptual selectivity in late-blind and very-low-vision subjects. *Perception, 32*, 499–511.

Hiris, E. (2007). Detection of biological and nonbiological motion. *Journal of Vision, 7*, 1–16. https://doi.org/10.1167/7.12.4

Hiris, E., Humphrey, D., & Stout, A. (2005). Temporal properties in masking biological motion. *Perception & Psychophysics, 67*, 435–443.

Huddleston, W., Lewis, J. W., Phinney, R. E., Jr., & DeYoe, E. A. (2008). Auditory and visual attention-based apparent motion share functional parallels. *Perception & Psychophysics, 70*, 1207–1216. https://doi.org/10.3758/PP.70.7.1207

Iverson, J. R., Patel, A. D., Nicodemus, B., & Emmorey, K. (2015). Synchronization to auditory and visual rhythms in hearing and deaf individuals. *Cognition, 134*, 232–244. https://doi.org/10.1016/j.cognition.2014.10.018

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*, 201–211.

Johansson, G. (1975). Visual motion perception. *Scientific American, 232*, 76–88.

Johnson, S. P. (2010). How infants learn about the world. *Cognitive Science, 34*, 1158–1184. https://doi.org/10.1111/j.1551-6709.2010.01127.x

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34*, 169–231. https://doi.org/10.1017/S0140525X10003134

Jousmaki, V., & Hari, R. (1998). Parchment-skin illusion: Sound-biased touch. *Current Biology, 8*, R190.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Ferrar, Strauss & Giroux.

Kappers, A. M. L., & Bergmann-Tiest, W. M. (2015). Tactile and haptic perceptual organization. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 621–638). Oxford, UK: Oxford University Press.

Katz, D. (1925). *Der Aufbauder Tastwelt (The World of Touch)* (L. E. Krueger, trans. & Ed.). Hillsdale, NJ: LEA Associates.

Keetels, M., Stekelenburg, J., & Vroomen, J. (2007). Auditory grouping occurs prior to intersensory pairing: Evidence from temporal ventriloquism. *Experimental Brain Research, 180*, 449–456. https://doi.org/10.1007/s00221-007-0881-8

Kelley, L. A., & Kelley, J. L. (2014). Animal visual illusion and confusion: The importance of a perceptual perspective. *Behavioral Ecology, 25*, 450–463. https://doi.org/10.1093/beheco/art118

Kellman, P. J., & Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cognitive Psychology, 23*, 144–221. https://doi.org/10.1037/0033-295X.114.2.488

Kershenbaum, A., Sayigh, L. S., & Janik, V. M. (2013). The encoding of individual identity in Dolphin signature whistles: How much information is needed? *PLoS One, 8*, 1–7. https://doi.org/10.1371/Journalpone.0077671

Kitagawa, N., Igarashi, Y., & Kashino, M. (2009). The tactile continuity illusion. *Journal of Exerimental Psychology: Human Perception and Performance, 35*, 1784–1790. https://doi.org/10.1037/a0016891

Klatzky, R. L., & Lederman, S. J. (2010). Multisensory texture perception. In M. J. Naumer & J. Kaiser (Eds.), *Multisensory object perception in the primate brain* (pp. 211–230). New York, NY: Springer, LLC.

Kohler, W. (1929). *Gestalt psychology*. New York, NY: Liveright.
Koppensteiner, M., Stephan, P., & Jäschke, J. M. P. (2016). Shaking *takete a*nd flowing *maluma*. Nonsense words are associated with motion patterns. *PLoS One, 11*, e0150610. https://doi.org/10.1371/journalpone.0150610
Lacey, S., Lin, J. B., & Sathian, K. (2011). Object and spatial imagery dimensions in visuo-haptic representations. *Experimental Brain Research, 213*, 267–273. https://doi.org/10.1007/s00221-011-2623-1
Lederman, S. J., & Klatzky, R. L. (1987). Hand movements: A window into haptic object recognition. *Cognitive Psychology, 19*, 342–368.
Lee, S.-H., & Blake, R. (1999). Visual form created solely from temporal structure. *Science, 284*, 1165–1168.
McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.
McDermott, J. H., Schultz, A. F., Undurraga, E. A., & Godoy, R. A. (2016). Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature, 535*(7618), 547–550. https://doi.org/10.1038/nature186635
McPhedran, R. C., & Parker, A. R. (2015). Biomimetics: Lessons on optics from natures school. *Physics Today, 68*(6), 32–37. https://doi.org/10.1063/PT.3.2816
Mishra, J., Martinez, A., & Hillyard, S. (2013). Audition influences color processing in the sound-induced visual flash illusion. *Vision Research, 93*, 74–79. https://doi.org/10.1016/j.visres.2013.10.013
Nahoma, O., Berthommier, F., & Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *Journal of the Acoustical Society of America, 132*, 1061–1077. https://doi.org/10.1121/1.4728187
Newcombe, N. (2011). What is Neoconstructivism? *Child Development Perspectives, 5*, 157–160. https://doi.org/10.1111/j.1750.8606.2011.00180.x
Nishida, S. (2011). Advancement of motion psychophyiscs: Review 2001–2010. *Journal of Vision, 11*, 1–53. https://doi.org/10.1167/11.5.11.
Overvliet, K. E., Krampe, R. T., & Wageman, J. (2013). Grouping by proximity in haptic contour detection. *PLoS One, 8*, e65412. https://doi.org/10.1371/journal.pone.0065412
Palmer, S. E., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review, 1*, 29–55. https://doi.org/10.3758/BF03200760
Parise, C. V. (2016). Crossmodal correspondences: Standing issues and experimental guidelines. *Multisensory Research, 29*, 7–28. https://doi.org/10.1163/22134808-00002502
Parise, C. V., & Spence, C. (2009). "When birds of a feather flock together": Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS One, 4*, e5664. https://doi.org/10.1374/journal.pone.0005664
Pavlova, M., & Sokolov, A. (2000). Orientation specificity in biological motion perception. *Perception & Psychophysics, 62*, 889–899.
Pawluk, D., Kitada, R., Abramowicz, A., Hamilton, C., & Lederman, S. J. (2011). Figure/ground segmentation via a haptic glance: Attributing initial finger contacts to objects or their supporting surfaces. *IEEE Transactions on Haptics, 4*(1), 2–12. https://doi.org/10.1109/ToH.2010.25
Peterson, M. A. (2015). Low-level and high-level contributions to figure-ground organization. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 259–280). Oxford, UK: Oxford University Press.

Plaiser, M. A., Bergmann-Tiest, W. M., & Klappers, A. M. L. (2009). Salient features in 3-D haptic shape perception. *Attention, Perception, & Psychophysics, 71*, 421–430. https://doi.org/10.3758/APP.71.2.421

Plantinga, J., & Trehub, S. E. (2014). Revisiting the innate preference for consonance. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 40–49. https://doi.org/10.1037/a0033471

Quinn, P. C., & Bhatt, R. S. (2015). Development of perceptual organization in infancy. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 691–712). Oxford, UK: Oxford Univesity Press.

Rahne, T., Deike, S., Selezneva, E., Brosch, M., König, R., Scheich, H., Böckmann, M., & Brechmann, A. (2007). A multilevel and cross-modal approach towards neuronal mechanisms of auditory streaming. *Brain Research, 1220*, 118–131. https://doi.org/10.1016/j.brainres.2007.08.011

Recanzone, G. H. (1998). Rapidly induced auditory plasticity: The ventriloquism aftereffect. *Proceedings of the National Academy of Sciences, 95*, 869–875. https://doi.org/10.1073/pnas.95.3.869

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology, 89*, 1078–1093. https://doi.org/10.1152/jn.00706.2002

Roseboom, W., Kawabe, T., & Nishida, S. (2013). The cross-modal double flash illusion depends on featural similarity between cross-modal inducers. *Scientific Reports, 3*, 3437. https://doi.org/10.1038/srep03437

Ruxton, G. D. (2009). Non-visual crypsis: A review of the emprical evidence for camouflage to senses other than vision. *Philosophical Transactions of the Royal Society B, 364*(1516), 549–557. https://doi.org/10.1098/rstb.2008.0228

Schellenberg, E. C., Adachi, M., Purdy, K. T., & McKinnon, M. C. (2002). Expectancy in melody: Tests of children and adults. *Journal of Experimental Psychology: General, 131*, 511–537.

Sekuler, A. B., & Bennett, P. J. (2001). Generalized common fate: Grouping by common luminance changes. *Psychological Science, 12*, 437–444. https://doi.org/10.1111/1467-9280.00382

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research, 14*, 147–152.

Sidhu, D., & Pexman, P. (2018). Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review, 25*, 1619–1643. https://doi.org/10.103758/s13423-017-1361-1

Spelke, E. S., & Kinsler, K. D. (2007). Core knowledge. *Developmental Science, 10*, 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Stevens, M., & Merilaita, S. (2009). Animal camouflage: Current issues and new perspectives. *Philosophical Transactions of the Royal Society B, 364*, 423–427. https://doi.org/10.1098/rstb.2008.0217

Takahashi, K., Saiki, J., & Watanabe, K. (2008). Realignment of temporal simultaneity between vision and touch. *NeuroReport, 19*, 319–322.

Teki, S., Chait, M., Kumar, S., Shamma, S., & Griffiths, T. D. (2013). Segregation of complex acoustic scenes based on temporal coherence. *eLife, 2*, 16. https://doi.org/10.7554/eLIFE.00699

Thayer, A. H. (1909). *Concealing-coloration in the animal kingdom: An exposition of the laws of disguise through color and pattern: Being a summary of Abbot H. Thayer's discoveries.* New York, NY: Macmillan.

Thomas, J. P., & Shiffrar, M. (2013). Meaningful sounds enhance visual sensitivity to human gait regardless of synchrony. *Journal of Vision, 13*(14), 1–13. https://doi.org/10.1167/13.14.8

Van Aarsen, V., & Overvliet, K. E. (2016). Perceptual grouping by similarity of surface roughtness in haptics: The influence of task difficulty. *Experimental Brain Research, 2016*, 2227–2234. https://doi.org/10.1007/s00221-016-4628-2

Van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences* (PhD Doctoral). Eindhoven University of Technology, Eindhoven, NL.

Van Polanen, V., Bergmann-Tiest, W. M., & Kappers, A. M. L. (2012). Haptic pop-out of moveable stimuli. *Attention, Perception, & Psychophysics, 74*, 204–215. https://doi.org/10.3758/s13414-011-0216-5

Vidal, M. (2017). Hearing flashes and seeing beeps: Timing audiovisual events. *PLoS One, 12*, e0172028. https://doi.org/10.1371/journal.pone.0172028

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics, 72*, 871–884. https://doi.org/10.3758/APP.72.4.871

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & Von de Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin, 138*, 1172–1217. https://doi.org/10.1037/a0029333

Warren, R. M. (1999). *Auditory perception: A new analysis and synthesis*. Cambridge, UK: Cambridge University Press.

Watanabe, K., & Shimojo, S. (2001). When sound affects vision: Effects of auditory grouping on visual motion perception. *Psychological Science, 12*, 109–116.

Watanabe, O., & Kikuchi, M. (2006). Hierachical integration of individual motions in locally paired-dot stimuli. *Vision Research, 46*, 82–90. https://doi.org/10.1016/j.visres.2005.10.003

Watson, D. G., & Humphreys, G. W. (1999). Segmentation on the basis of linear and local rotational motion: Motion grouping in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 70–82.

Webster, R. J., Hassall, C., Herdman, C. M., Godin, J.-G., & Sherratt, T. N. (2013). Disruptive camouflage impairs object recognition. *Biology Letters, 9*, 0501–0506. https://doi.org/10.1098/rsbl.20130501

Werner, L. A. (2012). Overview and issues in human auditory development. In L. A. Werner, R. R. Foy, & A. N. Popper (Eds.), *Human auditory development* (pp. 1–19). New York, NY: Springer.

Wertheimer, M. (1923). Untersuchungen zur Lehre van der Gestalt. *Psychologishe Forschung, 61*, 301–350.

Whitaker, T. A., Simões-Franklin, C., & Newell, F. N. (2008). Vision and touch: Independent or integrated systems for the perception of texture? *Brain Research, 1242*, 60–72. https://doi.org/10.1016/j.brainres.2008.05.037

Winkler, I., Denham, S., Mill, R., Bom, T. M., & Bendixen, A. (2012). Multistability in auditory stream segregation: A predictive coding view. *Philosophical Transactions of the Royal Society B, 367*, 1001–1012. https://doi.org/10.1098/rstb.2011.0359

CHAPTER 3

Multistability

3.1 Introduction

As I argued in Chaps. 1 and 2, perceiving is hard. Fundamentally, the proximal sensations at the receptors are underdetermined. They do not specify a unique thing (the *distal* object). There always is ambiguity; there is so much visual, auditory, and tactual information that any scene can be understood in more than one way. For example, a ball seen increasing in size might be approaching or it may be inflating. What rescues us from this confusion is that there are neurological and cognitive processes that capitalize on the physical constraints on the shapes, movements, and sounds of objects in our physical world to yield a single percept. In addition, objects are not presented in isolation and that context may act to bring about one percept. I am certain that part of the match between the physical and perceptual world is partly due to evolution and partly to experience as discussed in Chap. 2.

Normally, one possible percept will be *stronger due the above factors*, and it dominates. But, in other instances two or more possible percepts are *equally strong*. In those instances, given enough time, the resulting percept can bounce around the various possibilities. The physical stimulation is unchanging, but conscious awareness fluctuates and the perception of one of the incompatible percepts suppresses all of the others. Transitions may be marked by very brief periods of indeterminate, mixed, or intermediate appearances. But each percept eventually becomes unified and complete; it is **not** an amalgam of the two or three possible competing percepts.

Such scenes, drawings, or sounds have been termed *reversing* figures or *multistable* figures. Each percept may not occur equally often, but each one is a plausible interpretation of the stimulus. I do not think that reversing or

Electronic Supplementary Material: The online version of this chapter (https://doi.org/10.1007/978-3-319-96337-2_3) contains supplementary material, which is available to authorized users.

S. Handel, *Perceptual Organization*,
https://doi.org/10.1007/978-3-319-96337-2_3

multistable pictures should be considered merely curiosities, since all perception allows for alternative potential outcomes. That said, there are ambiguous stimuli that never undergo alternations. A color on the green/yellow border does not reverse.

The original explanation for the perceptual reversals was that they were due shifts of attention, for example, focusing on different faces of the Necker cube (see Fig. 3.2). But the current conceptualization is that the reversals can be due to many factors in addition to shifts of attention. These range from intrinsic properties of the stimulus, to knowledge of the alternatives, to random fluctuations of neural firing, and to neural fatigue resulting in shifts in firing rates. In the following sections we will describe the various types of visual, auditory, and tactual reversing figures, particularly those in which parts of one stimulus are reconfigured to produce a different percept. Then, we will show how neural processes and higher-level cortical processes affect the alternation.

3.2 Visual Multistability

3.2.1 Multistable Static Figures

In vision, the classic methodology is to use ambiguous static pictures, as illustrated in Figs. 3.1 and 3.2. In some cases participants are given prior information about the multiple percepts, while in others they are naïve to the possibilities. Typically, the participant is presented with a single stimulus and asked to indicate whenever the percept shifts. The rivalry is between two alternative interpretations of the same stimulus, is it two faces or one vase.

One type of reversing figure occurs when the two alternatives seem to be at the same depth. In Fig. 3.1, these two reversing percepts are A, D, E, F, G, H, J, K, L, and I. For A, F, and J the alternatives arise from two different orientations of the same perceived object. For I, the triangles can be seen pointing in three alternative directions so that it is a tristable figure. For D, E, G, H, and K, the alternatives arise from grouping the component parts in different ways to bring forth two distinct objects (e.g., a duck versus a rabbit in E).

The second type are figure-ground reversals; two alternate objects appear at different depths, one in front of the other, and the reversal brings the behind object in front and thereby changes the perception. The reversals in B, C, and L occur when the black and white segments reverse in depth. For B, when the white segment is seen in front it appears as a vase and the black segment is just unimportant background. If the black segment is seen in front, it appears to be the silhouettes of two men facing each other and the white segment is background. The border goes with the part in front, it shapes the object as discussed in Chap. 2.

Consider the classic case of a reversing figure, the Necker cube shown in Fig. 3.2. The six lines can be interpreted as being on the two-dimensional surface, or projected into three dimensions. In (A), we have the basic configurations (also see Fig. 3.1A) with six transparent sides. The basic Necker cube reverses in three dimensions easily; in one configuration one of the square faces

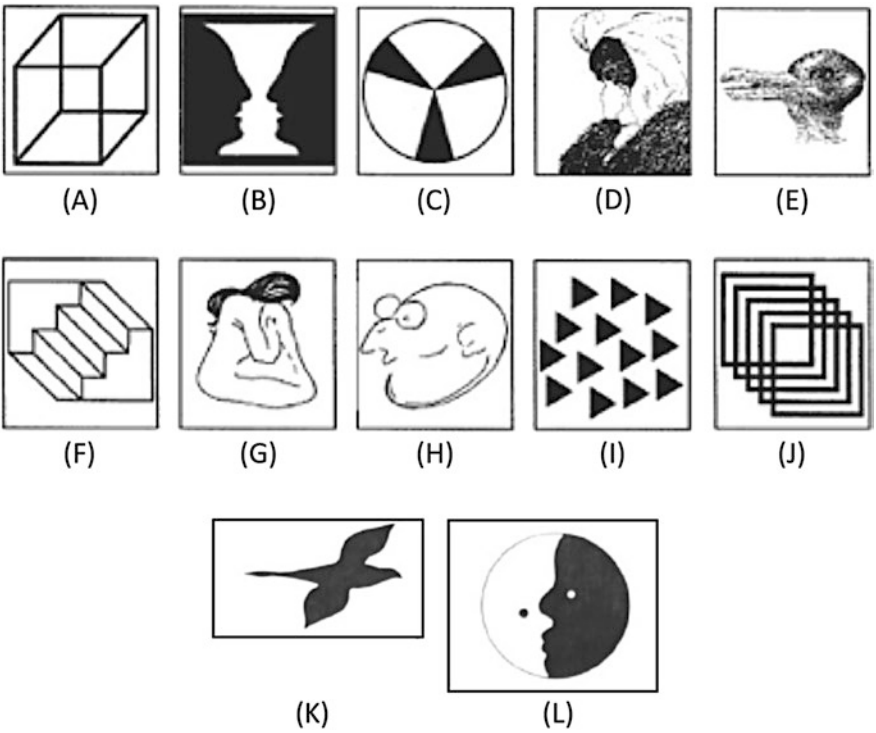


Fig. 3.1 A set of well-known reversing figures. (A) Necker Cube; (B) Face/Vase; (C) Maltese Cross; (D) Wife/Mother-in-Law; (E) Duck/Rabbit; (F) Staircase; (G) Man/Girl; (H) Rat/Man; (I) Triangles; (J) Overlapping Squares; (K) Goose or crow; (L) Gog and Magoog. (Adapted from Long & Toppino, 2004; Fisher, 1968)

appears to point down to the left, in the second configuration the other square face appears to point up to the right. If one of the two square faces is made darker but transparent in 3.2A1, it is still seen in three dimensions and reversals occur easily. If the darker face is opaque, it is still three-dimensional but few reversals occur probably due to the occlusion of one of the corners and one of the connecting edges. The perception of 3.2A as three-dimensional is probably due to two factors. First, all of the corners consist of 90° angles, which, combined with two obtuse angles greater than 90°, signify depth. Second, there are several places where lines cross, which would not be a probable outcome for a two-dimensional object. The more important factor, however, is probably the corner angles because the figure still looks three-dimensional if the crossing lines are eliminated, though now the cube does not reverse (A2).

In 3.2B, the faces meet at a point, four edges still connecting the vertices of the faces. The transparent Necker cube in 3.2B can reverse between two or three dimensions (or reverse in three dimensions). If one face is made transparent so that the connecting edges remain visible (3.2B1), it can still be seen in two or three dimensions. However, removing the hidden edges makes the cube three-dimensional (3.2B2).

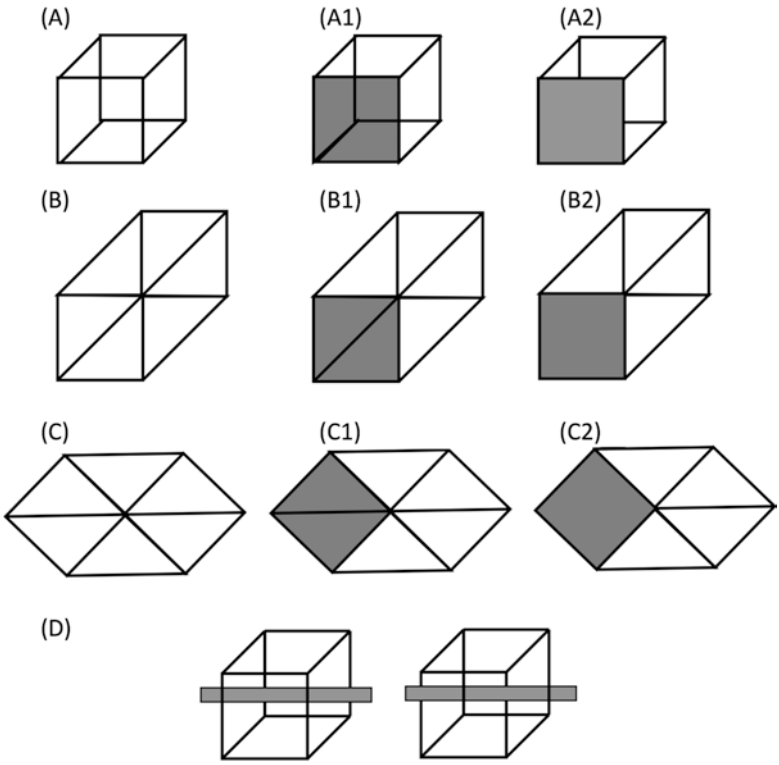


Fig. 3.2 A Necker cube illustrated at different orientations. In (A) through (C), all six faces are transparent. If one face is shaded, different orientations and occlusions give rise to quite different perceptions. In (D), an intersecting rod makes one orientation predominant

If the cube is rotated 45°, it is invariably seen in two dimensions (3.2C) even if one face is made opaque (3.2C1) and the connecting edge is removed (3.2C2). The explanation for the two-dimensional percept for 3.2C and 3.2D is usually based on the "most probable" principle multiplying the prior probabilities by the likelihood that each hypothesis would yield the proximal sensations, that is, Bayesian calculations. Namely, if the distal object was three-dimensional the proximal stimulus would occur only in one improbable orientation and therefore the best bet is to perceive the distal stimulus as a flat two-dimensional object. Such cases, that is, three-dimensional objects oriented to appear two-dimensional, have been termed "accidental coincidences" and thus are unlikely events (Rock, 1983). We can also use the same principle to explain why the preferred orientation of 3.2A and 3.2B is downward to the left. Typically, we would look at a cube at a downward angle on a flat surface so that the front surface would face down and that prevalent viewpoint would predict the preferred orientation of 3.1F and 3.1J. Finally, in 3.2D a rod fixes the orientation of the cube. Adding context stabilizes our perceptions.

YouTube Video

Archival Gibson: There are several videos in the category "Archival Gibson" that illustrate J.J. Gibson's seminal contributions. The videos seem primitive; they were produced before computers and were constructed using single frame photography. But, they illustrate important aspects of visual perception

Bob Shaw-1974. Symmetry and Event Perception: Shows how the perception of the Necker cube varies as a function of orientation

3.2.2 Multistable Dynamic Figures

The simplest multistable case comes about if two lights, tones, or vibrators are flashed alternatively at a fixed distance apart with a constant blank temporal interval between the stimuli. (Fig. 3.3A). The fundamental perceptual issue is determining how the first stimulus relates to the second. The perceptual systems must construct the most plausible correspondence. If the temporal interval is very short, then both stimuli seem to flash at the same time so that they probably represent different objects. If the temporal interval is very long, then the two stimuli flash intermittently and there is no reason to connect the two. However, if the spatial distance and the temporal interval are chosen appropriately, the light seems to move smoothly between the two positions even though the motion is not seen until after the second stimulus is presented. At that timing between the onsets, the illusion of a stimuli moving back and forth alternates with the accurate perception of the alternation. Korte's third law (Korte, 1915) formalized the relationship that to achieve smooth motion the onset interval must increase as the distance increases. This holds true for touch and vision at the best timing for apparent motion and at threshold for audition (Lakatos & Shepard, 1997). However, Gepshtein, and Kubovy (2007) varied the time interval between the lights and found that while Korte's third law was true for higher speeds (i.e., shorter onset intervals) between the lights, a tradeoff between speed and distance occurred at slower speeds. Here, as the distance interval increases, the onset interval must decrease.

If the interval between the first stimulus (left light) and the second stimulus (right light) is shorter than the reverse interval between the right light back to the left, then the perceived motion is left to right (Fig. 3.3B). The shorter interval gives rise to the dominant motion. Freeman and Driver (2008) made use of this asymmetry to show that a tone can affect the direction of the apparent motion, another example of temporal ventriloquism. Specifically, the intervals between the two lights were identical. Then they placed one tone so that it lagged the onset of the first tone slightly and a second tone so that it led the onset of the second light by the same amount. A lag-lead sequence had the effect of making the perceived interval duration shorter and that led to motion from the lag light to the lead light. The effect of the tone offsets was equal to that caused by actually shifting the timing of the lights (without the tones).

The perception of *apparent* motion can be very strong and will occur even if the two lights are different colors or sizes. I have even seen apparent motion between a slide of a mouse and a slide of an elephant, where a trunk grows near

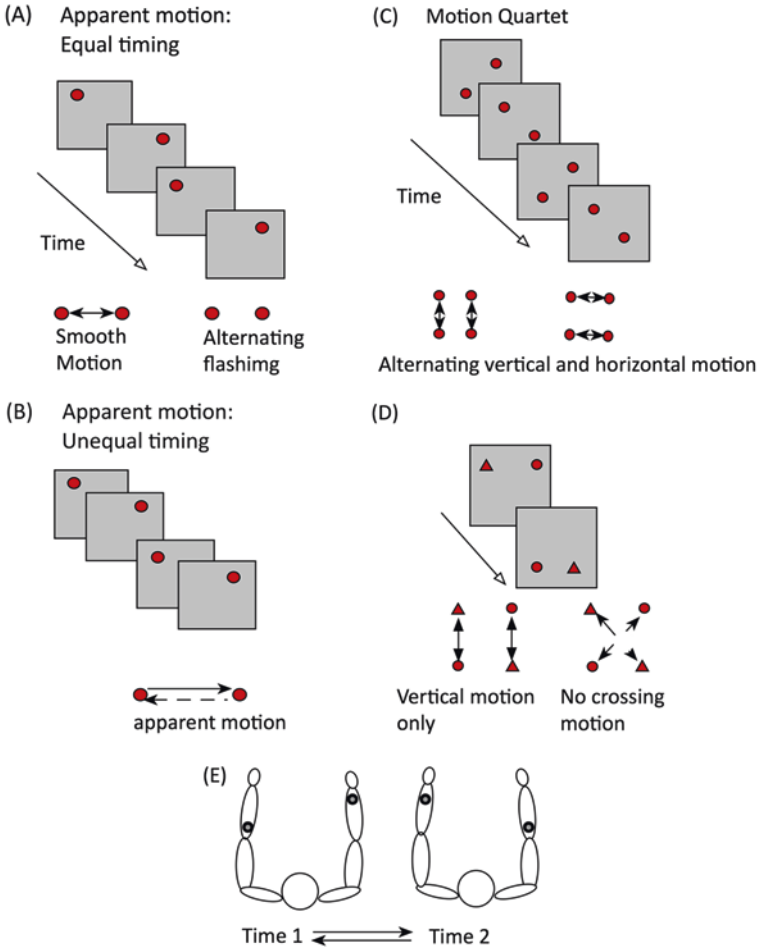


Fig. 3.3 (A) If the same object is flashed alternatively, two perceptions appear: either (1) a smooth movement between the objects or (2) the alternating flashing objects. If the leftward and rightward timing is the same, the speed of the leftward and rightward movement is identical. (B) If the rightward timing is shorter than the leftward timing, the light seems to go quickly to the right and then slowly back to the left. (C) If two pairs of identical diagonal objects are flashed alternatively, either (1) horizontal or (2) vertical motion occurs. Even if the objects are different as in (D), only vertical motion takes place. Crossing motion that connects the same kind of shape does not happen (Web designers beware). In (A), (B), (C), and (D) the different percepts switch back and forth. (E) Apparent motion also occurs between two limbs. The perception crosses the body midline

the elephant and shrinks near the mouse. (I am certain that if the second slide were that of a zebra, then I would have seen stripes appear and disappear in the motion). The visual system solves the correspondence problem by assuming that the two stimuli represent the same object in spite of the physical dissimilarity.

The perception of motion can be so strong that magicians trick audiences into believing that a coin has been pitched between two hands by making the appropriate tossing and catching movements at the right timing as described by Macnik, Martinez-Conde, and Blakeslee (2010).

Furthermore, it is possible to create apparent motion between two vibratory actuators or a light and a tactual actuator on a fingertip (Harrar, Winter, & Harris, 2008). In contrast to apparent motion for two lights, the apparent motion between the vibrators occurred over a shorter range of distances between the stimuli. Apparent motion also occurred between a light and vibrator although the distance between the light and vibrator did not change the strength of the perceived motion. This could suggest a different mechanism for cross-modal apparent movement.

A more complex dynamic visual configuration is based on four lights fixed on the corners of an imaginary square (Fig. 3.3C). In one of the two alternating frames, two diagonal dots are presented and in the other frame the two other diagonal dots are presented. Three percepts occur: (1) the dots are perceived to move back and forth horizontally; or (2) move up and down vertically; or (3) flash periodically. The three percepts can oscillate particularly when the vertical and horizontal distances between the lights are equal. The individual motions are nearly always identical to minimize changes in the configuration. Transforming the square into a tall vertical or a long horizontal rectangle can change the strength of each percept so that strongly horizontal and vertical movement will occur respectively. The motion occurs between the two closest dots and that results in the slowest movement corresponding to gradual movements in the environment. As previously argued, these percepts can be understood in terms of the strong predisposition to perceive a three-dimensional world. It is interesting to note that several apparent motion percepts do not occur: the dots rarely rotate around the square, and the dots do not split apart. Moreover, objects do not appear to move on paths that cross one another (avoiding collisions) even if that results in motions between dissimilar shapes (Fig. 3.3D).

It is possible to create an analogous array using four vibrators. The four vibrators could be placed on one forefinger (Carter, Konkle, Wang, Haywood, & Moore, 2008), or two vibrators could be placed on each forearm (Liaci, Bach, van Elst, Heinrich, & Kommeler, 2016). The latter is illustrated in Fig. 3.3E. The tactual results parallel those for vision with some interesting differences. The kinds of motions are identical: there are parallel vertical or horizontal movements. This occurs even though for the array illustrated in Fig. 3.3E, the vertical movements are within each forearm, while the horizontal movements are between the forearms, crossing the body midline. But, variation in the vertical and horizontal distances among the vibrators has much less effect on the type of movement than for the visual displays. It is possible to create 100% vertical or horizontal movement visually by the placement of the lights, but that does not occur for tactual arrays. The reason for this difference is unclear.

Given that apparent motion is similar for all three modalities, it is possible to explore the interactions when apparent motion occurs in two or three modalities simultaneously. In the basic configuration, there are two lights, two tones,

and two vibrators, one placed to the participants left and one placed to the right. The spatial arrangement is identical: the light is placed in front of the audio speaker and the participant grasps the vibrators in front of the speaker. Moreover, the timing of the stimuli in each modality is identical. The stimuli in each modality are excited left then right or the reverse and the task is simply to judge whether the stimuli moved left or right. (There is only one cycle). In some instances only one stimulus was presented, but in other instances two or three stimuli were presented simultaneously. The two or three stimuli could move *congruently* left to right (or right to left) or the stimuli could move *incongruently* so that one stimulus moved left to right but the other stimulus (or stimuli) moved right to left. (This stimulus presentation does not give rise to alternating perceptions because there is only one cycle, but it does make use of the appearance of motion). These options are illustrated in Fig. 3.4.

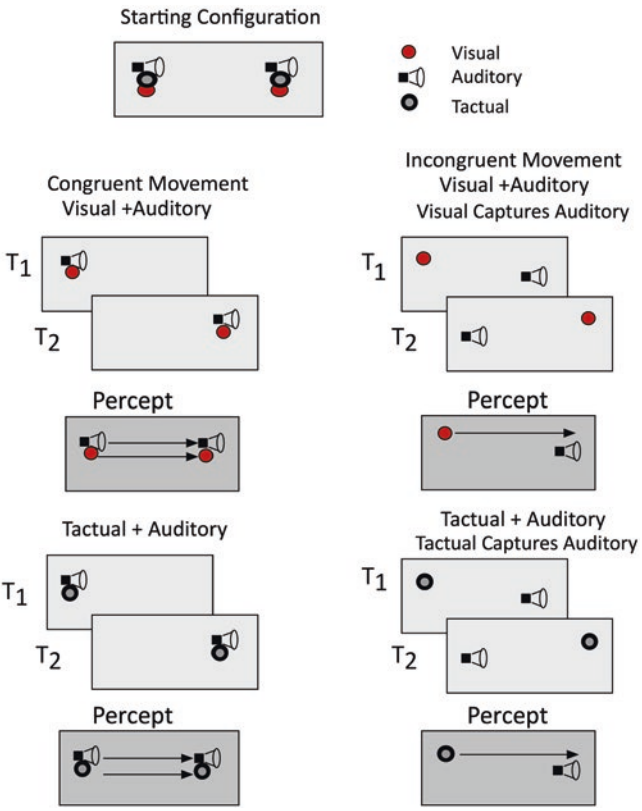


Fig. 3.4 The visual, auditory, and tactual stimuli are stacked on both sides of the display. If the stimuli from two modalities move congruently (either left to right or right to left), the percept is integrated and participants can judge the direction of either modality. If the stimuli move incongruently in opposite directions, the visual and tactual stimuli capture the auditory one. The auditory stimuli either are misjudged to move in the same direction as the visual or tactual stimuli or their direction cannot be judged

If the stimuli in one, two, or three modalities were congruent, then the perception of the direction was essentially perfect. If the stimuli were incongruent, the perception of direction was asymmetric and differentiated the modalities. Vision is the dominant modality. If the visual light and auditory tone or vibrator motions were incongruent, participants were able to perceive the direction of the light perfectly in spite of the tone or vibrators, but were unable to report the direction of the tone or vibrators. The direction of the light captured the directions of the tone or vibrator. In similar fashion, the direction of the touch stimuli captured the direction of the tones although the direction of the tone did affect the direction of the touch somewhat. At least for tones and vibrators, increasing the salience or discriminability of either modality did increase the strength of the capture (Occelli, Spence, & Zampini, 2010). If the direction of the tones was incongruent to the bimodal presentation of lights and touches, the capture of the tones was greater than either lights or touch alone (Sanabria, Soto-Faraco, & Spence, 2005; Soto-Faraco, Spence, & Kingstone, 2004).

Another way to illustrate the influence of a second sense on the perception of multistable figures is by means of binocular rivalry. If different scenes are presented to each eye, say a series of vertical lines to the left eye and a series of horizontal lines to the right eye, the percept oscillates between the two orientations. Rarely do the two scenes fuse into a crisscross pattern. If the observer places one hand on a tactual grid scored with either vertical or horizontal lines, that orientation dominates visually. Switching the orientation of the hand on the tactual surface switches the orientation of the dominant visual percept (Klink, van Wezel, & van Ee, 2012).

3.3 Auditory Multistability

One common approach to study auditory multistability is to make use of the stream segregation paradigm. In the simplest case, two sounds of different frequencies are recycled. If the frequencies are similar or the time interval between the offset of one and the onset of the next is long, the percept is of one oscillating pattern. As the frequency difference is increased, and/or the time interval is shortened, the continuous pattern breaks up and is perceived as two independent sequences or streams; one stream of the low-pitch tones, one stream of the high-pitch tones **(Examples are found in Sound File 2.9)**. For these sequences, sound similarity dominates. In between, there are frequency and time intervals in which the percept switches between the two. In general, the initial percept is that of a single integrated cycling of the tones, another instance of the preference to perceive one source or object. The two interleaved streams percept appears later, as an alternative. After the initial swap, further swaps between the one- and two-stream percept occur after random time intervals. Clearly, the multistability of streaming and apparent visual movement (i.e., movement between the dots versus independent flashing dots, Fig. 3.3A) is analogous both in stimulus configuration and outcome.

The streaming paradigm can be expanded to include three frequencies (Thossen & Bendixen, 2017). The three tones 400 Hz, 635 Hz, and 1008 Hz

cycle at the rate of five tones/sec. There are many possible ways of organizing the three tones. For example, the 400 Hz tone could form one stream and the 635 Hz + 1008 Hz tones could form the second. Moreover, the single-tone stream could be heard in the forefront or the two-tone stream could be heard in the forefront. In general, the streams were based on frequency, the 400 Hz tone versus the two higher pitches or the 1008 Hz tone against the two lower pitches. What is most relevant here is that the organization is multistable, and the average time between reversals was roughly 16 sec. The number and pattern of reversals differed greatly among the listeners, which make sense given the number of alternative organizations.

Sound File 3.4: Multistability for three tone sequences

A more complex example of binaural rivalry occurs with sequences of alternating pitches (Brancucci & Tommasi, 2011). Originally discovered by Deutsch (1974), a pictorial representation of the stimulus sequence is shown in Fig. 3.5. The identical alternations of tones are presented to each ear such that the sequences are out of phase with each other. The very surprising outcome for the vast majority of listeners is that one ear hears only the low tones and the other ear hears only the high tones. The perceived rate of each tone is one-half the actual rate. The tones switch ears at random, although they seem to switch ears more frequently at the faster presentation rate. There is no agreed upon explanation (Fig. 3.5).

A second type of auditory multistability occurs if a sound or word is repeated continuously. For example, Warren and Gregory (1958) have constructed a sequence that simply repeats the word "farewell." As one listens, the word becomes "welfare" and there then is a continual switch between the two words.

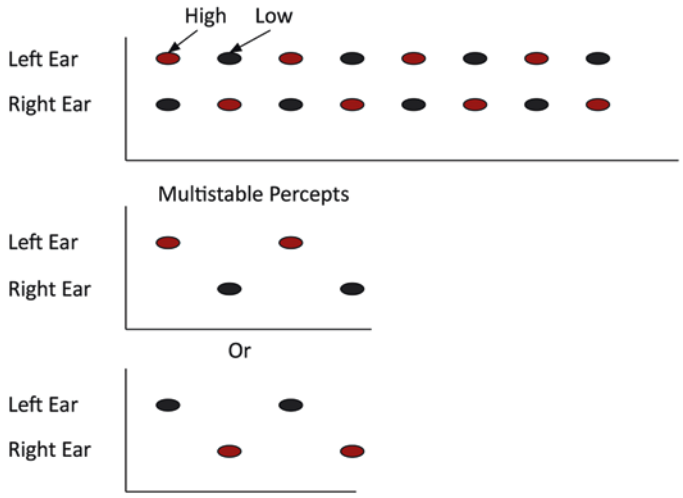


Fig. 3.5 A representation of the out-of-phase low/high tone sequences. The typical perception is either the high tone in the left ear and the low tone is the right ear or vice versa. The two possibilities switch back and forth

Sound Files 3.5: Binaural rivalry creates multistability as illustrated in Fig. 3.5

This is an interesting twist because the two possibilities are due to a timing swap; there is a different starting point for each word, it resembles focusing on a different point in a visual figure. Moreover, as can be seen in Fig. 3.6, the shift to "welfare" involves linking the two syllables across a relatively silent gap. A second example occurs when repeating the word "ace." The perception shifts back and forth between "ace" and "say." This sort of multistability seems analogous to perspective shifts in vision.

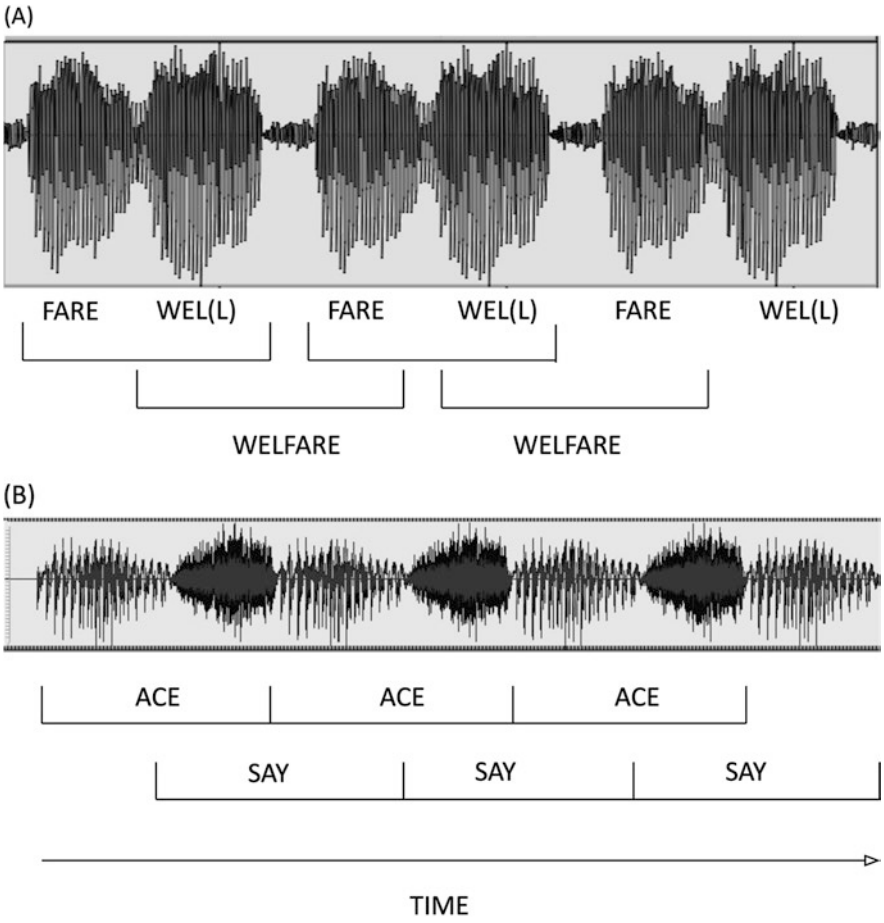


Fig. 3.6 (A) An acoustical representation of three repetitions of the word "welfare." After listening for a short period, the percept changes from "farewell" to "welfare" and then oscillates back and forth. (B) An acoustical representation of two representations of the word "ace." The percept changes from "ace" to "say" and then switches back and forth. www4.uwm.edu/APL/demonstrations.html. These demonstrations are derived from (Warren, 1999)

Sound Files 3.6: Multistability due to repetition of one and two syllable words as diagrammed in Fig. 3.6

3.4 The Nature of the Reversals: No Single Explanation

Taken together, all of this implies that there are many processes underlying these reversals. There are competition mechanisms based on adaptation as well as mutual inhibition at multiple neural-processing levels. Some occur at the initial stages of visual and auditory processing based on neural fatigue and binocular rivalry. Others occur at later cortical areas involved with attention and recognition, the initial local processes embedded in later more global ones. There are always interactions among the levels so that feedback from higher levels prompts reanalysis of lower level percepts. The remainder of this chapter will attempt to tease apart these components. Much of this follows from the reviews by Alais and Blake (2015) and Long and Toppino (2004).

3.4.1 "Bottom-Up" Passive and Automatic Peripheral Processing

As mentioned previously, the original explanation for reversing figures centered on the idea that eye movements aimed at different areas of the figure led to the reversal. It is clearly true that eye movements can expedite and cause reversals, but reversals readily occur without any eye movements at all. For example, if one stares at a reversing picture for a long time and then looks at a blank wall, there will be an afterimage of that picture which will also reverse even though eye movements could not have caused it.

After World War II, there was renewed interest in reversing figures because they seemed to be explained by the electrical brain forces postulated by Gestalt theorists as causing the perception of organized wholes. According to the Gestalt ideas of electrical flow, continuous viewing leads to the satiation and fatigue of those circuits. The fatigue increases the resistance, inhibiting the current in these circuits so that the electrical flow moves to different areas, generating the perceptual switch. Over time, the satiation in the new area increases, so the theory goes, while the satiation in the previous area dissipates, and the electrical flow reverts to its original position and the percept similarly reverts back to the original.

Explanations based on brain currents shifted to those based on inhibition among receptors at the eye or ear because recordings of single cells in cortical regions revealed that neural coding at the receptors created localized responses to particular stimulus features, that is, a horizontal bar versus a vertical one. Such neurons would mutually inhibit each other and the dynamics of the inhibition would create the reversals. Each set of neurons would encode one of the incompatible possible percepts and fatigue or satiation of one set would lead to a reversal.

Among the evidence for a peripheral explanation is that the number of switches increases over time. The recovery from satiation would be incomplete after each cycle and diminish over time. Each swap from percept A to B would occur at shorter intervals because the satiation point for A would be reached faster having begun at a higher recovery level due to less recovery (from the previous reversal). Percept B would then reappear more quickly, as shown in

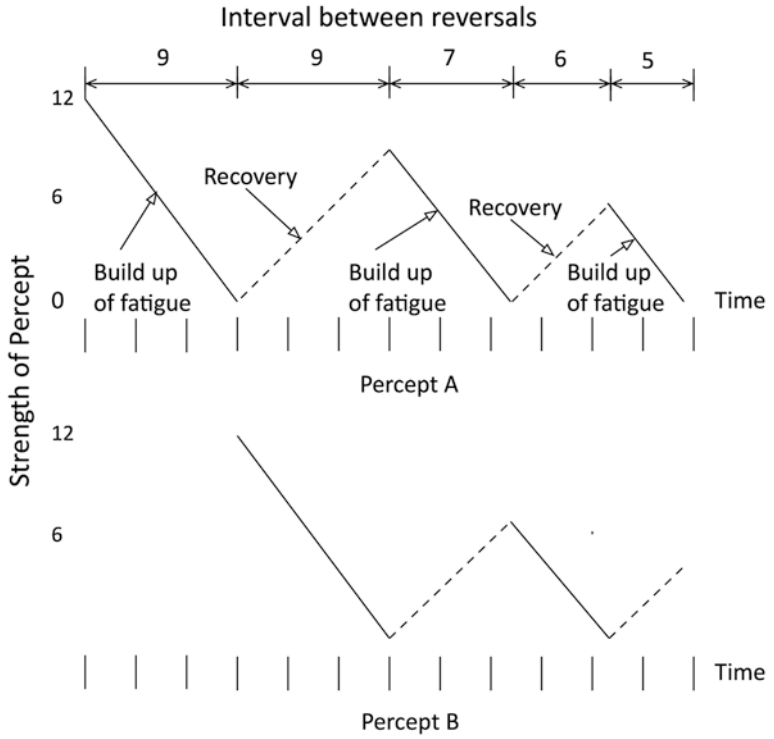


Fig. 3.7 If the fatigue rate and recovery rate are different, then the rate of alternation between the two percepts increases. In the figure, fatigue increases at four steps/time-unit while recovery reduces fatigue at the rate of three steps/time-unit. Percept A builds up to the satiation level over nine time-units where the strength of the percept drops to zero. The percept then switches to B that appears for nine time-units until its strength goes to zero. After Percept B satiates, Percept A reappears but has not recovered fully its former strength so that when it reaches the satiation level after only seven time-units and then Percept B reoccurs. This alternation continues and as the recovery level increases the interval between reversals decreases

Fig. 3.7. If one percept is made more prominent, the duration of the dominant percept does not decrease, but the duration of the weaker percept does.

A related phenomenon is that after a change in retinal location of the reversing figures the original longer intervals between reversals reappear. As stated above, continuous viewing leads to more rapid reversals, but if the same figure is moved to a new retinal position, the timing among reversals reverts to the original timing, as if the figure had never been seen. Fatigue and satiation appears to be limited to a restricted area of the retina.

A second phenomenon that supports peripheral processing (but also top-down processes) occurs when multiple figures are presented at the same time. If two Necker cubes are presented adjacent to each other, they reverse independently of one another (Figure 3.8A). This should not be a surprise as each cube

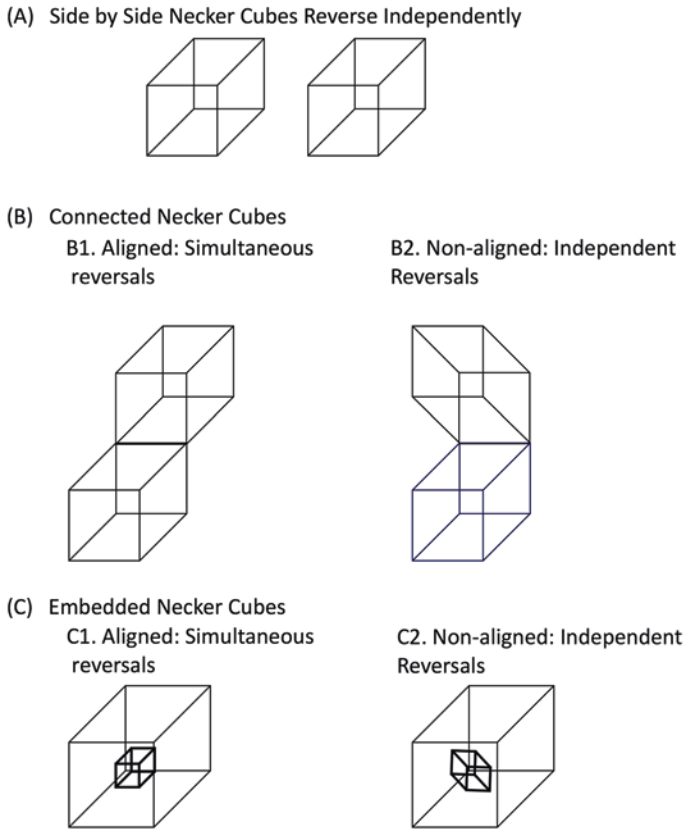


Fig. 3.8 The spatial position and alignment of two Necker cubes determines whether the reversals are independent. (Adapted from Adams & Haire, 1958, 1959)

is localized at a different retinal position and there is no connection between the two cubes. But, if the two cubes are connected, the cubes reverse together if they are aligned as in B1, yet reverse independently if they are not aligned, as in B2. In similar fashion, if an embedded cube is aligned with the larger one, reversals occur simultaneously (C1), but if the embedded cube is not aligned, then the smaller and large cube reverse independently (C2). The independent reversals in B2 and particularly C2 preclude an explanation based only on eye movements, but the simultaneous reversals in B1 and C1 suggest that parts of the visual field that appear grouped undergo the same reversals.

In spite of the large differences in procedures, stimuli, and the nature of the rivalry, the rate and pattern of visual and auditory alternations is identical, suggesting the same underlying neural processes (Brascamp, van Ee, Pestman, & van den Berg, 2005; Pressnitzer & Hupe, 2006). Analyses of the timing of the reversals suggest that the switches occur independently. What this means is that whatever the interval between one switch, the next switch interval could be longer or shorter with equal probability. There are no sequences of fast or slow

reversals except by chance. If the frequency of the intervals between switches is tabulated, the best-fitting probability distribution to model the shifts are based on the accumulation of small independent "micro-switches" between A and B until a threshold is reached and then the reversal occurs. Furthermore, if there were simultaneous visual and auditory presentation, for example, visual apparent movement and auditory stream segregation, reversals were independent. The visual reversals occurred at different times than the auditory reversals, with no central mechanism linking the two. These outcomes support the conclusions from Chap. 2 that multisensory cross-modality organization is not as strong as within modality organization.

A third phenomenon illustrating peripheral effects concerns tristable visual figures with three alternative percepts, such as the triangles in Fig. 3.1I that can point in three directions, The perceptual question is whether the three perceptual directions cycle, A-B-C-A-B-C, or undergo alternative unbalanced transitions such as A-B-A-C-A-B-A-C in which one percept serves as a "hub." Wallis and Ringelhan (2013) used a more complex display and found that cycling happens almost exclusively. They argued that the cycling maximizes the interval between the repetitions of any orientation thereby minimizing the fatigue of all the possible percepts

In sum, there is strong evidence that localized peripheral effects can bring about the reversals. Continuous viewing or listening to the same proximal stimulus fatigues a set of cells, so that the percept shifts to other non-fatigued cells that were previously firing at low rates. This shift in active cells gives rise to a different percept; however, the new cells become fatigued in turn so that the percept shifts once again to the firings of the set of now "recovered" original cells. There is reciprocal inhibition. But the peripheral processes do not explain all of the factors involved in the switches. Below we will describe some of the evidence that illustrates the influence of higher-level cognitive processes.

3.4.2 "Top-Down" Active Cognitive Control

The first phenomenon illustrating cognitive control is the ability to control the rate of reversals. If participants are asked to hold on to one percept, then their rate of reversals is roughly half that of passive attending. Conversely, if they are instructed to maximize the rate of alternation, the reversals occur five times the rate under passive attending, from one switch every five seconds to one switch every second. It is important to note that even when participants are asked to hold one percept, they are unable to do so; the alternative always appears. In some experiments (Kornmeier, Hein, & Bach, 2009) participants are asked to hold on to one of the two possible percepts, rather than either percept as above. In following these instructions, participants increase the duration of the primary percept, and tend to decrease the duration of the secondary percept.

A second phenomenon that illustrates top-down processing is that reversals occur only if the participants realize that there is an alternative percept. The participant's prior experience with the alternatives and their intention to reverse the figure are critical. For example, Fisher (1967) has created a set of ambiguous drawings that range from a clearly visible gypsy head with a barely visible girl with a mirror

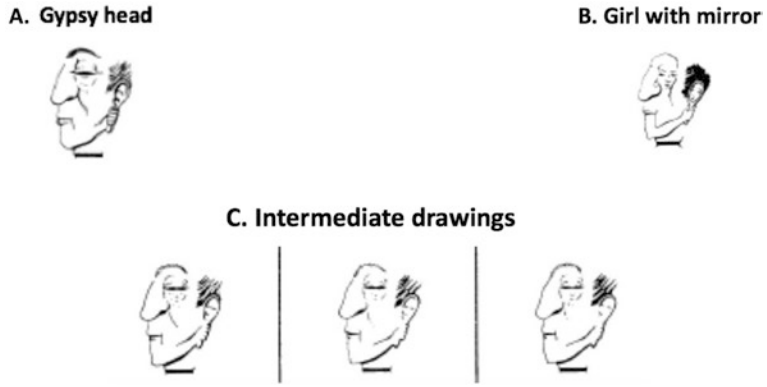


Fig. 3.9 The gypsy head (A) and girl with mirror (B) show the end points of the transformation. It is very difficult to perceive the alternative drawing without first seeing the other end point. But, after seeing both endpoints, it is easy to alternate between the two in the intermediate drawings in (C). (Adapted from Fisher, 1967)

at one end, to a clearly visible girl with mirror and barely visible gypsy head at the other end (Fig. 3.9A, B, & C). I found myself unable to reverse either end drawing without previously viewing a clear representation of the other end drawing.

A third phenomenon is that of set or expectancy. As described in the previous chapter, expectancy and prior learning heavily influence all aspects of perceiving, which the use of reversing figures can point out explicitly. After participants preview an unambiguous version of the figure for short time periods, they are far more likely to see that possibility when viewing an ambiguous version. However, longer viewing periods decrease the probability of seeing the previewed alternative, illustrating the effect of peripheral satiation.

Pastukhov, Vonau, and Braun (2012) have investigated reversals in the perception of rotating donuts. If the outline of the circumference of a donut is rotated around its vertical axis, a striking impression of the donut rotating in depth occurs. The direction of rotation often spontaneously reverses, from clockwise to counterclockwise and vice versa. These possibilities are illustrated in Fig. 3.10.

Pastukhov et al. (2012) were interested in the points at which the donut swapped the direction of rotation since logically the swap should occur at any angle with equal probability. However the majority of alternations occurred at the two orientations where the donut appeared depth symmetric, either flat or edge-on so that only a single surface was visible. The authors argue that the choice of those orientations reflect prior experience with the transitions between objects in the environment. To go from one object representation to another requires prior experience with the validity of that transition.

Finally, Ward and Scholl (2015) investigated whether fleeting cues, possibly unconscious, can bring about switches in the percept. The silhouetted spinning dancer can be seen rotating in either direction, but it is often difficult to reverse and for some individuals it never reverses at all. In the experiment, briefly presented short, white lines indicated which leg was in front as shown in Fig. 3.11, and following those cues there was an increase in the number of reversals.

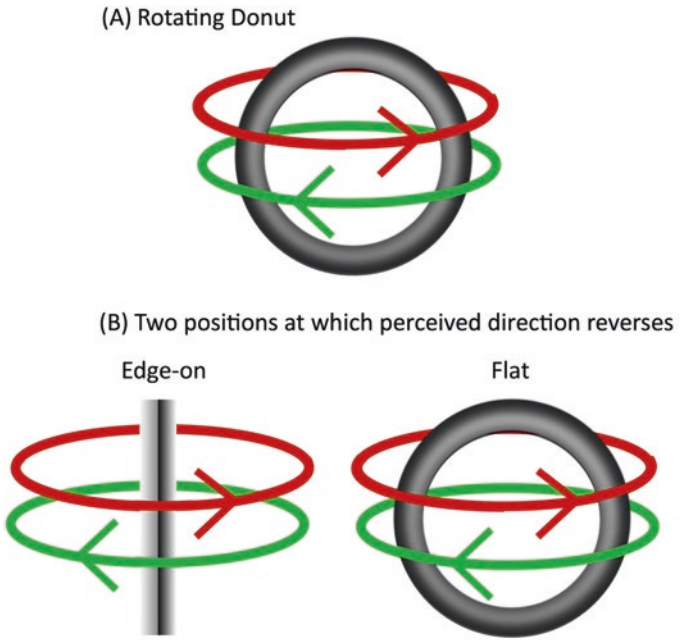


Fig. 3.10 The direction of rotation of the donut reverses spontaneously. There were two places at which the majority of reversals occurred: if the donut was edge-on and when the donut was flat

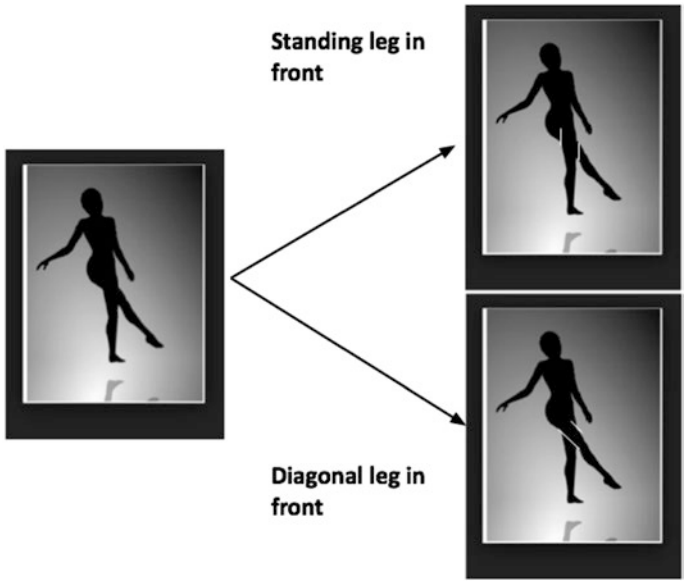


Fig. 3.11 Briefly flashing the white lines indicating which leg is in front can bring about a rotation switch even if the observer is unaware of the lines. (Adapted from Ward & Scholl, 2015)

Spinning Dancer
Wikipedia: http://en.wikipedia.org/wiki/File:Spinning_Dancer.gif
http://www.yale.edu/perception/dancer/

YouTube Videos
Improved 'solution' for spinning dancer girl illusion-alternating directions by Peter Wassink
Bistable perception by Mariushart

3.5 Summary

To return to the basic issue, many different distal objects could generate the same proximal image at the eye, ear, or hand and any distal object can generate many different proximal images. The major difference among perceptual theories lies in how this ambiguity is resolved. As described in the previous chapter, two theoretical positions have emerged. Gestalt theory emphasizes that percepts tend to be the simplest given the actual stimulation. Underlying this notion of *prägnanz* is the belief that the percept follows the operation of a uniform nervous system. Statistical theory on the other hand emphasizes that percepts tend to be the ones most likely to occur in specific situations and that people attend to the most reliable sensations, the Bayesian assumption. Perceptions that originally required conscious calculations become rapid, accurate, and unconscious with practice. As is clear, it is often difficult to distinguish between these two alternative explanations, difficult to define simplest or coherent, and difficult to derive a catalog of all the possible objects in the environment. It is extremely rare for the exact same sensation to occur twice.

As in also clear, no single mechanism is responsible for shifting perceptions, and the interaction among neural and cognitive processes, which of course are also neural, has been a constant theme. Fluctuations in neural responses at both anatomically early and late processing are strongly correlated with the recurrent percepts (Sterzer, Kleinschmidt, & Rees, 2009). But it is unclear whether these fluctuations actually cause the reversals. The dominant stimulus at any time point flows through the nervous system linking with all the connotations of the stimulus. The alternative but suppressed stimulus is disrupted at several points in the pathway although patterns of activity continue to carry information about that percept. There is constant feedback; the present perceptions guide actions that alter perceptions that guide further actions.

We can imagine a simple model for multistability. Suppose there are two alternative percepts, each of which can be represented as a depression in a perceptual energy module, as in Fig. 3.12. There is a barrier between the percepts, and energy is necessary to shift to the other depression and thereby change the percept. This energy comes from a series of small random inputs from different parts of the cortical tracts such as neural satiation, switches in viewing orientation, expectancies, attention shifts, inputs from other senses, and so on. It is a mixed bag of inputs, bottom-up and top-down, some of which seem to lower

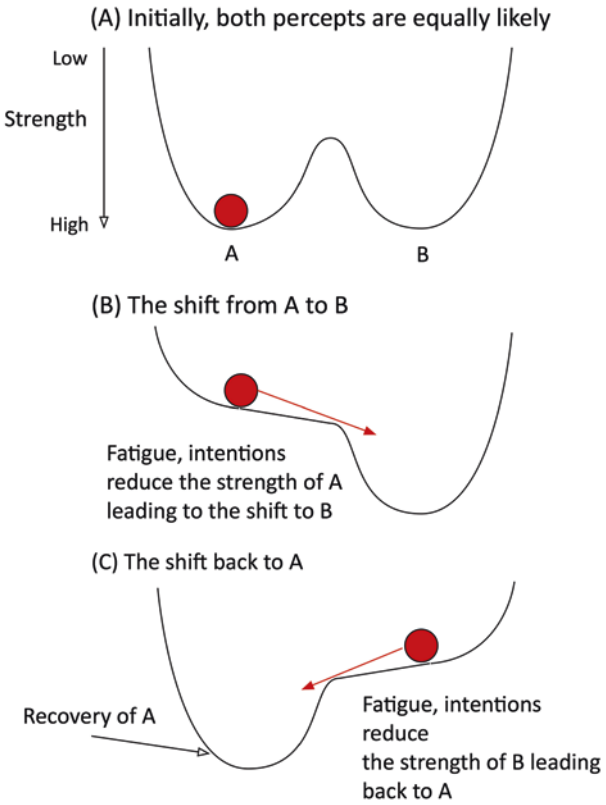


Fig. 3.12 Initially, both percepts could be equally likely based on the depth of the depression. The barrier between the two maintains the initial perception. If the inputs to Percept A reduce the depth of the depression so that it is higher than the barrier, the percept will shift to B. Then, inputs will reduce the depth for Percept B, and at the same time Percept A will recover its strength. The height becomes greater than the barrier and that results in a reversal back to Percept A

the barrier and others which seem to lift the percept. On top of this we should expect each to have different timings between inputs. The percept shifts when the sum of these random inputs reaches a critical value so that the depression becomes higher (i.e., becomes weaker) than the barrier. Fluctuations in the inputs can thus account for the overall randomness of the oscillation between the percepts (Braun & Mattia, 2010).

Models like this are found in several domains. For example, ecologists use such models to explain the shift from one state (e.g., algae-infested turbid water) to another state (clear water), that is, tipping points, (Popkin, 2014). This explanation reinforces the argument from Chap. 1 that perceptions result from the interaction of many neural processes. There is no "single actor" who is in charge (Fig. 3.12).

References

Adams, P. A. A., & Haire, M. (1958). Structural and conceptual factors in the perception of double-cube figures. *American Journal of Psychology, 71*(3), 548–886. https://doi.org/10.2307/1420250

Adams, P. A. A., & Haire, M. (1959). The effect of orientation on the reversal of one cube inscribed in another. *American Journal of Psychology, 72*(2), 296–299. https://doi.org/10.2307/1419384

Alais, D., & Blake, R. (2015). Binocular rivalry and perceptual ambiguity. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 775–798). Oxford, UK: Oxford University Press.

Brancucci, A., & Tommasi, L. (2011). "Binaural rivalry": Dichotic listening as a tool for the investigation of the neural correlates of consciousness. *Brain and Cognition, 76*, 218–224. https://doi.org/10.1016/jbandc.2011.02.007

Brascamp, J. W., van Ee, R., Pestman, W. R., & van den Berg, A. V. (2005). Distributions of alternation rates in various forms of bistable perception. *Journal of Vision, 5*, 287–298. https://doi.org/10.1167/5.4.1

Braun, J., & Mattia, M. (2010). Attractors and noise: Twin drivers of decsions and multistability. *NeuroImage, 52*, 740. https://doi.org/10.1016/j.neuroimage.2009.12.126

Carter, O., Konkle, T., Wang, Q., Haywood, V., & Moore, C. (2008). Tactile rivalry demonstrated with an ambiguous apparent-motion quartet. *Current Biology, 18*, 1050–1054. https://doi.org/10.1016/j.cub.2008.06.027

Deutsch, D. (1974). An auditory illusion. *Nature, 251*, 307–309.

Fisher, G. H. (1967). Measuring ambiguity. *American Journal of Psychology, 80*(4), 541–557.

Fisher, G. H. (1968). Ambiguity of form: Old and new. *Perception & Psychophysics, 4*(3), 189–192.

Freeman, E., & Driver, J. (2008). Direction of visual apparent motion driven solely by timing of a static sound. *Current Biology, 18*, 1262–1266. https://doi.org/10.1016/j.cub.2008.07.086

Gepshtein, S., & Kubovy, M. (2007). The lawful perception of apparent motion. *Journal of Vision, 7*(8), 9, 1–15. https://doi.org/10.1167/7.8.9

Harrar, V., Winter, R., & Harris, L. R. (2008). Visuotactile apparent motion. *Perception & Psychophysics, 70*, 807–817. https://doi.org/10.3758/PP.70.5.807

Klink, P. C., van Wezel, R. J. A., & van Ee, R. (2012). United we sense, divided we fail: Context-driven perception of ambiguous visual stimuli. *Philosophical Transactions of the Royal Society B, 367*, 932–941. https://doi.org/10.1098/rstb.2011.0358

Kornmeier, J., Hein, C. M., & Bach, M. (2009). Multistable perception: When bottom-up and top-down coincide. *Brain and Cognition, 69*, 138–147. https://doi.org/10.1016/j.bandc.2008.06.005

Korte, A. (1915). Kinematoskopische Untersuchungen (Cinematoscopic investigations). *Zeitschrift fur Psychologie, 72*, 193–216.

Lakatos, S., & Shepard, R. N. (1997). Constraints common to apparent motion in visual, tactile and auditory space. *Journal of Exerimental Psychology: Human Perception and Performance, 23*, 1050–1060.

Liaci, E., Bach, M., van Elst, L. T., Heinrich, S. P., & Kommeler, J. (2016). Ambiguity in tactile apparent motion perception. *PLoS One, 11*(5), e0152736. https://doi.org/10.1371/journal.pone.0152736

Long, G. M., & Toppino, T. C. (2004). Enduring interest in perceptual ambiguity: Alternating views of reversible figures. *Psychological Bulletin, 130*, 748–768. https://doi.org/10.1037/0033-2909.130.5.748

Macnik, S. L., Martinez-Conde, S., & Blakeslee, S. (2010). *Sleights of mind*. New York, NY: Henry Holt.

Occelli, V., Spence, C., & Zampini, M. (2010). Assessing the effect of sound complexity on the audiotactile cross-modal dynamic capture task. *The Quarterly Journal of Experimental Psychology, 63*, 694–704. https://doi.org/10.1080/17470210903068989

Pastukhov, A., Vonau, V., & Braun, J. (2012). Believable change: Bistable reversals are governed physical plausibility. *Journal of Vision, 12*, 1–16. https://doi.org/10.1167/12.1.17

Popkin, G. (2014, September 26). On the edge. *Science, 345*, 1552–1554.

Pressnitzer, D., & Hupe, J. M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current Biology, 16*(13), 1351–1357. https://doi.org/10.1016/j.cub.2006.05.054

Rock, I. (1983). *The logic of perception*. Cambridge, MA: MIT Press.

Sanabria, D., Soto-Faraco, S., & Spence, C. (2005). Assessing the effect of visual and tactile distractors on the perception of auditory apparent motion. *Experimental Brain Research, 166*, 548–558. https://doi.org/10.1007/s00221-005-2395-6

Soto-Faraco, S., Spence, C., & Kingstone, A. (2004). Congruency effects between auditory and tactile motion: Extending the phenomenon of cross-modality dynamic capture. *Cognitive, Affective, & Behavioral Neuroscience, 4*, 208–217.

Sterzer, P., Kleinschmidt, A., & Rees, G. (2009). The neural bases of multistable perception. *Trends in Cognitive Science, 13*, 310–318. https://doi.org/10.1016/j.tics.2009.04.006

Thossen, S., & Bendixen, A. (2017). Subjective perceptual organization of a complex auditory scene. *Journal of the Acoustical Society of America, 141*, 265–276. https://doi.org/10.1121/1.4973806

Wallis, G., & Ringelhan, S. (2013). The dynamics of perceptual rivalry in bistable and tristable perception. *Journal of Vision, 13*, 1–21. https://doi.org/10.1167/13.2.24

Ward, E. J., & Scholl, B. J. (2015). Stochastic or systematic? Seemingly random perceptual switching in bistable events triggered by transient unconscious cues. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 929–939. https://doi.org/10.1037/a0038709

Warren, R. M. (1999). *Auditory perception: A new analysis and synthesis*. Cambridge, UK: Cambridge University Press.

Warren, R. M., & Gregory, R. L. (1958). An auditory analogue of the visual reversible figure. *American Journal of Psychology, 71*, 467–473. https://doi.org/10.2307/1420267

CHAPTER 4

Rhythm and Timing

4.1 Introduction

Rhythm and timing are inherent in so many aspects of perception, and rhythmic behavior is so common and easy that there does not seem any need to create a framework to understand it. Yet, it is not easy to create an adequate one. We construct our perception of time, and perhaps it is based on the succession of events.

One method is to study the perception of simple auditory rhythms directly. The traditional approach is to consider the perception of auditory rhythms as a cognitive problem and attempt to abstract the rules underlying their organization. At the simplest level, we have the perception of simple isochronous rhythms made up of different sorts of sounds. Alternatively, we have the perception of rhythms based on different timings between the onsets of identical adjacent sounds (for rhythms, it is the onset-to-onset interval that is critical, not the offset-to-onset interval critical for stream segregation). Combining the sound and timing differences creates rhythms that approach those found in music. The understanding of these simple rhythms would be clearly based on the Gestalt principles of organization described in Chap. 2.

An alternate approach to rhythm perception is to enlist the accompanying motor sensations and consider the interplay between the two. This has been termed *embodied rhythm perception* (Maes, Leman, Palmer, & Wanderley, 2014). People tap their feet, clap in time, and sway to rhythms, and these movements are indubitably tied to our perception of those rhythms. By considering those movements timed to the *beat*, it is possible to understand the expressiveness of different rhythms. For physical actions there are the rhythms of hands, arms, bodies, and other individuals that move at different tempos. In addition, there

Electronic Supplementary Material: The online version of this chapter (https://doi.org/10.1007/978-3-319-96337-2_4) contains supplementary material, which is available to authorized users.

S. Handel, *Perceptual Organization*,
https://doi.org/10.1007/978-3-319-96337-2_4

are the multiple timings involved in all kinds of physical activities, ranging from hitting a tennis ball to jumping over a ditch. It is a mistake to argue that rhythm occurs at one level, when each rhythmic level supports the others. The emergent rhythm is due to the interplay of rhythms at different levels.

A second method is to consider the ways that rhythms slice up the continuous ongoing sensations of music, speech, and vision into units of varying duration. For music, there are the rhythms of individual notes, groups of notes, contours of notes, and themes. For speech, there are the rhythms of syllables, words, and sentences. It seems that speech is broken initially into longer, slower units at roughly five syllables/sec, and then the acoustic variations within those units are analyzed at a higher frequency to identify the particular syllable. For vision, timing will affect apparent movement as well as other multistable movements; the timing between the positions of the individual lights in point light displays described in Chap. 2 (Johansson, 1975) determines which motions are perceived. Visually, we perceive smooth motion in movies being projected at 24 frames per second not on account of a sluggish visual system. A sequence of random movements in successive frames is seen as a jumble of unconnected flashes. It is the contour of the movements, rather, that leads to smooth motion. At a slower frame rate, the motion looks jerky once again; it is the combination of the space and time contours that are critical.

Still another method is to stretch concepts like rhythm and timing metaphorically to the visual arts. We can think of rhythms as creating a set of slots into which notes can be placed. In Western music at least, the slots are equally spaced, and as explained later, subsets of more widely separated but still equally spaced slots receive accents. These accents create expectancies that important notes will occur at particular rhythmic intervals and lead listeners to hear the note sequence in synchrony to the accents (Jones, 1976; Large & Jones, 1999). In vision, we can think of space as being filled with a recurring spatial grid, and designs being created by "darkening" some of the grid lines or spaces (DeLong, 2013; Tetlow, 2013). By varying the size and type of grid, it is possible to create a bewildering assortment of designs, whose repetitions in many instances are analogous to the repeating nature of the auditory rhythms. Understanding visual rhythms entail the same principles as the Gestalt approach to auditory rhythms.

At stated above, what is common to all of these aspects of rhythm is the notion that rhythmic organization occurs at several levels simultaneously. Neural codes operating on different time scales might encode complementary stimulus features. We may use the longer intervals of slow cortical network oscillations to create phrase information, namely, the start and end of complex events, and simultaneously use shorter intervals of faster oscillations to analyze the variability within those segmented events. The perceived rhythm comes from the interaction between the timing of the sounds and the internal timing of the listener. Multiple timings allow us to attend to expected events, but it also allows us to react to unexpected changes in the timing of events. It ensures flexibility.

4.2 Auditory Temporal Rhythms

Before starting it is important to distinguish between the physical and phenomenal meanings of the term *rhythm*. We can define the rhythm in terms of the acoustical characteristics of the tonal sequence: the frequency, duration, intensity, and timbre of the individual tones and the timing between the onsets of adjacent notes. We can also define rhythm in terms of the perceptual response. Clapping, finger tapping, and/or swaying to the beat show the perceived rhythm. Finally, we can define rhythm in terms of the notated musical score that indicates the relative timing of the notes. The physical, phenomenal, and notated rhythms may not have a simple match. That is, the beat of a rhythm can occur on an actual tone, but the beat can also fall on a silence.

It seems appropriate to focus on the experience of rhythm. Rhythm is inside us and arises in a specific context. We can specify the physical acoustical characteristics of the sounds that support the experience of rhythm, but at every level we must relate those characteristics to the phenomenal and muscular feelings of listeners to the experienced rhythm. There is no single component of the acoustical signal that can uniquely predict the felt rhythm.

The experience of rhythms involves the feeling of regularity and grouping among the sounds as well as the accentuation and differentiation of those sounds within the groups. Weaker elements are attached to stronger accented ones creating a feeling of regularity among the stronger sounds that influence body motions. The obvious periodic physiological process such as breathing, heartbeats, and walking initially led researchers to believe that such physiological responses underlay the perceptual response. The tempo (i.e., the rate of the events) of heartbeats or breathing rates was thought to set a baseline, so that the tempo of rhythms above that rate would be described as fast and ones below were thought to be sluggish. It soon became quite clear that although physiological changes and movements accompanied the perception of rhythms, they were not the cause of that perception.

Yet, physiological processes do set limits on perceivable rhythms and the ability to synchronize. When synchronizing, listeners anticipate the next sound and respond before it appears; they are not simply reacting to the appearance of that sound and then responding. The critical sense of rhythmic regularity disappears as the onset-to-onset interval approaches 100 msec (10 elements/sec), and this is about the fastest tapping rate that individuals can maintain when synchronizing to a sequence of sounds. Beyond this rate, individuals are even unable to determine if they are synchronized to the sounds. (The 100 msec interval is also the minimum interval between musical notes). At the other extreme, as the onset-to-onset interval approaches 2–2.5 sec, the sense of regularity again disappears and the sounds appear to be discontinuous and isolated. That interval is about the longest that allows for successful synchronization. Synchronization to flashing lights is far more restricted at the faster presentation rates. Individuals are unable to synchronize when the presentation rate is faster than 2/sec, although the limitations at the slower rates are the same as those to audition.

While these limits on rhythmic perceiving and synchronizing are quite broad, the onset-to-onset intervals between 400 and 600 msec (roughly two elements/sec) seem natural and preferred for rhythms. Individuals spontaneously tap at these rates and can more accurately reproduce intervals in this range.

In what follows, we will initially concentrate on the rhythms in Western European music with a single regular beat. These follow a strong-weak beat or strong-weak-weak beat sequence that allows listeners to tap or march along in synchrony. It is a simple rhythmic structure, but it allows listeners to keep in time when the rhythm speeds up or slows down for expressive purposes. The 1:2 and 1:3 timing ratios seem natural, easy to produce, and are preferred across a wide range of musical cultures. The steady beats allow individuals to synchronize in groups; music that is performed solo often does not have a steady beat. Moreover, many movements and motions follow the same strong-weak alternation, so that musical rhythms may be just one example of a more fundamental pattern. Next we will study polyrhythms that contain two or more rhythmic lines so that listeners can select one of them to act as the beat. Finally, we will briefly consider non-regular rhythms that are found in many musical traditions.

4.2.1 Isochronous Pulse Trains

The simplest rhythmic stimulus consists of a series of equally spaced, identical sounds. Early work found that even without physical differences, the sounds grouped into units of two, three, or four depending on their tempo. Listeners are able to vary their tapping rate, at faster rates creating groups of four or eight sounds, or at slower rates creating groups of two or even one element. Even without a physical rhythm, the phenomenal rhythm exists at several levels. In all cases, the first note of the group appears stronger and accented, and the intervals within a group appear shorter than the intervals between the final element of one group and the initial element of the next group as shown in Fig. 4.1A, B & C (Bolton, 1894; Woodrow, 1909). These were termed "subjective rhythms" due to the strong perceptual rhythm in spite of no differences among the elements. (In reality, all rhythms are subjective).

Subsequent research, although keeping the onset-to-onset intervals identical, made the sequences more complex by varying the physical characteristics of some of the elements. If every second, third, or fourth element is made louder, then that louder element is perceived to be the first element of the groups (Fig. 4.1D & E). The intervals within the groups are perceived to be shorter than the intervals between the final softer note of the group and the initial louder sound of the next group. In similar fashion, if every second, third, or fourth sound is increased in duration with the offset-to-onset interval remaining constant, then the longer duration sound is perceived to be the last sound of the groups (Fig. 4.1F & G). Again, the intervals within groups are perceived to be shorter than the intervals between the final longer duration sound of one group and the initial shorter sound of the next group. Finally, if one of the onset-to-onset intervals is physically lengthened, the groups are created by that interval

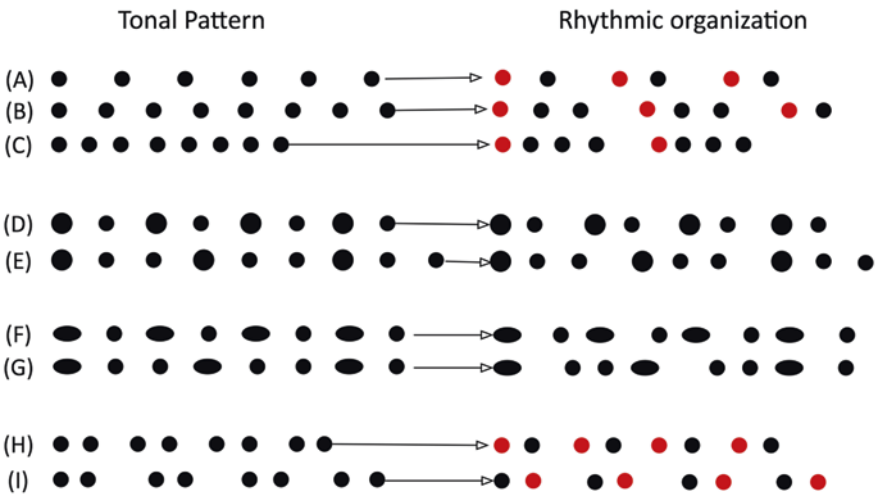


Fig. 4.1 The perception of simple rhythms. Isochronous sequences of identical elements are organized into groups of two, three, or four depending on the rate of the elements (A, B, & C). The initial elements are perceived as accented shown in red. Moreover, the elements seem to be grouped in time so that the interval between the last element in the group and the first element in the next group is perceived to be longer than intervals within the group. If every second or third element is louder, then that element is perceived to start the group (D & E), but if those elements increase in duration, they are perceived as the last element in the group (F & G). As the interval between groups of elements increases, the accent shifts from the first element (red in H) to the last one (red in I)

Sound Files 4.1: Rhythmic grouping created by intensity, duration, and inter-element interval differences

(Fig. 4.1H & I). If the difference between the two onset intervals is small, the first element of the group is perceived to be louder, while if the difference is large, the last element of the group is perceived to be louder.

Although these outcomes seem too simple to explain the complex rhythms of music and speech, I think they form the backbone for understanding those rhythms in much the same ways that Gestalt principles form the backbone for understanding complex visual scenes. In fact, the rhythmic principles are quite analogous to the Gestalt ones. It is so commonplace that it is easy to miss that people spontaneously organize sounds (and visual or tactual) elements into rhythmic groups, which become figures against a homogeneous background of time. The rhythm creates a sound source. In this process, grouping reorganizes the entire sequence. If the intensity of some sounds is increased, it brings about a change in the perception of the intervals between the sounds. Conversely, if the interval itself is changed, it brings about a change in the perception of the intensity of the sounds. Rhythm is relative timing, as the timing and accentuation of any sound is determined relative to the timing and accentuation of all the other sounds; often there is no simple correspondence between the acoustic signal and the perceived rhythm. The perception of an accented sound may be the result of an acoustic change at a distant point in that sequence.

4.2.2 Beats and Meters

As described above, listeners hear stronger accented elements even in isochronous pulse trains. The accents tend to occur at every two, three, or four elements and, at least for Western listeners, occur at equal intervals. Moreover, listeners can alter their tapping rate at will, tapping rapidly every two elements, more slowly every four elements, and slowest every eight elements. (Within limits, people can also vary the limb used to tap the rhythm). This ability to hear a sequence at different time intervals yields a hierarchical representation that results in the perception of beats and meters.

Beats refer to that sense of equally spaced accented elements and usually occur at the onset of tonal elements. However, they can also be felt without an actual element. Meter is the sense of a periodic sequence of subjectively stronger and weaker beats that characterize music; it arises from the beats at different time scales. Meter creates a sense of temporal regularity that tells the listener when to expect the next stronger beat. Dancing, tapping, and marching are all timed to the meter. There are faster meters at lower levels in the hierarchy in which beats occur on every element, or on every two or three elements, and slower meters at higher levels in which beats occur every four, six, or eight elements. Typically, lower pitches are used to play the slower beats, and higher pitches are used to play the faster beats. At each level the beats are presumed to be equal so that it is the addition of the beats across the levels that give rise to the strength of each element. Without the faster beats, the slower beats are simply recurring accents, and without the slower beats, the faster ones also are simply recurring accents. Meter is the mental construct that fuses the beats at different levels together and organizes time.

Any theory of rhythm must explain why some elements become strong beats and why some become weak beats in order to account for the sense of meter. Furthermore the theory must account for the grouping of the beats, which strong beats and weak beats go together. To untangle the two components of meter, beat strength and grouping, we start by considering the grouping of the elements and how that feeds into the perceived location of the beats. The grouping process brings about the sense of strong and weak accents, which are integrated to form the beat. The accents may not occur at the equal intervals of the beat so that there is a constant back and forth between the group and beat organizations.

4.2.3 The Grouping Hierarchy

Many factors determine the partitioning of a sequence of elements into groups. In nearly all instances, each element is placed in one and only one group (as is true for all kinds of perceptual organizations as detailed in the previous chapters), although in many real cases a single element may seem to bridge two groups.

1. First, there is the principle common to all perceptual and cognitive processes to place elements into equally sized groups and to avoid groups of just one element. For the isochronous sequences, the first element of groups of two to four elements is perceived to be accented (Sound Files 4.1A, B, C).
2. Second, for non-isochronous sequences composed of identical elements, the longer onset-to-onset intervals between adjacent elements split the elements into groups. Povel and Essens (1985) and Povel and Okkerman (1981) propose several rules to account for the perceived accents in such sequences: (a) accents would occur on isolated elements; (b) accents would occur on the first element of a two-element group if the temporal interval was short, but on the second element if the interval was longer (as in Fig. 4.1H versus 4.1I and Sound File 4.1H versus 4.1I); and (c) the first and last elements would be accented if the groups consisted of three or more elements. Consider the following 16-unit rhythm (the rhythm would be recycled at the end). The x's represent sound and the dots represent silences equal in duration to the sounds. Applying these rules to the same sequence yields the following accents shown in capitals:

   PHYSICAL RHYTHM x x x . . x . . x x . . . . x x . . . .
   GROUPING RHYTHM X x X . . X . . X x . . . . x X . . . .

3. Third, for isochronous and non-isochronous sequences composed of differing elements, groups tend to occur at points of change. Changes in timbre, frequency range, intensity, and/or tempo may act to slice the sequence into discrete groups in a variety of musical genres, e.g., Turkish makam music investigated by (Mungan, Yazici, & Kaya, 2017). As described in Fig. 4.1, in sequences composed of different intensities, louder elements begin groups and the interval between the prior softer and louder next element is heard as longer than intervals among the softer elements (Fig. 4.1D & E and Sound Files 4.1D & E). For sequences composed of differing frequencies, listeners expect adjacent tones to be close in frequency, so that large frequency shifts lead to the perception of group boundaries at the shift (Fig. 4.2A). In similar fashion, reversals in direction of the frequency changes lead to the perception of a group boundary at the turn-around (Fig. 4.2B). Another general rule is that groups are formed from repeating sequences, regardless of the nature of the repeats (Fig. 4.2C). Every musical culture contains repeating segments and the repetitions allow the listener to look forward in the music. Finally, one more cue to a group boundary is the increase in duration of a tone at the end of phrases (Fig. 4.2D).

Vimeo https://vimeo.com/120517523, Video discussing the importance of melodic repetition.

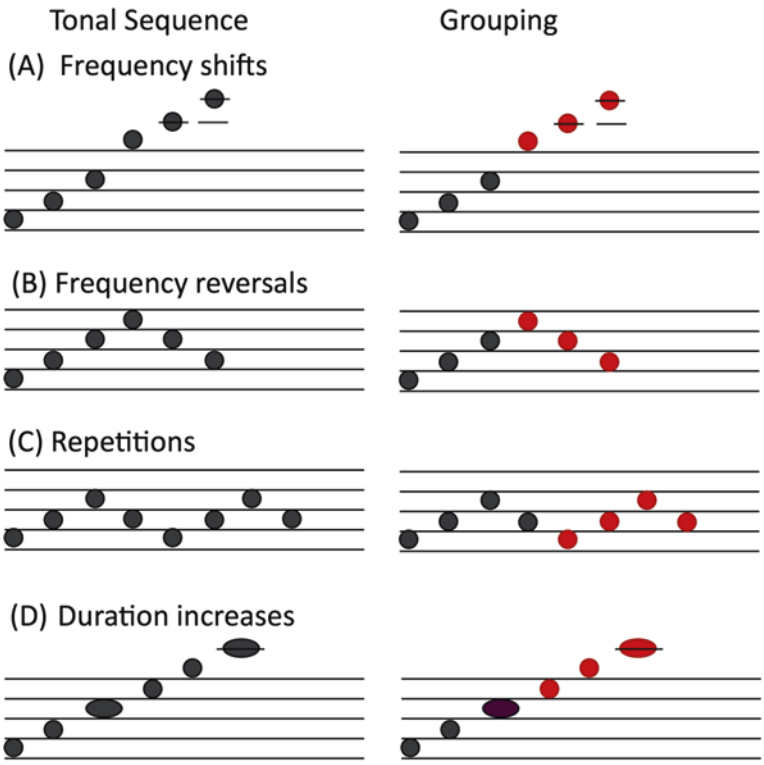


Fig. 4.2 Frequency shifts, frequency reversals, repetitions, and duration lengthening are four properties of tonal sequences that lead to the grouping of the tones (black versus red notes)

Sound Files 4.2: Tonal grouping due to frequency shifts and reversals, repetitions, and tonal duration lengthening as shown in Fig. 4.2A–D

In non-musical settings, repetition can aid the identification and localization of the source. In many species, repeating sequences help maintain contact and cohesion among individuals within a social group and may even serve to identify the social group itself (Zwamborn & Whitehead, 2017). For example, when male humpback whales congregate to mate in warm water, each whale in the region sings pretty much the same song. The song does change each season, but the change is contagious and each whale soon begins to sing the same new song. If you put a hydrophone in the water there is a cacophony of whale sounds, so that if a male is going to successfully attract a female its song must be easy to locate. The repetition reduces the masking of the song due to the background noise (Brumm & Slater, 2006). The first factor that helps identify the same whale and its location is the repetition of a set of sounds, termed a phrase. The second factor is that the sounds in successive phrases are nearly identical, so that the song evolves slowly. An example is shown in Fig. 4.3. The first phrase is the simple alternation RA and that phrase may be repeated 10–20 times. Invariably, the next phrase RRCCC doubles the same R sound and combines it with three

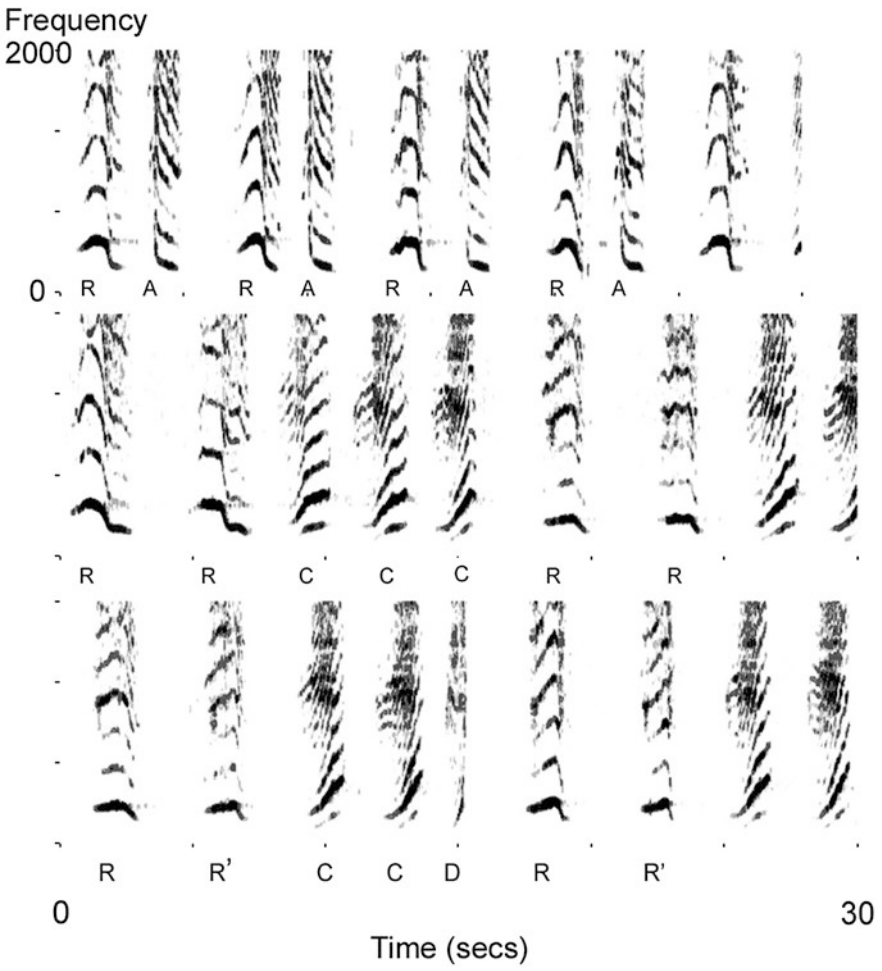


Fig. 4.3 Three successive phrases from a humpback whale recorded in Hawaii during 1994. (Adapted from Handel et al., 2012)

Sound Files 4.3: Three phrases in the humpback whale song (Fig. 4.3)

repetitions of the C sound. This phrase is repeated two or three times and it then evolves into the RR'CCD in which the second R sound is slightly modified and the third C sound is replaced by the D sound. The same pattern occurs in all parts of the song: each phrase is repeated and successive phrases substitute new sounds for some of the old ones (Handel, Todd, & Zoidis, 2012). (I think this minimizes the memory load for the whale, but that is another story.) Sound files 4.3A,B, and C give an example of each phrase.

Given the many factors that affect grouping, it is no wonder that although some sequences are invariably grouped in the same way, others lead to a wide variety of outcomes. Nonetheless, whatever the resulting grouping, it should meet several criteria: (1) only adjacent notes can be placed in a group; (2) no

note can be skipped; and (3) with rare exceptions, any note can only be placed in a single group. In similar fashion, in the hierarchical organization of the groups, any longer group can be split at lower levels, but only adjacent groups can be combined at higher levels.

4.2.4 The Meter Hierarchy

4.2.4.1 Single Rhythm Lines

The splitting of the element sequence into groups creates the pattern of perceived accents and timing. The construction of the meter can be thought as fitting the regular timing of the strong-weak or strong-weak-weak beats to the grouping. The "best" meter would happen when the periodic beats coincide with the onsets of those accented elements. Thus, the beats should occur on the more intense elements, the longer duration elements, on the initial element of large frequency shifts or reversals, and on the initial element of groups of repeating melodic phrases. Needless to say, it is all but impossible to satisfy all of these preferences. Fitting a meter is nearly always a compromise.

A useful way of imagining fitting the metric hierarchy to the grouping pattern of the sequence of elements is by using a hierarchical timing grid based on a series of isochronous dots so that each represents beat strength. At the highest level (the fastest) of the hierarchy, each dot represents the timing of the shortest element, for example, an eighth or sixteenth note in musical terms, so that at this level each note or silence has equal beat strength. At the next level down, the isochronous dots are aligned with every second or third beat of the lowest level. To create a strong-weak duple meter every second beat is represented by one timing dot to represent its added metric importance. To create a strong-weak-weak triple meter, every third beat is represented by a timing dot to represent its added importance. At the third level down, the isochronous dots are aligned to represent every fourth (duple) or sixth beat (triple). The timing dots at this level further show the relative strength of each beat. At lower levels, the isochronous timing dots again and again split the dots at higher levels in half to give the overall strength of each beat. The depth of the stack of dots indicates the strength of the beat.

We can start with nine-element patterns embedded in a 12-element repeating unit depicted in Fig. 4.4. Pattern A is xxx-xxx-xxx- where x represents an element and a dash represents a silent interval equal to an element in duration is shown in (A). (To be more specific, if we present the pattern at a comfortable rate of four elements per sec (i.e., 250 msec between element onsets), an element might be 125 msec long with a silent offset-to-onset interval between adjacent elements of 125 msec, to create the full 250 msec onset-to-onset interval between elements. The silent interval would therefore be 375 msec, composed of the 125 msec silent offset-to-onset interval following an element plus the 250 msec silent onset-to-onset interval). For this pattern, the stronger accents would fall on the first element of each run of three elements, and a weaker accent would fall on the last element of each run of three elements **X**x**X**-**X**x**X**-**X**x**X**-. Given these accents,

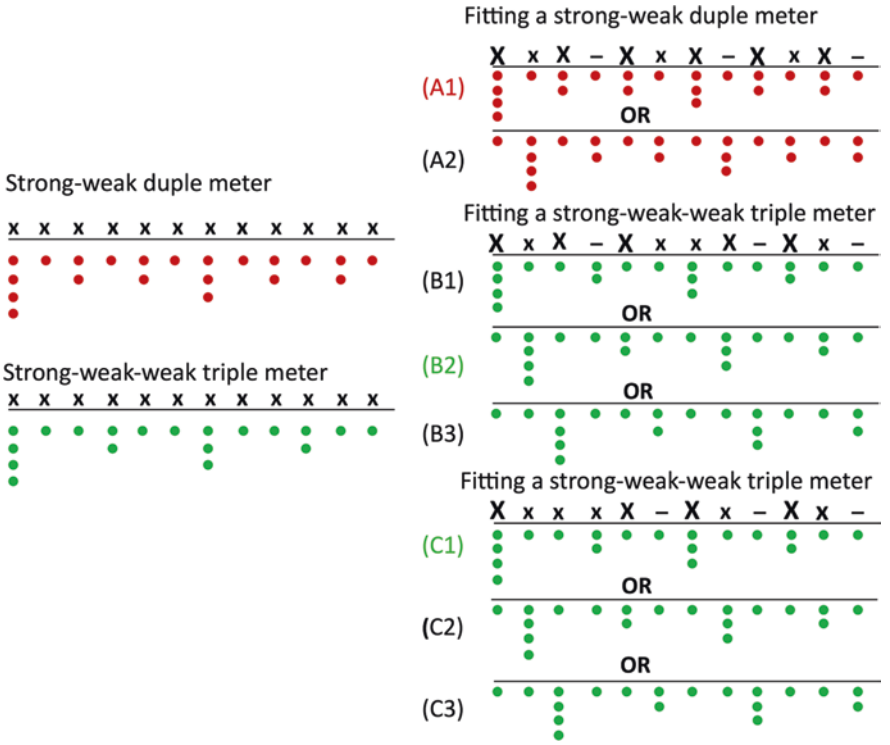


Fig. 4.4 The beat hierarchy of strong-weak (red) and strong-weak-weak meters (green) is illustrated for repeating 12 element patterns. The two alternative strong-weak beat meters are shown for the nine sounds plus three silences in a 12-element pattern xxx-xxx-xxx- in (A1) and (A2). The three alternative strong-weak-weak meters are shown for xxx-xxxx-xx- in (B1, B2, & B3) and xxxxx-xx-xx- in (C1, C2, & C3). In all cases, the goal is to align the strongest beats shown as the depth of the hierarchy with the first element of a run of two elements and the first and last element of a longer run. The best fits are in color: (A1 red), (B2 green), and (C1 green)

the best-fitting meter would be a strong-weak duple meter starting on the initial element of each run (see A1 in Fig. 4.4). This positioning would optimize the next level down in the hierarchy because each strong-weak-strong-weak beat would consist of the run of three elements ending on a silent interval (xxx-). If the duple meter was shifted one element to the right (see A2 in Fig. 4.4), half of the strong-weak duples would begin on a silent interval. A triple strong-weak-weak meter will not work because wherever the meter starts the beats will fall at least once on a silent interval.

Even moving the position of one element can drastically change the meter. Consider Pattern B, xxx-xxxx-xx-, that still contains nine elements embedded in a 12-element repeating unit. The accents would be XxX-XxxX-Xx-. The optimal duple meter found above simply does not work as well here because the fifth strong beat would fall on a silent interval. Actually, there is no possible

shift for the strong-weak meter that does not result in a strong beat falling on a silent interval. A triple strong-weak-weak meter can result in the strong beats falling on actual elements if the meter starts at the second element of the initial run of three elements (see B2 in Fig. 4.4). However, this meter is not optimal for two reasons; first, the beat falls on the middle element not accented in a run of three elements, and second, at the next hierarchical level the six-element meter breaks the runs apart.

Another one element move can shift the best meter again. For Pattern C, xxxxx-xx-xx-, the accents would be XxxxX-Xx-Xx-. The best meter placement is again a triple meter starting on the initial element of the pattern (C1). No other triple or duple meter works as well in matching the accents to the beats. The amplitude × time representations of the rhythms in Fig. 4.4 are pictured in Fig. 4.5, and the sound files are found in Sound File 4.5.

The meter controls the perception of the entire pattern. Povel and Essens (1985) presented single-note patterns and accompanied the pattern either with a drumbeat every three elements (triple meter as in Sound File 4.6D1) or every four elements (duple meter as in Sound File 4.6D2). Listeners were unable to recognize that the patterns were identical when accompanied by the different metric drumbeats. What is important to understand is that the fitting of a meter to a grouping of the elements is usually a compromise between the best grouping and the best meter. It is rare that they perfectly correspond and it may take a while for the meter to emerge (Fig. 4.6).

We can see the fit between the grouping and meter structures for simple songs such as Stephen Foster's "Way Down Upon The Swanee River." In the first phrase, the shortest note is an eighth note, so that the lowest (fastest) level of the metric hierarchy should equal that note. With the exception of the dotted quarter note equal to three eighth notes, the remaining notes are all even multiples of the eighth note. The grouping structure would be based on the differences in duration of the notes, the pattern of frequency changes, and the clear repetitions of the groups. The grouping is perfectly aligned with a duple meter in which there are four strong-weak units.

"Down in the Valley" has an unusual time signature of 9/8, indicating that an eighth note gets one beat and there are nine beats per measure. The lowest level of the meter hierarchy would be based on the individual eighth notes, the next higher on three eighth notes, lining up with the dotted quarter notes, and the highest level would be based on nine eighth notes, lining up with the initial dotted quarter note of the first full measure, the dotted half note in the second and fourth measures, and the first dotted quarter note in the third measure. Another factor that influences the meter is the position of the important tonal elements. In extensive analyses of classical composers, the important tonal elements (i.e., the tonic and fifth) occur at the strongest metric points. This correspondence probably strengthens the musical expectancies that convey musical rhythms and emotions (Prince & Schmuckler, 2014) (Fig. 4.7).

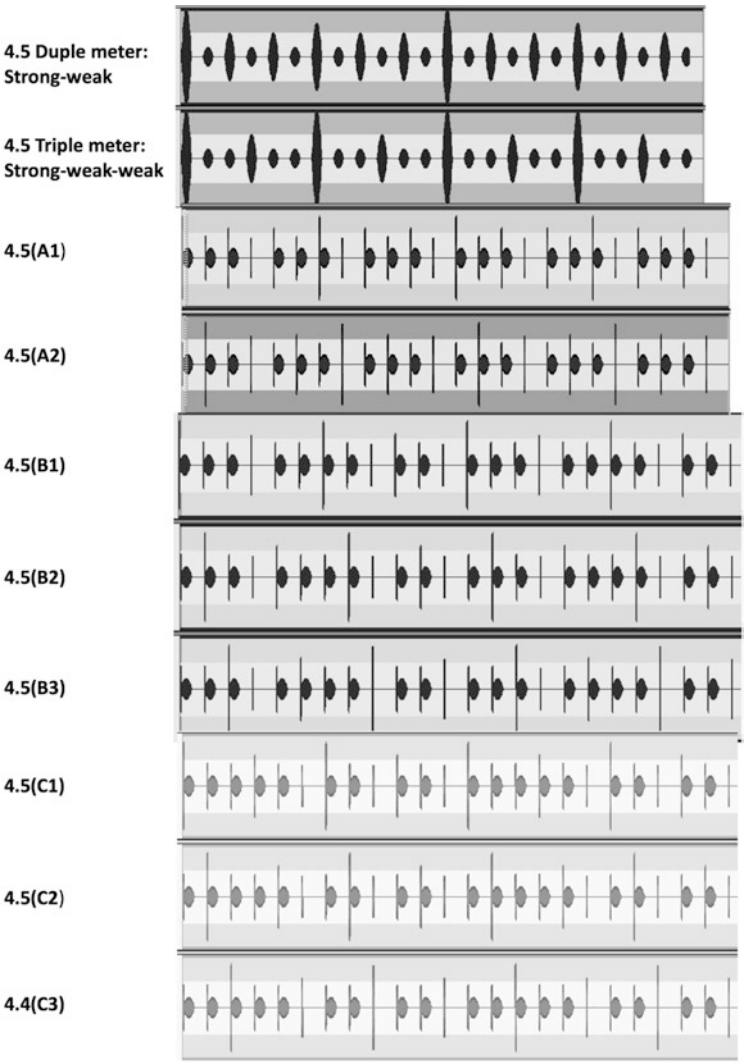


Fig. 4.5 Amplitude by time representation of the sound files for the rhythms displayed in Fig. 4.4. The presentation rate is 4/sec. Each tone is 125 msec and each beat (the thin vertical line) is 10 msec. The strength of the accents for the duple and triple meters is indicated by the amplitude of each sound (its height). The strength of each beat is indicated by the length of the beat line

Sound Files 4.5: Beats in strong-weak duple meters and strong-weak-weak triple meters for the sequences shown in Fig. 4.5

4.2.4.2 Multiple Rhythm Lines

In spite of the obvious strengths, there are clear limitations to this hierarchical analysis of rhythm. First, hierarchical analyses are restricted to one rhythmic line, so they cannot cope with thick polyphonic passages in which instruments are playing at different tempos and differing groupings. Second, these hierarchies

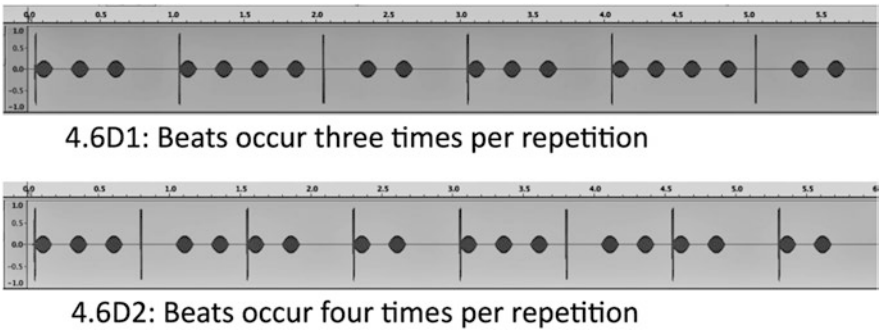


Fig. 4.6 The amplitude × time representation of the alternate beat structures. Sound Files 4.6D1 and 4.6D2 illustrate how incompatible meters can make the same pattern xxx-xxxx-xx- appear to be different rhythms as found by Povel and Essens (1985). It is also possible to hear the effect of shifting the meter by comparing 4.5B1, 4.5B2, and 4.5B3 to one another and 4.5C1, 4.5C2, and 4.5C3 to one another

Sound Files 4.6: Alternative beat structures shown in Fig. 4.6D1 and D2 determine the perception of the pattern

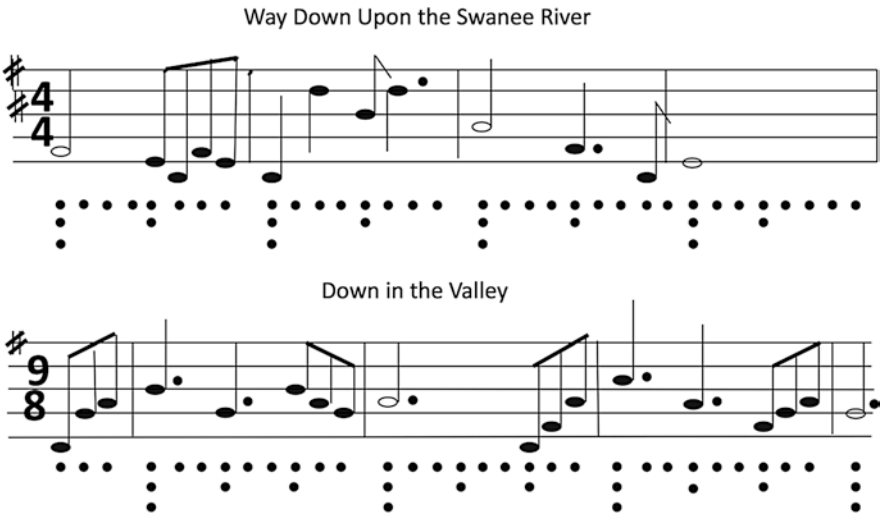


Fig. 4.7 The initial phrases of "Way down on the Swanee River" and "Down in the Valley." The number of dots under each beat indicates the importance of each beat. In these phrases, the important beats always coincide with the onset of a note, but that does not always occur in other songs. (Adapted from Fox & Weissman, 2007)

work for Western music with a strict periodic meter, but will not work for other musical genres with non-metrical rhythms in which the beat is not periodic.

4.2.4.2.1 Polyrhythms

Our argument above is that the meter, beat, and accent for a single rhythmic line arises from the interplay of rhythms at different levels. In essence, the rhythms at slower levels create the accentuation for the rhythm at faster levels.

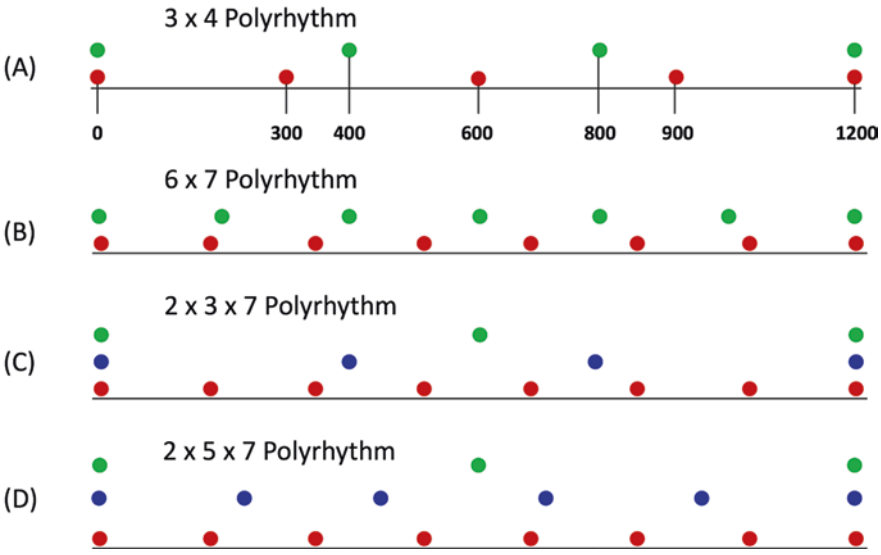


Fig. 4.8 In (A), the timing of three-pulse and four-pulse polyrhythms is shown for each of the pulse trains. In (B), (C), and (D), the two-pulse 6 × 7 polyrhythms, and the three-pulse 2 × 3 × 7 and 2 × 5 × 7 polyrhythms are shown. The pulse trains are synchronous at the first note of each cycle

An alternative approach is to create polyrhythms composed of two or three isochronous pulse trains in order to understand how the interplay of the levels represented by each pulse train determines the overall rhythm (Handel, 1984). Polyrhythms are composed of two or more isochronous pulse trains with identical elements that have conflicting tempos, for example, two per pattern repetition versus three per repetition (i.e., two tones per measure versus three tones per measure) but not two per repetition versus four per repetition.

An example of a 3 × 4 polyrhythm is shown in Fig. 4.8A. Based on a repetition rate of 1200 msec, the onset-to-onset interval for the three-pulse rhythm would be 400 msec, while the onset-to-onset interval for the four-pulse rhythm would be 300 msec. There would be one point at which the pulses from each rhythm would coincide; the order of the pulses alternates and the onset-to-onset intervals between the pulses from each rhythm would be either 200 msec or 100 msec. In polyrhythms with conflicting ratios, the onset timing between elements changes. For more complex polyrhythms like 6 × 7 (Fig. 4.8B), the number of different onset-to-onset intervals between the rhythmic pulses increases and this increase is even greater for three-pulse polyrhythms such as 2 × 3 × 7 (Fig. 4.8C) and 2 × 5 × 7 (Fig. 4.8D). Schematic illustrations of 2- and 3-pulse polyrhythms with different frequency components are shown in Fig. 4.9.

These sorts of polyrhythms can be heard in several ways: (a) one of the two or three pulse trains is heard as the meter; listeners would tap the elements of only one of the pulse trains and the elements of the other pulse train would fall off the meter; (b) the meter of the polyrhythm would occur only on the synchro-

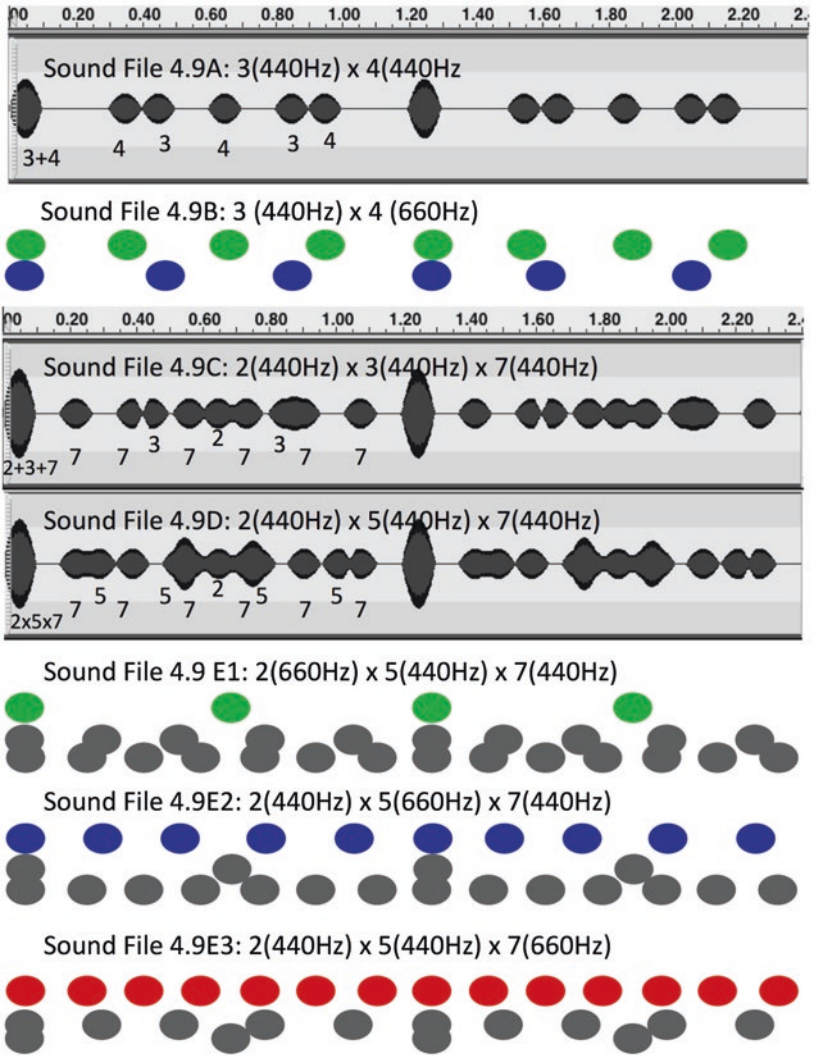


Fig. 4.9 The first panel shows the amplitudes across time of the identical notes in the 3 × 4 polyrhythm. The duration and amplitude of the notes is evident and the higher amplitude of the synchronous first note of each repetition is due to the sum of the two pulses. The next panel is a schematic of the two pulses at different frequencies. The third and fourth panels show the notes of the 2 × 3 × 7 and 2 × 5 × 7 polyrhythms with all notes at the same pitch. The final three panels illustrate schematically the timing of the pulses of the 2 × 5 × 7 polyrhythm when one of the pulses is presented at a different frequency

Sound Files 4.9: Two- and three-pulse polyrhythms pictured in Fig. 4.9

nous elements so that there would be one beat per measure; (c) the two pulse trains could be integrated into a *cross-rhythm* and the listener would tap each element, six for the 3 × 4 polyrhythm (not seven because of the synchronous element). If the polyrhythm were composed of three pulse trains, for example,

2 × 3 × 7, the possible ways of hearing the rhythm would multiply. In addition to the three ways listed above, there would be three ways in which listeners tap along with two of the three pulse trains with the third pulse train heard separately, 2 × 3 versus 7, 2 × 7 versus 3, and 3 × 7 versus 2. One might expect that as the complexity of the polyrhythm increased there would be a tendency to find a simpler meter based on one of the pulse trains or based on the repetition of the polyrhythm at the point where the three elements synchronize.

In experiments investigating the perception of polyrhythms, the task of the listeners was simply to tap along in any way they wished as if they were tapping on the steering wheel at a red light. Although there were differences among listeners, there are some consistent trends:

A) The fundamental factor determining the interpretation of the polyrhythms was the timing among the elements. Remember from the beginning of this chapter the preferred tapping rate occurred when the onset-to-onset interval was between 200 and 800 msec. For polyrhythms composed of two pulse trains, if the interval between the onsets of the elements of only one pulse train falls in the preferred tempo zone of 200–800 msec, there was an extremely strong tendency to tap along with that pulse train. If the intervals between the onsets of the elements for both pulse trains were within that 200–800 msec region, then the low pitch or more staccato pulse train was preferred. If the overall presentation rate of the polyrhythm was so slow that the onsets between the elements of both pulse trains are longer than 800 msec, most listeners shift to tapping every element of both pulse trains. Conversely, if the presentation rate is so fast that the onset intervals of both pulse trains become less than 200 msec, then most listeners switch to tapping once each repetition, on the synchronous elements.

B) The second factor is based on the configuration of the specific pulses within a polyrhythm. Although the outcomes for the 2 × 3 and 2 × 5 polyrhythms are similar, if those polyrhythms are combined with a pulse train in which the elements occur seven times per measure, for example, 2 × 3 × 7 and 2 × 5 × 7, the outcomes become quite different. For the 2 × 3 × 7 polyrhythm, listeners either tapped to the 2 × 3 cross-rhythm or the seven-pulse trains; they rarely tapped separately to the two- or three-pulse train. The seven-pulse train emphasized the similarity between the two- and three-pulse trains so that neither one of them was tapped alone. For the 2 × 5 × 7 polyrhythm, one group of listeners tapped the seven-pulse train at the slower presentation rates and the two-pulse train at the faster rates, and a second group tapped the two-pulse train at all rates. The seven-pulse train pulled the five-pulse train from the two-pulse train so that tapping to the two-pulse train was predominant and 2 × 5 cross-rhythms did not occur. If one of the three-pulse trains was a different frequency, the tendency was to tap to the cross-rhythm of the two-pulse trains with the same frequency at slower tempos, and tap to the different frequency pulse train at the faster tempos.

To summarize, the perception of the meter of a polyrhythm is contextual with respect to the timing of each pulse train, the configuration of the polyrhythm, and the frequency and intensity of the elements making up each pulse train. Each factor influences the rhythmic interpretation; its effect depends on the values of all the other factors. There is no single rhythm in a temporal pattern so that there can be no single stimulus factor that determines the perceived rhythm.

4.2.4.2.2 Non-metric Rhythms

Even though Western music is almost exclusively metric with equally spaced beats, other cultures make use of complex meters in which the bests do not occur at equal intervals (Fig. 4.10). For example, in Turkish music, common complex meters are five or seven beats broken into 2 + 3 or 2 + 2 + 3 beats (Dave Brubeck makes use of these meters in "Take Five" and "Blue Rondo a la Turk"). It is surprising that exclusive listening to the simple Western meters seems to interfere with the perception of more complex non-metric rhythms. For example, six-month-old Western infants can detect meter changes in both simple and complex Balkan meters, but 12-month-old Western infants cannot detect the same changes in the complex Balkan rhythms. At 12 months it is possible to reverse this loss by a short period of exposure to complex Balkan music, but the same period of exposure for adults does not reverse the loss (Hannon & Trehub, 2005). Recently, Kalender, Trehub, and Schellenberg (2013) found that the ability of adults to detect changes in complex Turkish meters did not require experience with specifically Turkish rhythms. As long as the adult had listened even sporadically to music that did

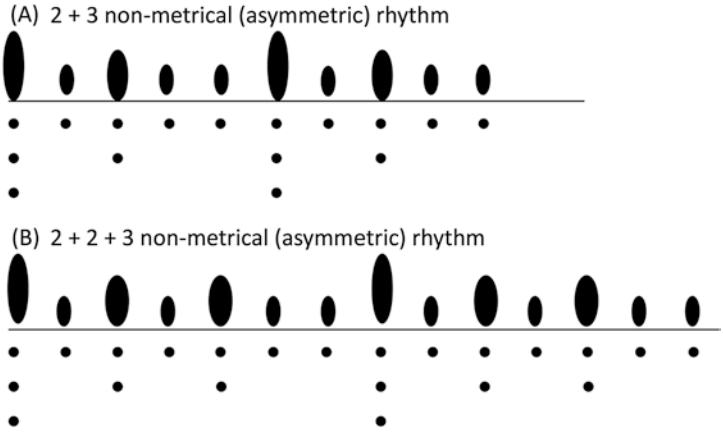


Fig. 4.10 Non-metric rhythms contain equally timed notes, but the beats do not fall at equal intervals. (Non-metric rhythms are sometimes termed asymmetric rhythms). (A) The rhythm 2 + 3 is shown in terms of the actual sounds (heard in Sound File 4.6A). The strong beats occur on the first and third tones. (B) The rhythm 2 + 2 + 3 is shown. The strong beats occur on the first, third, and fifth tones

Sound Files 4.10: The 2+3 and 2+2+3 non-metric rhythms drawn in Fig. 4.10

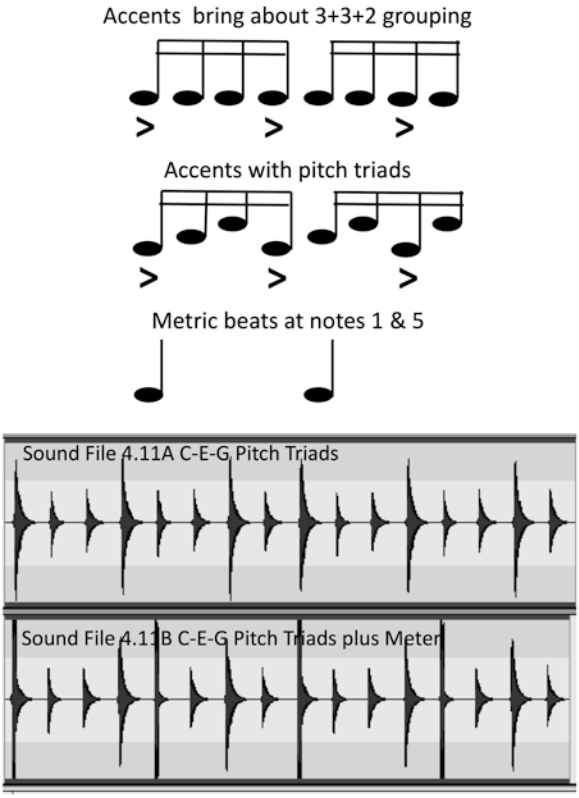


Fig. 4.11 Accents characteristic of bluegrass and ragtime music. If metric beats are added, the rhythm seems to shift over time

Sound Files 4.11: Ragtime rhythms based on the major triad and duple meters shown in Fig. 4.11

make use of non-metric rhythms, for example, Indian music, they were able to detect the changes. It seems that sensitivity to non-metrical rhythms is hard to lose, but then again it is hard to regain.

A second way to create a non-metric rhythm is to accent off the beat. Start with a simple eight-note measure in which the beats would normally fall on the first and fifth note. If the first, fourth, and seventh notes are accented, then a 3 + 3 + 2 grouping is created within a standard eight-note duple meter as illustrated in Fig. 4.11. If the grouping is played as pitch triads it creates a rhythmic phrase that is typical of ragtime and bluegrass music. Furthermore, if this rhythmic phrase is combined with a regular beat on notes one and five, it results in a sense of a shifting rhythmic structure (Fig. 4.11).

4.2.5 Beats, Embodied Rhythms, and Relative Movements

At the beginning of this chapter, two seemingly contrasting conceptualizations of rhythm were presented. The first made use of classic Gestalt principles of grouping and the second, while far more vague, made body movements and

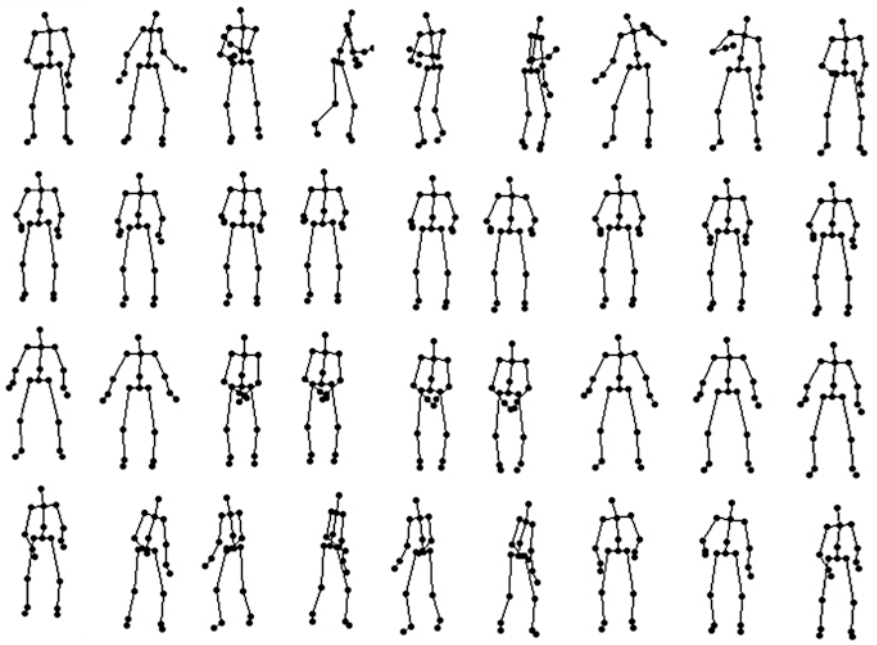
rhythm synonymous. The motor regions create the timing intervals used to pace the body movements. At this point, it is possible to synthesize two points of view and tie both to notions of relative motion perception due to Gunnar Johansson (Chap. 2). What is common to all is that rhythm occurs hierarchically at several levels. The Gestalt principles leading to beats and meters start at the fastest, highest level with each element receiving a beat. At the lower, slower levels, each second or third element receives a beat and the summation of those beats yields a two- or three-beat meter. In thinking about embodied rhythms, we can imagine that the overall dancing movements are built out of the slower motions of the body and torso, the faster movements of the arms and legs, and the still faster motions of the hands. Leman and Naveda (2010) have identified spatiotemporal reference frames, which they term "basic gestures," that match periods of the meter. These repetitive motions link musical properties to body movements. Thus, each level of the motions or gestures can be thought of as being analogous to each level of the beats. Johansson's work demonstrates that our perception of each level of the body motions is relative to the slower movements of other levels. If we did not perceive movements this way, then the rhythmic movements of the hands would seem random because they would be entangled in the movements of all the other body parts. (This is reminiscent of the general problem of auditory perception; namely partitioning the sound wave into coherent sources).

In Johansson's analyses the slower movements of the more massive body parts were subtracted from the faster motions of the lighter body parts in order to see the trajectories of those faster movements. Toiviainen, Luck, and Thompson (2010) mimicked Johansson's lighted dot figures, but instead of people walking or running, they had dancers move rhythmically to music at different tempos. Furthermore, instead of partitioning the motions of one body part from another, they partitioned the overall motion into different "timings" or periodic movements. For example, the slowest motion of a dancer could be a simple back-and-forth swaying. In addition, the dancer could also rotate the torso or bounce up and down at the knees, could also swing the arms at a faster rate, and could also wiggle fingers at even a faster rate. The dance consists of all these movements occurring simultaneously. The motion of the fingers would be the combination of the movement due to swaying, the movement due to the bouncing, the movement due to the swinging arms, and the movement due to the wiggling. The analyses used by Toiviainen et al. (2010) isolated the movements that occurred at the slowest rate, those that occurred at twice that rate, and those that occurred at four times that rate. Adding those movements together makes the original motion reappear. (This is the same way that the quality of a non-changing sound is analyzed, for example, a square wave (see Chap. 5). The amount of energy at the fundamental frequency and at multiples of the fundamental frequency is measured. Playing those frequencies at the derived amounts of energy reproduce the original sound).

A simulation of the various movement levels is shown in the following video. Each column represents a different dancer. The lowest row is the slowest motion and each higher row represents movement that is twice as fast. The top

row is the sum of all the movements (I am partial to the dancer on the far right). It is easy to see that the body parts with the highest mass and inertia move at the slowest rates and the lightest parts move at the highest speeds. The movements are "locked together" in timing to simple ratios of two or possibly three with different music, and I would argue that we perceive those movements relative to a common time base and relative to a common set of motions.

I thank Dr. Toiviainen for kindly providing the movie.



In this single frame, the dots on the dancer's limbs show the location of the lights used to record the movements of the dancers.

Note: The movie is found in the supplementary material

4.2.6 Do Animals Have Rhythm?

Obviously animals make rhythmic movements for locomotion, capturing prey, and so on. Furthermore, birds increase the rhythmic consistency of their song if threatened by interlopers. The research question has been whether animals can learn to synchronize to an external beat. Learning to move in time with the beat is not simple. Children can respond to different tempos far before they can synchronize. Children try to move in synchrony, but they need help to learn to do so. Four- and five-year-olds synchronize poorly, if at all. Synchronization success shows up only after ages seven or eight (Repp, 2013). Gaining the ability to abstract the beat and synchronize a body motion to it can be a long process. The simple rhythms found in infant songs and parental instructions start the process,

and the ability to attend to the levels of rhythm increase with all types of experience (for a broader view of rhythm in animals see Fitch (2018)).

It is still an open question whether animals can synchronize body motions to the beat. It has proven difficult to provide an answer for several reasons.

1. In constructing stimuli to test for synchronization in other species, humans make use of their own beat perceptions. Obviously, we cannot be certain that our perception matches that of other species (remember when discussing camouflage in Chap. 2, we imposed our sensitivities in evaluating the nature and effectiveness of the camouflage, not those of the predators). Probably all species can respond to simple auditory pulses, but most likely lack the cognitive capacity to abstract the beat in more complex sound environments. Moreover, as mentioned above, even human children need extensive practice to match the beat, and without similar training very few species are likely to achieve matching even if they are ultimately capable of doing so.
2. To demonstrate beat matching, the animal must have sufficient control of a limb or body part to effectively synchronize at the correct tempo. Humans can synchronize arm movements at faster tempos than body movements so that it is critical to choose a response that potentially could match the tempo of the beat. It is possible to train a seal to move its head in time with a sound, but not its flipper (Wilson & Cook, 2016).
3. Beat matching in the artificial experimental situations depends on the animal's motivations and attention span. Most of the demonstrations of beat matching involve species of parrots, highly social animals that bond and vocally mimic their owners. The popularity of the videos of Snowball (Patel, Iversen, Bregman, & Schulz, 2009) may have led researchers to limit their search for other species that may beat match, particularly since matching may not occur unless there is extensive training and practice. However, other social animals like dogs do not show beat matching even after extensive training.
4. Beat matching may be far more extensive in the animal's natural environment. Animals may synchronize their calls in social groups so that there are alternating and simultaneous utterances, that is, turn taking. Even bats produce duet-like social calls with one bat responding within 1/3 sec of the end of another's call (Vernes, 2017).

YouTube
Snowball™-Our Dancing Cockatoo (watch?v=N7IZmRnAo6s)
Snowball™-Another one Bites the Dust (watch?v=cJOZp2ZftCw)

Wilson and Cook (2016) make a strong argument that beat matching and other ways of synchronizing body movements to external stimuli is more widespread among animals than previously imagined. They argue that rhythmic

behavior depends on the voluntary control of motor behaviors and the learned coupling of those behaviors to sensory stimuli. If the four constraints listed above are overcome, the entrainment of motor behavior and rhythmic stimuli could be found in a wide range of animals.

Furthermore, there is strong evidence that animals are sensitive to rhythmic tempo. For example, elephant seals can distinguish among individuals based on the tempo and the spectrum of the pulses of their stereotyped mating calls. Each seal has a preferred tempo acquired when young and maintained through adulthood. Across the colony, the mating calls consist mainly of individual pulses presented at rates from roughly 1–3/sec. It is important to note that the preferred tempo for the dominant male can be any value within that range; it is not the fastest or slowest rate. Marthevon, Casey, Reichmuth, and Charriuer (2017) recorded the call of the dominant males in the colony, and during playback the peripheral males made avoidance movements. But, when the authors either increased or decreased the tempo without changing the spectrum, the peripheral males greatly reduced their avoidance and even disregarded the call. It would be interesting to determine how quickly the peripheral males learn the new relevant tempo when a new dominant male emerges.

4.3 Timing

Nearly everything in this book concerns grouping. For the purely static, spatial images in the first three chapters, generalized Gestalt laws attempt to predict which elements combine to form objects or surfaces. For the purely temporal rhythms so far in this chapter, similar Gestalt laws attempt to predict which sound units become the strong and weak beats, thereby creating the meter that brings about the overarching grouping of the sounds to form sources. Here we want to consider two related issues. First, we know that stream segregation and the preferred meter for polyrhythms changes as the tempo is varied. But up to this point we have tacitly assumed that a single rhythm would be perceived identically at all tempos. Here we will consider whether this is true.

Second, space and time are interlocked in several ways. The timing of sensations can determine the nature of the resulting spatial representation. As described below, tactile pulses presented at differing onset-stimulus intervals can lead to the perception of a "hopping" motion on the skin. This hopping or leaping perception has been termed sensory saltation and can also be found for visual and auditory presentation. By reversing our perspective, it is possible to use temporal order judgments to infer how different spatial representations are coordinated.

In addition, spatial configurations change over time. The perceptual problem is to determine the optimal matching of the elements in different rhythms or in each successive visual image, that is, the *correspondence* or constancy problem. Historically, the term perceptual constancy was used to describe the abstraction of visual and auditory objects in spite of diverse orientations, movements, sizes, octaves, timbres, and so on. Here we have used the term correspondence in two ways. First, correspondence refers to the grouping of parts of an interrupted object or source. Second, correspondence refers to the identity of an object or

source in a different context, making it equivalent to constancy. For example, the moving lights on joints are perceived as dancing figures in the demonstrations pioneered by Johansson (Johansson, 1973). The matching problem is simple here because the light points move in predictable arcs with small changes in each image. Tracking an individual firefly in a field of fireflies is far more difficult because the interval between flashes and the motion direction is more erratic. Even when the parts of a simple stimulus move coherently as in the Ternus configuration, the timing between images can change the perceived coherence.

4.3.1 Tempo and Rhythmic Organization

What we need to do is distinguish between two aspects of rhythmic organization. When we initially view a scene or listen to a string of sounds, our initial impressions are global and diffuse. The initial division of the visual array into closed contours is probably based on the principle of uniform connectedness: identical colors, textures, and motions are joined together. Further looking yields finer detail, figure-ground organization and three-dimensional objects. In similar fashion, the initial division of the auditory array into connected sequences would be based on pitch, timbre, and temporal proximity due to stream segregation. Further listening to one stream would differentiate the intervals between successive tones and possibly lead to the emergence of a stable metric. What is therefore common to both looking and listening is the initial splitting of the field into global parts sorted by common Gestalt principles. Following this split, each piece is further analyzed, eliminating alternative possibilities resulting in the final percept.

At the beginning of the chapter, we discussed that rhythms seem most natural between presentation rates of 7.5 elements/sec and 0.5 elements/sec. Beyond these rates, the sense of grouping and timing disappears. Two rhythms are drawn in Table 4.1. In both, there are five tones in a 16-element repeating pattern. The number of stars under each position indicates the metric strength of each position. In the first, all five tones occur at the stronger metric positions for a four-beat strong-weak- strong-weak meter, namely 1, 5, 9, 11, and 13. In the second, the tones occur at weaker metrical positions, namely 1, 4, 8, 10, and 12.

At the fastest tempo of 7.8 elements/sec, there is a sense of three groups, two of single tones and one of multiple tones, but only a very weak sense of repetitive timing. At the two slower tempos of 3.9 and 1.6 elements/sec, the 1-1-3- groups become predominant and there is a strong sense of rhythmic repetition, namely, two single tones followed by a group of three tones. At the two slowest speeds, 1.25 and 0.5 elements/sec, the grouping is far weaker and the elements seem unconnected. To me, there seems to be little difference in the grouping of the metric and non-metric versions.

Here we will consider if the perception of non-metric patterns is the same at different presentation rates. To do this, we present two rhythms at different tempos and ask listeners to determine if the rhythms are the same or different. The experimental strategy is simple: first, present the two rhythms at the same tempo to make sure that listeners correctly judge that the two rhythms are

Table 4.1 Two five-note rhythms constructed in a 16-element sequence. The number of stars indicates the metric strength of each position. The code 1-1-3 indicates that there are two groups of a single element followed by a group of three elements. The spacing between the second single element and the group of three elements determines whether the rhythm is metric or non-metric

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Metric	X				X				X		X		X			
Non-metric	X			X				X		X		X				
	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
	*		*		*		*		*		*		*		*	
	*				*				*				*			
	*								*							

Sound Files 4.12: The metric rhythm x---x---x-x-x--- and non-metric rhythm x--x---x-x-x---- played at different tempos

Table 4.2 Two pairs of five-note rhythms used to test whether rhythms are perceived in the same way at different tempos

Rhythm	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
1-1-1-2-	x			x			x			x		x				
1-2-2-	x			x		x				x		x				
1-1-2-1-	x			x			x		x			x				
1-1-3-	x			x				x		x		x				

Sound Files 4.13: Two pairs of non-metric rhythms shown in Table 4.2 played at different tempos

different, then present the rhythms at different tempos. If listeners can still correctly judge that the rhythms are different that shows that rhythms maintain their organization across tempos, but if listeners cannot judge whether or not the rhythms are different that would suggest that rhythms are perceived in different ways as tempo varies.

Two examples are given in Table 4.2. In the first, the rhythm 1-1-1-2- is compared to the rhythm 1-2-2-. As can be seen in Table 4.2, the only difference between the two occurs at elements six and seven so that 1-1-1-2- is heard as three individual sounds followed by a pair of sounds, while the rhythm 1-2-2- is heard as a single sound followed by two pairs of sounds. If the two rhythms are played at the same tempo, the percentage of different judgments was 67, but if the two were played at different tempos (3.9 tones/sec and 1.8 tones/sec), the percentage of different judgments was 33, less than chance performance. Two-thirds of the listeners thought they were identical. In the second, if the rhythms 1-1-2-1- and 1-1-3- are played at the same tempo it was relatively easy to determine that the rhythms were different, 80 percent. But if the rhythms were played at the different tempos, only 44 percent of the judgments were that the rhythms differed, again less than chance (Handel, 1993).

We can understand these results in terms of the Gestalt grouping principles. Suppose the first rhythm is presented at the faster rate so that the interval between each step of the rhythm is 80 msec. The onset-to-onset interval among

the single tones is 240 msec and the interval between tones heard in a group is 160 msec. At the slower rate the interval between each step is 175 msec so that the onset-to-onset interval between single tones is 525 msec and the interval between tones heard in a group is 350 msec. The interval among tones in a group at the slower tempo is longer than the interval between single tones at the faster tempo. To correctly judge whether the two rhythms are the same, listeners must make a figure-ground reversal so that intervals that formerly signified a group now signify a single tone. These outcomes suggest that this reversal is not easy.

In general, rhythmic organization at slower tempos, 1–3 tones/sec, is more flexible and rhythms are often organized beginning with the initial sound. At faster tempos, however, regardless of the initial sounds, the rhythm is reorganized so that the longest silent interval ends the rhythm. For example, if we started the rhythm 1-1-3- at the eighth element making it 3-1-1-3-1-1-. At a slower tempo listeners could retain the initial 3-1-1-organization, hearing the group of three tones starting the rhythm. But at faster tempos listeners shift to a 1-1-3 rhythmic organization, starting with the two single tones so that the longest silent interval ended the rhythm. I would speculate that at the slower tempos rhythms are not tightly embodied into body movements so that there is greater flexibility in how the rhythm is organized. At the faster tempos, there is a tight connection between the rhythm and body movements so that the organization of the rhythm reflects the neural and muscular constraints imposed by the limbs. These constraints would be more prominent when drummers are tapping syncopated rhythms at fast tempos (Barton, Getz, & Kubovy, 2017).

Sound Files 4.14: Organization of rhythms at different tempos

4.3.2 Sensory Saltation

If a stimulus is presented at one location, and a second one is presented close in time and space, the first stimulus is perceived to shift its location toward the location of the second stimulus. The timing between the two stimuli changes the perceptual space of the body; our spatial representation is a function of the temporal patterning of the sensations. This illusion was first discovered for tactual presentation (Geldard & Sherrick, 1972). What made this illusion so interesting is that if multiple taps were presented to the first location on an arm and followed by a single tap at a second location further down the arm, the taps seemed to hop down the arm. The perception was not a continuous motion as found for apparent motion discussed in Chap. 3, but a series of leaps or hops. This led to calling the illusion the cutaneous rabbit.

To understand the "hopping," it is best to start by considering the "reduced rabbit" paradigm. There are only two stimuli: the first stimulus occurs at one point on the skin and following a variable delay the second stimulus occurs at a different position on the arm. The tactual stimulation at both locations is identical, and in different studies the duration of the stimuli ranges from 5 to 50 msec. The precise characteristics of the stimuli are not critical and the same outcomes occur for physical taps, electric pulses, and even "hot" spots gener-

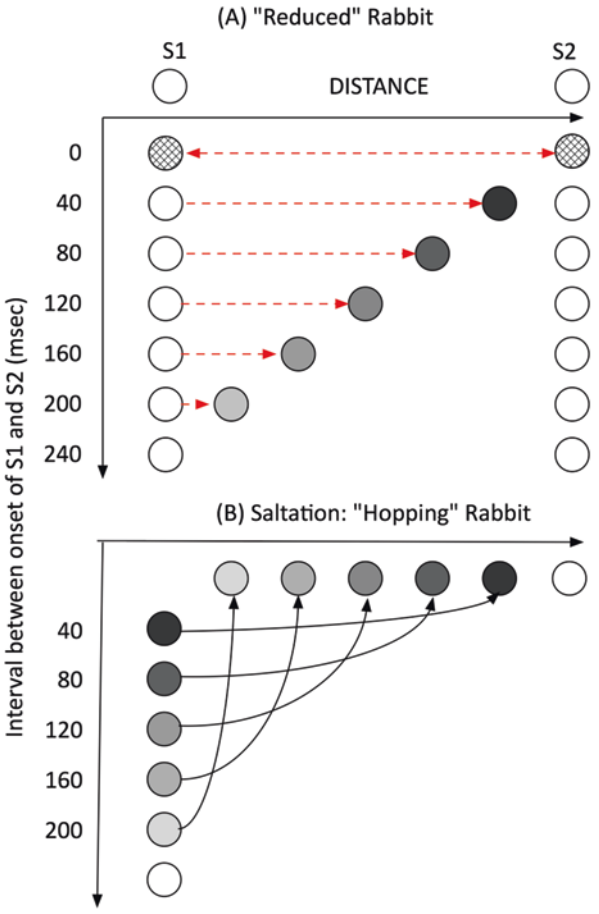


Fig. 4.15 Representation of the "reduced" rabbit and "hopping" rabbit configurations. The reduced rabbit stimulus consists of two pulses at different points on the arm, hand, or torso. (A) If the pulses at S1 and S2 are presented simultaneously, only one pulse is perceived. At the shortest intervals, the pulse at S1 is perceived to be adjacent to S2. As the interval lengthens, the shift in the position of S1 decreases, and at intervals roughly greater than 200 msec no shift occurs. (B) If multiple pulses occur at equal intervals, the shift in position of S1 relative to S2 due to the difference in timing spaces out the pulses, leading to the perception of pulses hopping down the limb or torso

ated by a CO_2 laser (Trojan et al., 2006). (Usually there is a prior stimulus signaling that the trial will begin in about 1 sec, but it does not affect the illusion).

The outcomes are shown in Fig. 4.15A. The two stimuli, S1 and S2, are identical and located a short distance apart. If the interval is greater than 240 msec, the two stimuli are perceived separately, at their actual positions. As the interval decreases, the illusionary displacement of S1 toward S2 increases. Participants report that the strongest perception of movement occurs when onset interval is between 60 and 90 msec. If the stimuli are presented synchronously or at a very short interval (< 25 msec interval), then only one stimulus is perceived at one location. This is illustrated in Panel A, Fig. 4.15.

Now suppose that five pulses are presented at position S1. The first pulse occurs 200 msec before the pulse at S2. The remaining four pulses at S1 occur at 40-msec intervals: 160 msec, 120 msec, 80 msec, and 40 msec before the pulse at S2. Based on the outcomes from the "reduced" rabbit, the initial pulse

that was 200 msec before the pulse at S2 will shift just a short distance toward S2. Each of the following pulses at S1 will shift further toward S2 because each is closer in time to S2. That will lead to the perception of discrete "hops" ending at the position of S2. This transformation of timing differences into spatial distances is shown in Fig. 4.15, Panel B. This is a strong illusion. Cholewiak and Collins (2000) made use of multiple vibrators placed along an arm to create a veridical hopping movement. People were unable to discriminate between the true movement and the illusionary hopping rabbit.

Saltation can be produced in the visual and auditory modalities, and Trojan, Getzmann, Moller, Kleinbohl, and Hölzl (2009) found it possible to create saltation between tactual and auditory stimuli. Nonetheless, there are spatial and temporal limits; the spatial distance beyond which saltation does not occur differs across the body. The saltation area is circular on the hand but elongated on the arm. Most importantly, saltation does not cross the body midline. This differs from tactual apparent motion that easily does cross the midline.

Finally, saltation also reflects the notions of belongingness found for all other perceptual acts. The strength of the hopping movement is reduced if the quality of the pulses or the rhythm of the pulses at S1 is varied. The coherence of the pulses at S1 does matter.

4.3.3 Temporal Order Judgments

Sensory saltation demonstrates the intricate connection between timing and space. The interval between two sensations affects the felt position of those stimuli. A second timing task, the judgment of temporal order, illustrates the somewhat flexible connection between body postures, limb positions, and external space.

In the typical temporal-order judgment task, two stimuli are presented separated by a short period of time, and participants simply judge which was presented first. Across many experiments involving auditory, tactual, and visual stimuli, the minimum asynchrony to correctly judge order is about 20 msec (Hirsh & Sherrick, 1961). The actual values differ due to the specific experimental conditions, but on the average the tactual sense requires the smallest interval for correct judgments. However, just crossing the hands dramatically alters perception of tactile temporal order. Judgments were often inverted at longer stimulus-onset asynchronies (100–200 msec.) so that the incorrect hand was judged to be leading, and correct judgments often required much longer stimulus-onset asynchronies (400 msec–1 sec). This was a finding unique to the tactile modality, because we can cross the feet, arms, and hands (tactile sense organs) in space, but not the eyes or ears (Yamamoto & Kitazawa, 2015). The same effect occurred if one hand and one foot were crossed; in fact, nearly every combination of crossed limbs or fingers worked. Moreover, as David Katz originally commented, tactile signals at the hand due to an extended probe are attributed to the movements at the tip of the probe. To investigate whether physically crossing the arms is necessary or whether stick crossing will suffice, Yamamoto, Moizumi, and Kitazawa

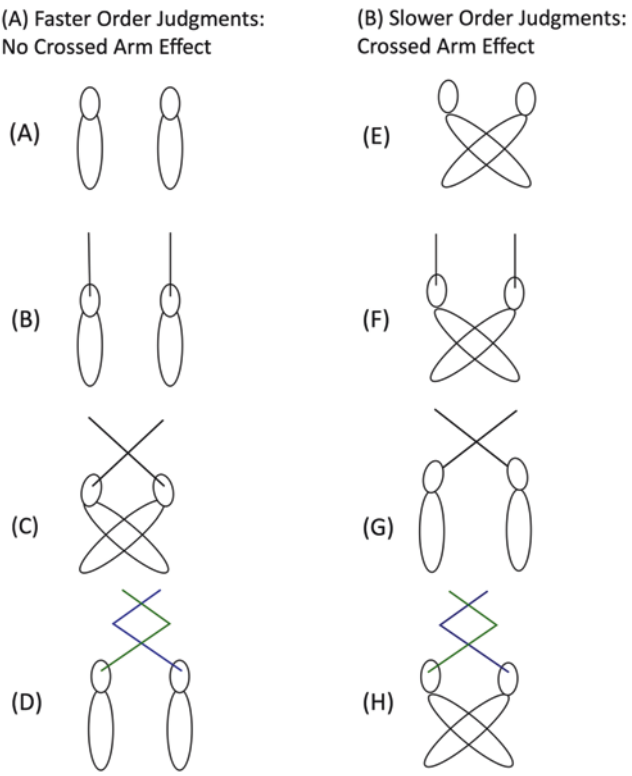


Fig. 4.16 Temporal-order judgments are faster and inversions do not occur if the arm or the arm + stick combination ends up on the ipsilateral side of the body. On this basis, in (C) the crossed sticks compensate for the crossed arms, and in (D) the double angles make the sticks end up on the ipsilateral side. But in (H), the double angles do not compensate for the crossed arms so that the stick end is on the contralateral side

(2005) created several variants of stick shape, arm crossing, and stick crossing shown in Fig. 4.16. In these experiments, participants grasped sticks placed on the vibrators arranged so that the vibrators did not affect the hands directly. The outcomes show that arm crossing (E or F) or stick crossing will bring about the slower-order judgments (G). It is interesting that the important factor determining the order effect is whether the end of the stick ends on the ipsilateral side of the body (C and D) or the contralateral side (G and H).

There are two important outcomes here. First, these results support Katz's contention that the tip of the probe acts like a sensory extension of the limb. In the time order judgment experiments, the probe tip is stationary and the position of the tip relative to the body is critical. But, recent research has emphasized also that movement of a probe allows one to explore objects in the same way as direct contact. If individuals use a wooden probe to touch an object, the perceived location of the touching point was intrinsic to the coordination system of

the arm and probe (Miller et al., 2018). The handheld tools act as a sensory extension (e.g., embodiment) of the user's body. In complementary research but still based on the premise of the purposive nature of touch (Gibson, 1966), Carello and Turvey investigated the mechanical properties of objects that provide the information (invariants in Gibson's terms) necessary to grasp and manipulate those objects (Carello & Turvey, 2017).

Second, the slower and inaccurate judgments for the crossed arm (and stick) experiment suggest one explanation based on the need to coordinate different spatial reference systems. There are at least two spatial representations for touch localizations. The first representation for the localization of the limbs is based on the body, while the second representation for the localization of the limbs is based on the external spatial coordinates. Typically the two are in coordination; the right hand is to the right of the left hand and the body midline, and the reverse is true for the left hand. But, the right hand can cross over the left one or move to the left of the midline and here the two representations must be brought back into registration. One representation has to be remapped into the other. There are several explanations for how this remapping occurs, but all explanations argue for the transition from internal body coordinates to external spatial coordinates for touch. The initial representation with respect to the body is rapidly remapped and transformed within 100–190 msec into the external one. Thus, tactile temporal order depends critically on the process of localizing tactile stimuli in external space, which develops over time.

There are several outcomes that support the idea that the "crossing" effect requires experience to attach the sensations of the skin (or of an external probe) to an external position. Individuals blind from birth do not show a crossover decrement, and for normal-sighted individuals the crossover decrement is much smaller when the hands are crossed behind the back where presumably visual experience is weaker.

We can understand the conversion to the external reference system as part of the need for the multisensory representation of space to be unified (Heed & Azanon, 2014). It would be impossible to act if the felt position of an object did not match the perceived visual position of that same object. The solution to the correspondence problem depends on the unity of the external reference. A different example of the "crossed arm" effect occurs an episode in the BBC/Masterpiece Theater series "Wolf Hall" written by Hilary Mantel. After young Thomas Cromwell picked up a heated tong in one hand, his master told him to put his hands in a watering trough and cross his hands. "Crossing would confuse the pain". Still another example is the crossed hand illusion. Cross one's arms, interlock the fingers by rotating the hands inward, and then rotate both hands upright. At this point the fingers of the left hand are pointing to the right and the fingers of the right hand are pointing to the left. People will then often confuse which fingers belong to each hand. The YouTube video makes this process easy to understand.

YouTube Video
The crossed hand illusion.

4.3.4 Visual Ternus Configuration

In the conventional Ternus display, three identical dots arrayed horizontally are displayed for a set amount of time, usually about 200 msec. The original array disappears, and after a blank interval the same three dots reappear but now offset to the left or right. After the identical blank interval, the original set of dots returns and the alternation continues. The sequence is illustrated in Fig. 4.17.

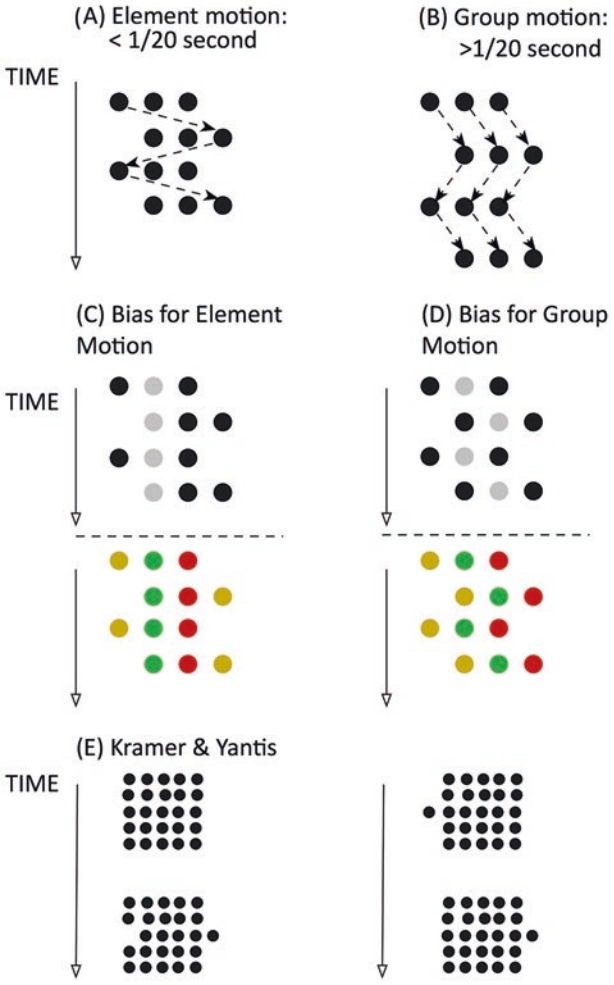


Fig. 4.17 (A) For the Ternus configuration, the perception that one element moves back and forth occurs if the interval between presentations is less than 1/20 sec (50 msec). (B) The perception that all three elements move back and forth as a unit occurs if the presentation interval is greater than 1/20 sec. (C) and (D) It is possible to bias the judgment toward either element or group movement by changing the stimulus configurations, as shown for black and gray dots or red and green dots. (E) The perceived "rigidity" of the dot configuration determines the type of motion. (Adapted from Hein & Moore, 2012; Kramer & Yantis, 1997)

Here, there are two percepts depending on the duration of the blank intervals. If the blank interval is less than 1/20 sec (50 msec) then the outside dot is seen to move to the other side of the two dots that do not appear to move. This has been termed *element motion* (Fig. 4.17A). If the blank interval is longer in duration, then the entire group of three dots is seen to shift to the left or right. This has been termed *group motion* (Fig. 4.17B). As the alternation continues, the dots undergo apparent motion and across a wide range of intervals, the percepts alternate between element and group motion, it is multistable.

Element motion and group motion represent two solutions to the correspondence problem. Several studies have attempted to determine how element similarity determines the blank interval at which the transition between element and group motion occurs. For example, Hein and Moore (2012) varied the order and position among the elements in the alternating stimuli in several ways. In some variations, the initial element of the first stimulus became the final element of the second stimulus so that element motion was fostered and continued at longer intervals (i.e., greater than 1/20 sec.). In the other variations, the second stimulus was identical to the first except offset in space so that group motion was fostered and it emerged at shorter intervals possibly eliminating element motion entirely. Examples based on black/white and color are shown in Fig. 4.17C & D. Apparent motion and the Ternus motion differ with respect to element similarity. Similarity has little effect on apparent motion but does affect the motion in the Ternus display.

Kramer and Yantis (1997) investigated the role of the spatial arrangements in determining the choice of element motion or group motion. Consider the two configurations shown in Fig. 4.17E. In the first, all the horizontal rows line up, suggesting that they form one unit. Only one dot moves and, given the perceived rigidity of the other dots, this sense of unity leads to the perception of element motion. In the second, one of the rows is originally offset to the left and then is shifted to the right in the second stimulus. This weakens the unity of the rows so that group motion increases.

Petersik and Rice (2006) summarize the evolution of explanations in terms of two opposing forces. If the inner elements of the three-element Ternus configurations are identical or seen to be similar, element motion occurs. The identity of the middle element fixes the location of the second configuration. Conversely, if the elements in each of the configurations are seen to be connected, group motion occurs.

Given that there are just two perceptions and that they are mutually exclusive, the Ternus paradigm has been used to investigate other aspects of temporal perception. Harrar and Harris (2007) constructed simplified Ternus configurations using just two stimuli. Both stimuli were either two lights or two tactual activators (i.e., unimodal) or one light and one tactual activator (i.e., bimodal). The same relationship between the stimulus-onset asynchrony and the type of apparent motion was found for all three conditions although the crossover

point between element motion and group motion was roughly 20% longer for the tactile presentation. The crossover point for the visual-tactual presentation equaled that for the visual presentation. Furthermore, as discussed in Chap. 3, the repeated presentation of one stimulus seems to fatigue that stimulus, leading to the increased perception of an alternate stimulus. Following this logic, to fatigue group motion Harrar and Harris (2007) presented the visual Ternus stimuli repeatedly at a long-onset asynchrony for 2.5 minutes. The important finding was that the fatigue procedure shifted the perception for the visual and visual-tactual stimuli toward element motion as expected, but it did not affect the perception of the tactual stimuli. These outcomes lead the authors to argue that there are no general timing mechanisms across modalities. This was the same conclusion from the study of visual and auditory apparent motion discussed in Chap. 3.

Chen (2009) made use of temporal ventriloquism to affect the perceived stimulus-onset asynchrony and thereby affect the shift between element and group motion. As discussed in Chap. 2, two tones that bound two lights (AVVA) seem to increase the asynchrony between the lights, while the same tones interspersed between the lights (VAAV) seem to decrease the asynchrony between the lights. By placing one tone before the first visual stimulus and the second tone after the second stimulus, the stimulus-onset asynchrony interval seemed longer so that the crossover occurred at a shorter interval. Conversely, if the tones were placed between the visual stimuli, the interval seemed shorter so that the crossover occurred at a longer interval. The effect did not occur if there was only one tone, showing once again that multisensory effects happen only if there is a belief that the sensations from each modality come from the same event.

These outcomes reveal how physical properties, spatial arrangements, and timing interact to solve the correspondence problem. To some extent, these aspects are functionally interlocked; it is possible to overcome a change in one property by a compensating change in another.

4.4 Visual Spatial Rhythms

4.4.1 The Visual Grid

It seems quite natural to feel musical rhythms and to experience the visual rhythms of dancers and athletes as they extend in time. However, it seems harder to be aware of the rhythms inherent in static visual designs and friezes. Starting with a small set of distinct patterns, such as shown in Fig. 4.18, complex designs can be made by combining and overlaying these patterns with each other in intricate ways. The patterns might be repeated, alternated, interleaved, rotated, reproduced at varying spatial scales, and/or offset laterally. Any one or combination of these processes can construct the layering characteristic of visual rhythms.

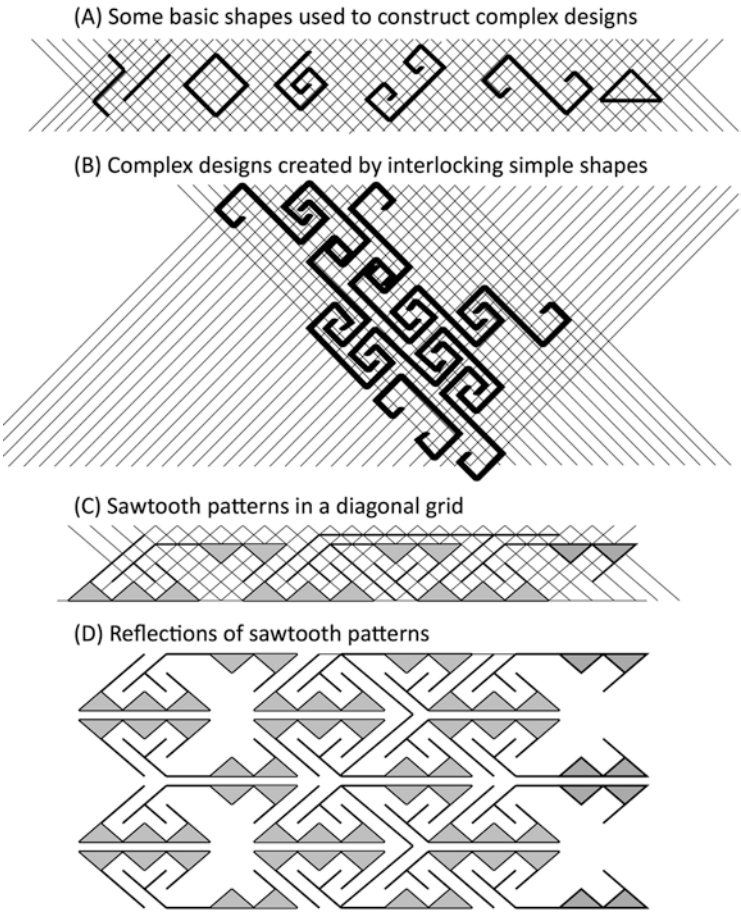


Fig. 4.18 Complex designs can be created on a diagonal grid by combining simple shapes in different groupings. A simple pattern in (C) can be reflected and repeated to create the complex design in (D). (Adapted from Tetlow, 2013)

When we analyzed the organization of auditory rhythms, the highest level could be termed the "metrical grid" and consisted of "dots" that represented each time point. The dots were equally separated in time, and the strong and weak beats represented by "dots" at lower hierarchical levels. By analogy, we can start with a "metrical" spatial grid with equally spaced lines and create a visual rhythmic pattern by darkening some of the metrical lines.

The first visual patterns were probably constructed using only a compass and a straight edge on an underlying grid (Tetlow, 2013). A variant of such a grid uses diagonal lines to fill the space with diamonds or squares at 45°. Now any shape can be described in terms of the grid units travelled before making a 90°

turn as shown in Fig. 4.18A. For example, +4, +4, +4, +4, where + indicates a right turn and – indicates a left turn, yields a small square. The sequence +1, +1+2, +2, +3, +3, +4, +4, and so on yields a square spiral and +1, +2, +3, +4, +10, +4, +3, +2, +1 yields a "C-shape." Mixing left and right turns such as +1, +2, +3, +7, −3, −2, −1 can produce an "S-shape."

What is important about these sorts of patterns is that the size can be varied so that smaller spirals might be embedded in larger ones, one pattern can be reflected or rotated to create different designs, and different patterns can be intermixed to create more elaborate designs. Examples of these possibilities are shown in Fig. 4.18B. Other construction rules can produce "sawtooth" shapes shown in Fig. 4.18C, which could also be used to create larger designs by reflecting or rotating the original pattern and connecting them together. Here, groups of three triangles alternate vertically with groups of two triangles as shown in Fig. 4.18D. What is common to diagonal grids is the ability to vary the size and to interleave patterns to create the final design. These are open to continuous elaboration, much like auditory rhythms.

A second example of the "openness" of visual designs is found in the Cathedral of Palermo described by Garofalo (2017). Smaller hexagons fitted within a larger hexagonal yield a pattern of hexagons and six-sided stars (Fig. 4.19). A third example is the Sierpinski triangle. We start with an equilateral triangle and then embed a single equilateral triangle one half the size of the original. The process is then continued; at the next stage the next three embedded triangles are one quarter the size of the original, and at the following stage, each of the nine embedded triangles are one eighth the size of the original also shown in Fig. 4.19. This process would continue; the triangles at each stage would be one half the size of the triangles at the previous stage. The Sierpinski triangle is a fractal because the identical pattern occurs at increasingly smaller scales. What is important to us is the analogy to metric rhythms. Each level of the metric grid splits the timing interval in half for double meter or thirds for triple meter so that the meter organizes each beat. Here, the overall figure is organized by the constant reduction of the embedded triangles.

4.4.2 Islamic Tiling Patterns

Islamic visual patterns share many properties with Celtic patterns. Islamic patterns are space filling and emphasize equilateral triangles, squares, and equilateral hexagons that can completely fill a space as circles cannot (Critchlow, 1976). In common with the Celtic patterns, the shapes are layered so that they occur in many orientations and sizes and that layering yields the sense of rhythm and repetition.

A good starting point would be the drawing shown in Fig. 4.20. Here are three circles within a larger circle, and within each of the smaller circles is one of the other space-filling shapes.

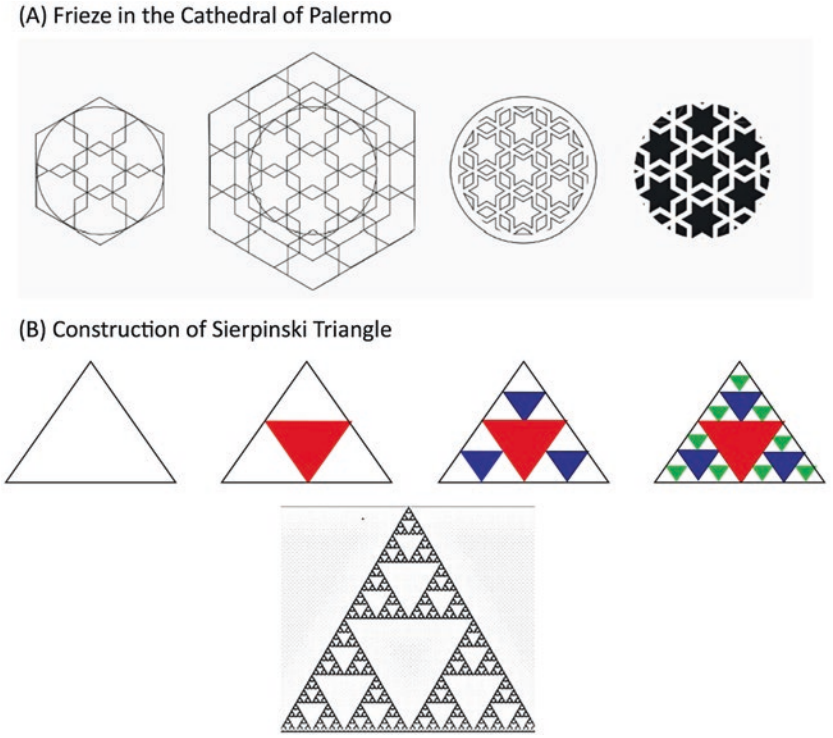


Fig. 4.19 (A) Frieze in the Cathedral of Palermo. In the second panel from the left, three, four, and five smaller hexagons are drawn within a larger hexagon. Due to this embedding, emergent stars result from the overlap of the smaller hexagons. The same kind of superposition of rhythmic lines leads to complex rhythms. (B) In the construction of the Sierpinski triangle, at each stage the embedded equilateral triangles are one half the size of those at the previous level. (Reproduced from Garofalo, 2017. CC license)

For our purposes, the importance of the three shapes is in their ability to be embedded within themselves. As illustrated in Fig. 4.20, the triangles, squares, and hexagons contained within circles can produce all of these space-filling designs at multiple sizes. This is also true for circles, even though there will be gaps in the surface.

Finally, there are combinations of the triangles, squares, and hexagons that fill the entire space. Variants of these combinations form the basis of many patterns found on walls and surfaces. One basic pattern is shown In Fig. 4.21A, and one shading scheme is illustrated in Fig. 4.21B. The same shading scheme is reproduced multiple times in Fig. 4.21C and can give a sense of the rhythm of the design.

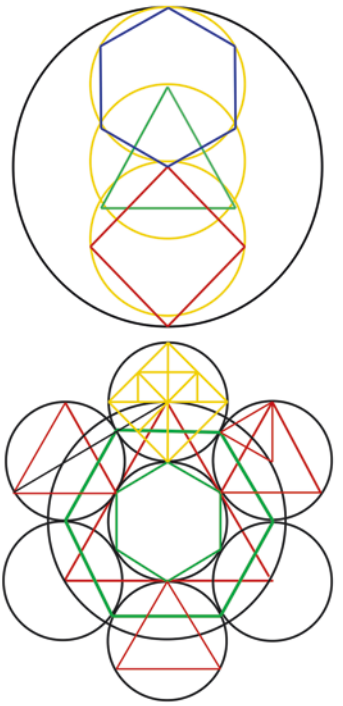


Fig. 4.20 The space-filling squares, triangles, and hexagon are embedded in a circle. Each shape has a particular religious significance. Squares (yellow), hexagons (green), and triangles (red) can be embedded in multiple sizes that yield new shapes making use of sides of other shapes. (Adapted from Critchlow, 1976. Pages 19 & 150)

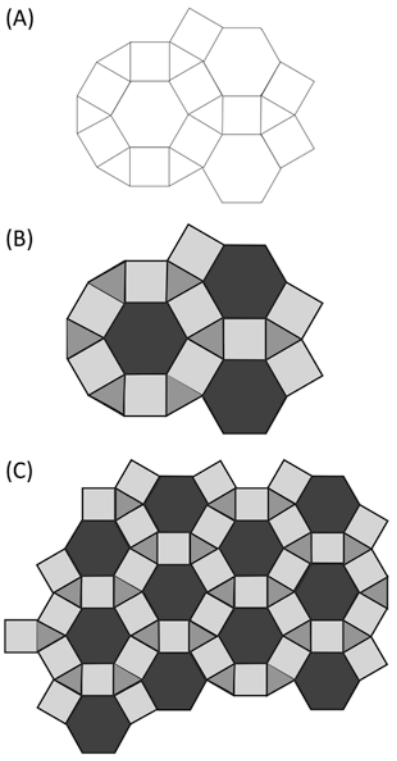


Fig. 4.21 The basic pattern is shown in Figure 4.21A. One possible shading design is illustrated in Figure 4.21B and the same shading design is reproduced multiple times at a smaller scale in Figure 4.21C. (Adapted from Critchlow, 1976. Pages 119 & 123)

4.5 Summary

Rhythms always seem to engage the listener or viewer, and create figures in time and space. Like all other perceptual acts, hearing and seeing rhythms are multilayered. One can zoom in and out; it is possible to react to every beat or to beats widely separated in time, and it is possible to view a narrow spatial field or a wide one. Listeners construct layers to hear the beat and meter even for an isochronous sequence of identical tones, and for more complex polyrhythms and non-metric rhythms. In similar fashion, viewers construct layers to group the dots into larger, more encompassing groups even for a series of equidistant identical dots and for more complex visual patterns that repeat in space. But, musical and visual beats do not simply match auditory or visual sensory features. The beats emerge from the interaction of sensory features and perceptual acts.

References

Barton, S., Getz, L., & Kubovy, M. (2017). Systematic variation in rhythm production as tempo changes. *Music Perception, 34*, 303–312. https://doi.org/10.1525/mp.2017.34.3.303

Bolton, T. L. (1894). Rhythm. *American Journal of Psychology, 6*, 145–238.

Brumm, H., & Slater, P. (2006). Ambient noise, motor fatigue, and serial redundancy in chaffinch song. *Behavioral Ecology and Sociobiology, 60*, 475–481. https://doi.org/10.1007/s00265-006-0188-y

Carello, C., & Turvey, M. T. (2017). Useful dimensions of haptic processing 50 years after "the senses considered as perceptual systems". *Ecological Psychology, 29*(2), 95–121. https://doi.org/10.1080/1040713.2017.1297188

Chen, L. (2009). *Crossmodal temporal capture in visual and tactile apparent motion: Influence of temporal structure and crossmodal grouping* (PhD). Muchen, Germany: Ludwig-Maximilians.

Cholewiak, R. W., & Collins, A. A. (2000). The generation of vibrotactile patterns on a linear array: Influences of body site, time, and presentation mode. *Perception & Psychophysics, 62*, 1220–1235.

Critchlow, K. (1976). *Islamic patterns: An analytical and cosmological approach.* London, UK: Thames & Hudson.

DeLong, L. (2013). *Curves: Flowers, foliates & flourishes in the formal decorative arts.* New York, NY: Bloomsbury USA.

Fitch, W. T. (2018). Four priniciples of biomusicology. In H. Honing (Ed.), *The origins of musicality* (pp. 23–48). Cambridge, MA: MIT Press.

Fox, D., & Weissman, D. (2007). *The great family songbook for piano and guitar.* New York, NY: Black Dog & Leventhal.

Garofalo, V. (2017). *Visual perception and Graphic analysis. The patterns and inlays in the Cathedral of Palermo.* Paper presented at the Inerbationaal and Interdisciplinary Conference IMMAGINI? Image and Imagination between Representation, Communication, Education and Psychology, Brixen, Italy.

Geldard, F. A., & Sherrick, C. E. (1972). The cutaneous "rabbit": A perceptual illusion. *Science, 178*(4057), 178–179.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston, MA: Houghton-Mifflin.
Handel, S. (1984). Using polyrhythms to study rhythm. *Music Perception, 1*, 465–484.
Handel, S. (1993). The effect of tempo and tone duration on rhythm discrimination. *Perception & Psychophysics, 54*, 370–382.
Handel, S., Todd, S., & Zoidis, A. M. (2012). Hierarchical and rhythmic organization in the songs of humpback whales (Megaptera novaeangliae). *Bioacoustics, 21*, 141–156. https://doi.org/10.1080/09524622.2012.668324
Hannon, E. E., & Trehub, S. E. (2005). Metrical categories in infancy and adulthood. *Psychological Science, 16*, 48–55.
Harrar, V., & Harris, L. R. (2007). Multimodal Ternus: Visual, tactile and visio-tactile grouping in apparent movement. *Perception, 36*, 1455–1464. https://doi.org/10.1068/p5844
Heed, T., & Azanon, E. (2014). Using time to investigate space: A review of tactile temporal order judgments as a window onto spatial processing in touch. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00076
Hein, E., & Moore, C. M. (2012). Spatio-temporal priority revisited: The role of feature identity and similarity for object correspondence in apparent motion. *Journal of Experimental Psychology: Human Perception and Performance, 38*(4), 975–988. https://doi.org/10.1037/a0028.197
Hirsh, I. J., & Sherrick, C. E. (1961). Perceived order in different sense modalities. *Journal of Exerimental Psychology, 62*, 423–432.
Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*, 201–211.
Johansson, G. (1975). Visual motion perception. *Scientific American, 232*, 76–88.
Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review, 83*, 323–335.
Kalender, B., Trehub, S. E., & Schellenberg, E. G. (2013). Cross-cultural differences in meter perception. *Psychological Research, 77*, 196–203. https://doi.org/10.1007/s00426-012-0427-y
Kramer, P., & Yantis, S. (1997). Perceptual grouping in space and time: Evidence from the Ternus display. *Perception & Psychophysics, 59*(1), 87–99.
Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review, 106*, 129–159.
Leman, M., & Naveda, L. (2010). Basic gestures as spatiotemporal reference rames for repetitive dance/music patterns in Samba and Charleston. *Music Perception, 28*, 71–91. https://doi.org/10.1525/mp.2010.1.71
Maes, P.-J., Leman, M., Palmer, C., & Wanderley, M. M. (2014). Action-based effects on music perception. *Frontiers in Psychology, 4*, 1–14. https://doi.org/10.3389/fpsyg.2013.01008
Marthevon, N., Casey, C., Reichmuth, C., & Charriuer, I. (2017). Northern elephant seals memorize the rhythm and timbre of their rivals' voice. *Current Biology, 27*, 2065. https://doi.org/10.1016/j.cub.2017.06.005
Miller, L. E., Montroni, L., Koun, E., Salemme, R., Hayward, V., & Farner, A. (2018). Sensing with tools extends intersensory processing beyond the body. *Nature, 561*(13 September), 239–242. https://doi.org/10.1038/s41586-018-0460-0
Mungan, E., Yazici, Z. F., & Kaya, M. (Uğur). (2017). Perceiving boundaries in unfamiliar Turkish Makam music: Evidence for gestalt universals? *Music Perception, 34*, 267–290. https://doi.org/10.1525/mp.2017.34.3.267

Patel, A. D., Iversen, J. R., Bregman, M. R., & Schulz, I. (2009). Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Current Biology, 19*, 827–830. https://doi.org/10.1016/j.cub.2009.03.038

Petersik, J. T., & Rice, C. M. (2006). The evolutions of a perceptual phenomenon: A case history using the Ternus effect. *Perception, 35*, 807–821. https://doi.org/10.1068/p5522

Povel, D., & Essens, P. (1985). Perception of temporal patterns. *Music Perception, 2*, 411–440.

Povel, D., & Okkerman, H. (1981). Accents in equitone sequencies. *Perception & Psychophysics, 30*, 565–572.

Prince, J. B., & Schmuckler, M. A. (2014). The tonal-metric hierarchy: A corpus analysis. *Music Perception, 31*, 254–270. https://doi.org/10.1525/MP.2014.31.254

Repp, B. (2013). Sensorimotor synchronization: A review of recent research (2006–2012). *Psychonomic Bulletin & Review, 20*, 403–452. https://doi.org/10.3758/s13423-012-0371-2

Tetlow, A. (2013). *Celtic pattern: Visual rhythms of the ancient mind* (pp. 26–27). New York, NY: Bloomsbury USA.

Toiviainen, P., Luck, G., & Thompson, M. R. (2010). Embodied meter: Hierarchical eigenmodes in music induced movement. *Music Perception, 28*(1), 59–70. https://doi.org/10.1525/mp.2010.28.1.59

Trojan, J., Getzmann, S., Moller, J., Kleinbohl, D., & Hölzl, R. (2009). Tactile-auditory saltation: Spatiiotemporal integration across sensory modalities. *Neuroscience Letters, 460*, 156–160. https://doi.org/10.1016/j.neulet.2009.05.053

Trojan, J., Stolle, A. M., Kleinbohl, D., Morch, C. D., Arendt-Nielsen, L., & Holzl, R. (2006). The saltation illusion demonstrates integrative processing of spatiotemporal information in thermoceptive and nociceptive networks. *Experimental Brain Research, 170*, 88–96. https://doi.org/10.1007/s00221-005-0190-z

Vernes, S. C. (2017). What bats have to say about speech and language. *Psychonomic Bulletin & Review, 24*, 111–117. https://doi.org/10.3758/s13423-016-1060-3

Wilson, M., & Cook, P. F. (2016). Rhythmic entrainment: Why humans want to, fireflies can't help it, pet birds try, and sea lions have to be bribed. *Psychonomic Bulletin & Review, 23*, 1647–1659. https://doi.org/10.3758/s13423-016-1013-x

Woodrow, H. (1909). A quantitative study of rhythm. *Archives of Psychology, 14*, 1–66.

Yamamoto, S., & Kitazawa, S. (2015). Tactile temporal order. *Scholarpedia, 10*. https://doi.org/10.4249/scholarpedia.8249

Yamamoto, S., Moizumi, S., & Kitazawa, S. (2005). Referral of tactile sensation to the tips of L-shaped sticks. *Journal of Neurophysiology, 93*, 2856–2863. https://doi.org/10.1152/jn.01015.2004

Zwamborn, E., & Whitehead, H. (2017). Repeated call sequences and behavioural context in long-finned pilot whales off Cape Breton, Nova Scotia, Canada. *Bioacoustics, 26*, 169–183. https://doi.org/10.1080/09524622.2016.1233457

CHAPTER 5

Color, Timbre, and Echoes: How Source-Filter Processes Determine Why We See What We See and Hear What We Hear

Imagine that you are looking though a cardboard tube and all you can see is a red spot on a wall. It could be a wall painted red illuminated by white light, or a white wall illuminated by a red spotlight. In similar fashion, imagine you hear a low-pitched pure tone. It could be one tone, but it could also be a complex sound that has been modified electronically or filtered by the environment.

This sort of ambiguity is inherent in all forms of perception. We have encountered this issue in all previous chapters. How can we decide if a sound we hear comes from one or more sources given that sound waves combine? How can we tell if an object that appears to get bigger is moving toward us or is expanding? Usually we hear but one sound and see a rigid object approaching even with little or no previous experience. In most cases though, it is the context that helps resolve these ambiguities. This is same argument made in Chap. 3 about resolving multistable percepts.

There are two issues here:

1. Why discuss color constancy, timbre recognition, and echolocation together given that they appear to be quite different properties and processes? Here we will briefly detail the source-filter model to better understand their commonalities.
2. How do we use contextual information to accurately perceive the properties of objects and events?

Before starting, let us reconsider auditory stream segregation to emphasize how the contexts in which we hear and see determine our organization of those stimuli. If two tones start and stop at the same time, invariably those tones are perceived as a single complex tone as illustrated in Fig. 5.1A. But as shown in

Electronic Supplementary Material: The online version of this chapter (https://doi.org/10.1007/978-3-319-96337-2_5) contains supplementary material, which is available to authorized users.

S. Handel, *Perceptual Organization*,
https://doi.org/10.1007/978-3-319-96337-2_5

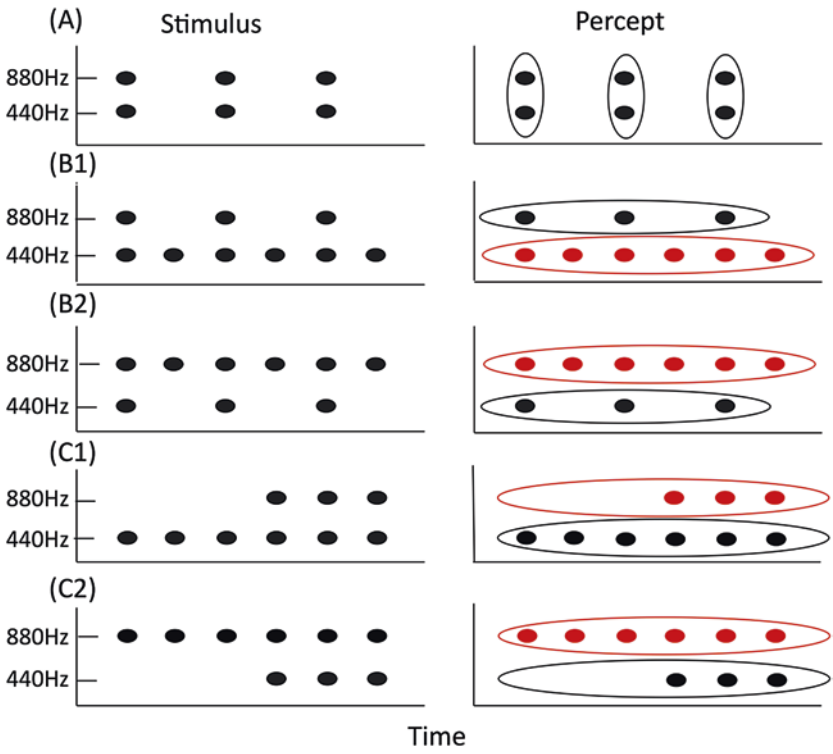


Fig. 5.1 If two sounds occur synchronously, the two sounds are combined as shown in (A). But if the sounds occur at different rates, then the sounds are heard separately (B1 & B2). Moreover, if one sound occurs before the other the sounds are heard separately even after they become synchronous (C1 & C2)

Sound Files 5.1: Examples of streaming as a function of context and presentation rate corresponding to Fig. 5.1

Fig. 5.1B and 5.1C if one of the two tones is presented at twice the rate of the other, the two-tone percept is split apart, and the two tones are heard individually. Either tone can used to split the complex. The "faster" tone provides the context, so that the combination tone is "understood" to be composed of one consistent repeating tone plus an intermittent one. Another way to split the two-tone complex is to initially present one of the two tones (Haywood & Roberts, 2013). In Fig. 5.1C1 the initial tones suggest that the lower-pitch tone is continuous and the higher-pitch tone comes from a different source, and in Fig. 5.1C2 the initial tones suggest that the higher-pitch tone is continuous. Again, this is the same issue discussed in Chap. 2, where within-modality organization usually was stronger than multisensory organization.

5.1 Color and Timbre

Equivalences between seeing and hearing are slippery; several are equally plausible. We could match color to pitch based on the physical fact that both are based on frequency. Moreover, color and pitch are the classic examples of

secondary qualities. The incoming visual electromagnetic vibrations and acoustic mechanical vibrations are neutral; they merely yield neural firings. They need to be interpreted to create color and pitches. Newton famously said, "For the rays, to speak properly, are not coloured. In them there is nothing else than a certain power and disposition to stir up a sensation of this or that colour." We can generalize this to other senses, "for the touch sensation to speak properly do not have shape or roughness," "for the vibrations, to speak properly, do not have pitch," and "for the peppers or perfumes, to speak properly, do not have tastes or smells." Thus, color, pitch and timbre, touch, taste, and smell are in the mind, not in the light vibrations, not in the air or skin vibrations, and not in the chemical composition. If we accept this correspondence, we can investigate the perceptual properties of color and pitch, especially the sensitivity and discriminability between colors and pitches.

While color and pitch are undoubtedly secondary qualities, I prefer to match color to timbre because my belief is that color and timbre are properties of objects. This matches our commonsense notions: color and timbre are "real" properties of the world that guide our actions in a world of things. In Chap. 2 we speculated how "blobs" of color lead an infant to split the visual field into rigid objects, and this suggests that different timbres lead infants to split the auditory scene into separate sources. The goal would be to describe the perceptual and cognitive processes that allow us to recover the color (i.e., the fixed surface reflectance of an object) and timbre (i.e., the resonances of an object) in spite of variation in the physical stimulation and environmental factors.

We will use the term *reflectance* for light and *resonance* for sound. The surface of any material contains molecules that can be excited by energy at specific wavelengths of the incident light. If the wavelengths of the incident light match those of the surface molecules, those incident wavelengths are absorbed, generate heat, but are not reflected. If the incident wavelengths do not match the wavelengths of the surface molecules, those wavelengths are briefly absorbed but then are radiated back from the surface and those wavelengths are what we see. The frequencies and intensities of the radiated wavelengths form the reflectance spectrum of the surface. Black objects that absorb all wavelengths are hot in the sun because they convert all of the incident energy into heat, and white objects that radiate back all incident wavelengths are cool.

With sound, we need to apply mechanical energy to get an object to vibrate. Suppose we set a wooden plate (resembling the top of violin) vibrating using a mechanical stimulus. As we change the frequency of the vibration, the plate will vibrate in different patterns depending on the frequency and the material and shape of the plate. The maximum amplitude of each type of plate vibration occurs at its *resonant* frequency. Thus, the electromagnetic frequencies of light that *do not* induce the surface molecules to vibrate yield the reflectance spectrum, but the mechanical pressure vibrations that *do* induce surface vibrations yield the air pressure waves at the resonant frequencies and those frequencies are what we hear.

We still have the problem of connecting the physical nature of color and sound production (i.e., the amount of energy at different wavelengths or frequencies)

to the subjective properties of color and timbre. One way is to suppose that objects have the ability or disposition to create color or timbre, but that those properties come about in a context that includes both a perceiver and an environment. A color-blind person, or one with cataracts, or a primate with different sorts of receptors, will perceive a different color. In similar fashion, an individual with hearing loss, or an animal with a different kind of auditory system, will hear a different sound.

We do not understand how the neural system creates the color or timbre in our heads. What we can do is correlate the firing of visual and auditory cortical cells to perceptual discrimination and similarity judgments, but at present we are at a loss as to how neural firings produce appearances and experiences (see Dedrick, 2015, for an elegant discussion of the philosophical issues). Moreover, the descriptions of color and timbre simply are attempts to mirror or echo the percepts, an example of sound source symbolism discussin Chap. 2.

Why are we able to perceive independent properties of objects? Why does the visual system "calculate" location, shape, and color; why does the auditory system "calculate" timbre, pitch, and loudness; and why does the haptic system "calculate" weight, roughness, and shape? The information required to assess each property that is intermingled in the stimulus would presumably require a separate neural pathway. Alternately, it is plausible that the sensory and perceptual systems could yield a Gestalt not accessible to analysis into independent attributes. I suggest that these properties can provide independent ways to break the visual and auditory fields into objects. For seeing and hearing, the constancy of one such property allows us to segment the visual and auditory scene into objects. For example, viewers can link together the surfaces of one solid object based on color in spite of changes due to illumination, motion, and location, or link the surfaces based on common motion in spite of changes in the other properties. For hearing, listeners can segment the auditory world into sources based on timbre in spite of changes in location, loudness, and pitch or location. For touching, surface properties such as roughness can link complex surfaces into one object or into abutting objects. Color and timbre help create the coherence of objects and their properties in a changing world. Such coherence makes imaginative comparisons of such derived properties like "looks just like a lemon but tastes like a pineapple" or "looks just like Bob Dylan but sounds just like Johnny Cash" so easy to imagine and understand.

It is not surprising that color and timbre are defined in similar fashion, but what is surprising is that both were originally defined by exclusion. Color was "that aspect of visual perception by which observers distinguish differences between equally bright, structure free fields of view of identical size and shape" (Kaiser & Boynton, 1996, page 315). A recent definition is that perceived color is "those characteristics of a visual perception that can be described by attributes of hue, brightness (or lightness) and colorfulness (or saturation or chroma)."

Timbre was "the quality of a sound by which a listener can tell that two sounds of the same loudness and pitch are dissimilar" (American National Standards Institute, 1973, page 56). Like the definition of color, timbre now can be defined in terms of the timing and distribution of energy at different

frequencies. Each of these definitions tells us that color and timbre are defined by discrimination, but neither tells us what color or timbre is. What is important here for color is that in some way it is invariant across changes in light; what is important for timbre is that in some way it is invariant across changes in pitch and loudness. Color and timbre become the result of perceptual acts.

5.2 Production of Color and Timbre: The Source-Filter Model

5.2.1 Ambiguity of Color and Timbre

The lesson from Chap. 2 was that the proximal stimulus at the eye and ear was underdetermined; the same proximal stimulus could have come from many different distal objects and sounds. Later we will consider how the contextual information allows us to choose among the possibilities, but right now it is worth considering the simple source-filter model that clarifies how ambiguous visual and auditory stimulation are created.

The source-filter model is conceptually simple. A **source** is characterized by energy at different wavelengths or frequencies. Wavelength (distance between peaks) equals speed of the wave divided by the frequency in hertz. The speed of light is roughly 300×10^6 meters/sec while the speed of sound is roughly 340 meters/sec.

The color spectrum ranges from blue (wavelength **(λ)** = 400×10^{-9}meters (nm) and frequency=7.5×10^{14} Hz) to red (wavelength **(λ)** = 700×10^{-9}meters and frequency = 4.3×10^{14} Hz).

The sound spectrum ranges from the low pitch 20 Hz tone (wavelength **(λ)** = 17meters) to the high pitch 20,000 Hz tone (wavelength **(λ)** = 0.017meters). Because wavelength and frequency are interchangeable, it is possible to use either to represent the energy distribution. What is confusing is that wavelengths are used to label colors, but frequencies label sounds.

In Fig. 5.2, the white **source** beam is composed of energy at all wavelengths, while the red **source** beam is composed of energy only at the longer red wavelengths. The wall **filter** reflects the incident wavelengths to different degrees. In the example here, the red wall absorbs all of the wavelengths except for the red ones, which are reflected off the wall, while the white wall reflects all of the wavelengths that strike it. In either case, only red wavelengths reach an observer.

The output at each frequency is equal to the source-input energy at that wavelength **(λ)** or frequency **(F)** multiplied by the percentage of energy reflected at that wavelength (or frequency). We do the multiplication at each wavelength or frequency separately.

Output Energy **(λ)** = Source Energy **(λ)** × Filter Percentage **(λ)** or

Output Energy **(F)** = Source Energy **(F)** × Filter Percentage **(F)**

We can imagine a similar outcome for sounds. Recall from Chap. 2 that complex harmonic tones are made up of the sum of simple tones that are multiples of the fundamental frequency F_0. One type of complex tone is termed,

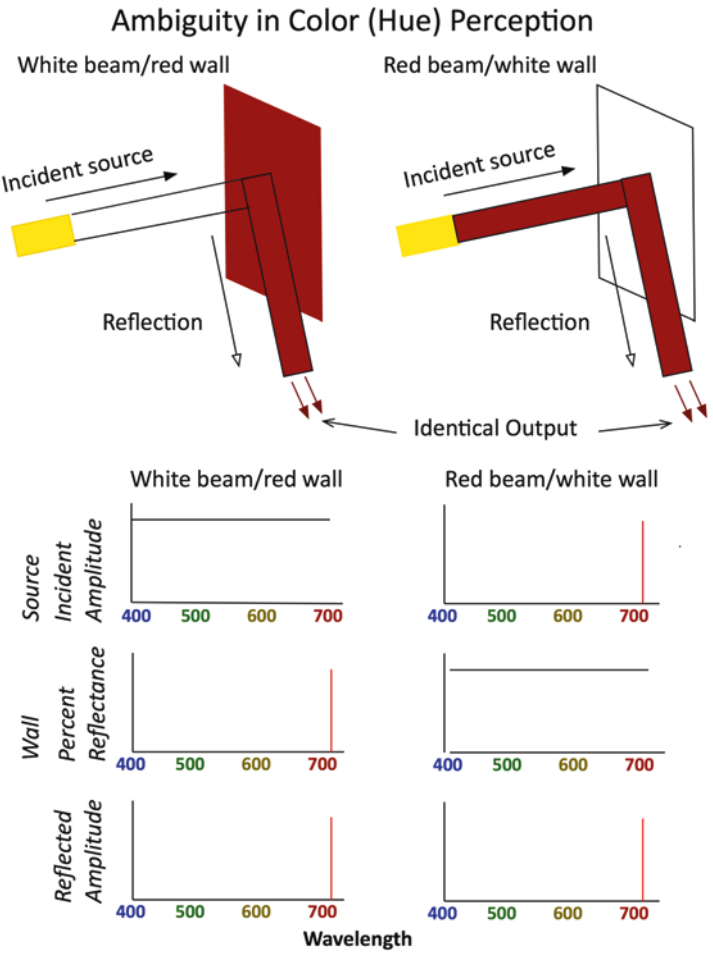


Fig. 5.2 A white beam source reflecting off a red wall yields the same sensation as a red beam reflected off a white wall. For the white beam/red wall, the source energy has equal energy at all wavelengths, but the red wall reflects only the light energy in the red region. The result is beam of red light. For the red beam/white wall, the source consists only of energy in the red region while the wall reflects all wavelengths equally. The reflected amplitude is found by multiplying the incident amplitude by the percentage reflectance at each wavelength. The result also is a beam of red light

for obvious reasons, a square wave, as shown in Fig. 5.3. A square wave is composed of the fundamental and all of the odd harmonics, $3F_0$, $5F_0$, $7F_0$, and so on. The amplitude of each harmonic is inversely proportional to its frequency; namely 1/3, 1/5, and 1/7 for the first three odd harmonics. In Fig. 5.3, the top panel shows a square wave that is the sum of many harmonics. The next panel illustrates the simple harmonic wave with the same frequency and the third panel illustrates $3F_0$ with 1/3 the amplitude. The next panel shows

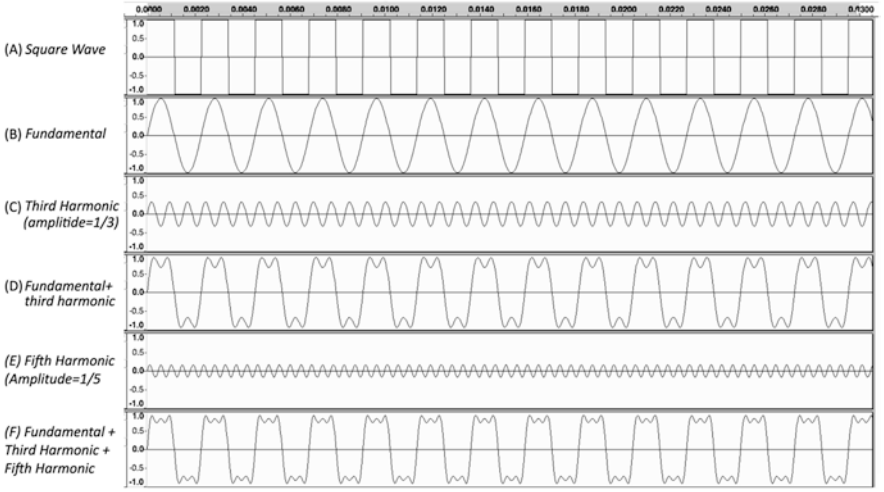


Fig. 5.3 The top panel shows the complete square wave. The second and third panels show the first components: a sine wave at F_0 and $3F_0$ with amplitudes of 1 and 1/3. The fourth illustrates the sum of F_0 and $3F_0$. The fifth panel shows $5F_0$ with amplitude 1/5 and the last panel illustrates the sum of F_0, $3F_0$, and $5F_0$

Sound Files 5.3: The fundamental and the odd harmonic frequencies that add to produce a square wave

the sum of F_0 and $3F_0$, not perfect but partway to producing a flat top. The next panel portrays $5F_0$ with 1/5 the amplitude. The last panel shows the sum of F_0, $3F_0$, and $5F_0$, a better approximation. But if the source square wave is passed through a low-pass filter that attenuates the higher frequencies all that is left is the original F_0. A simple pure tone therefore could just be either a pure tone or a complex tone that has been filtered.

Contextual information is essential to remove these ambiguities for both visual and auditory perception. We cannot discriminate between the white beam/red wall and red beam/white wall or the complex filtered tone and a pure tone without expanding our view, that is, without seeing a larger expanse of the wall or hearing additional sounds. What we see and hear is not only based on the signals that our eyes and ears send to our brain, but is also influenced strongly by the context of the visual and auditory scenes, on our previous knowledge, and our expectations (recall the discussion of prior probabilities and Helmholtz' concept of *unconscious inference*).

Now consider more realistic situations. Suppose we reflect different light sources off a white wall. The energy distribution of various illuminations is shown in Fig. 5.4. The wavelength distribution of morning daylight is continuous, with a peak toward the blue region (due to the sky); the distribution of evening light peaks at the red region, as do warm incandescent bulbs, while the wavelength distribution of fluorescent is peaky. The reflected wave reaching the eye from the white wall would closely match the incident light. If we reverted to looking at a white wall using only a small diameter tube, then the wall will

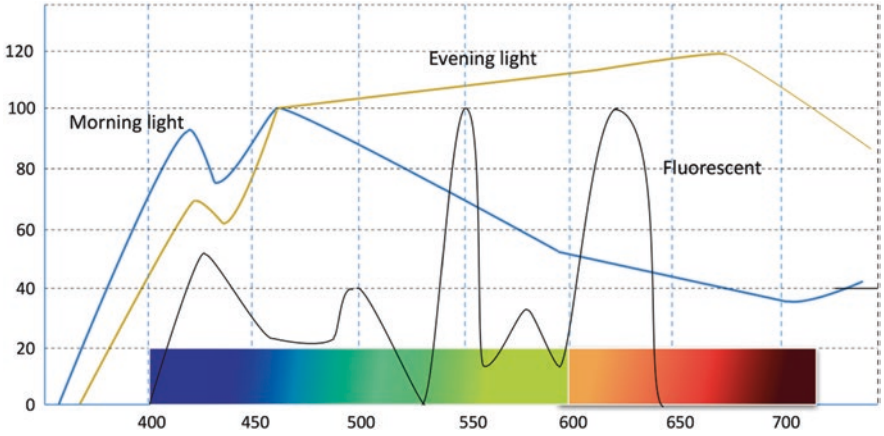


Fig. 5.4 The frequency spectra of morning daylight, evening light, and a typical fluorescent light

seem to look differently colored under those illuminations: bluish under normal daylight, reddish under the evening sky, and yellowish under a fluorescent tube (a combination of the various peaks). In no case would it look white. The reason is that being restricted to "tube" vision eliminates the possibility of judging the "true" color of the wall as distinct from the color of the illumination. The light reaching our eyes does not separate the illumination from the reflection. But the variously colored objects in our normally dense environment provide the context that allows us to split the source-filter output into the light source and the surface reflection, that is, the filter.

The sound production of a violin involves several interlocking steps. The initial bowing creates a set of discrete vibrations on the strings. These vibrations force the bridge into motion, which in turn creates vibrations in the top and bottom plates that act like the cone of a speaker radiating the sound that reaches our ears. There are frequencies at which the connected top and bottom plates vibrate maximally (i.e., the sound body resonances) and other frequencies at which the vibration amplitude is minimal.

Because the vibrations due to the bowing action on the string occur at discrete frequencies, the sound body vibrations due to the string vibrations also occur at discrete frequencies. As shown in Fig. 5.5, the match between the string frequencies and the sound body frequencies determines the amplitude of the output frequencies. The fundamental frequency F_0, roughly 220Hz, is the maximum vibration on the string, but because there is no body resonance at that frequency, it is not radiated to the listener.

If we wanted to perfectly represent the reflected color or radiated timbre, we would need many receptors each tuned to a specific frequency or wavelength. Even if we restrict ourselves to the visible spectrum omitting the infrared and ultraviolet light, we would need more than 300 different color-frequency receptors to represent the perceived colors. In similar fashion, if we restrict ourselves to frequencies yielding tonal perception (20 Hz to 20,000 Hz) we

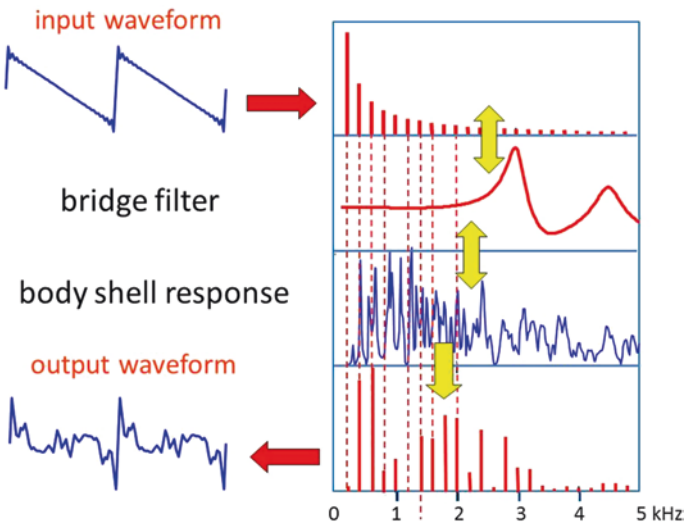


Fig. 5.5 The input waveform is modified by the bridge and body shell resonances to generate the radiated sound. The output is quite different than the input. The vertical dashed lines are the frequencies of the partials from the bowing. (Reprinted from Gough, 2016. Permission by author)

Sound Files 5.5: The input waveform due to bowing and the output waveform that is modified by the violin body

would need about 20,000 different frequency detectors. Admittedly, these numbers seem physiologically and evolutionary impossible. The strategy the visual and auditory systems use to overcome this problem is identical: make use of multiple eye receptors and ear receptors that are tuned (i.e., most sensitive) to different frequency ranges that match the physical properties of the light and sound vibrations and use the ratios of those receptors to derive the stimulating frequencies. Remember that the perceptual goal is to construct objects and guide actions; the percept does not have to be absolutely accurate.

5.2.2 The General Strategy

First, the basics. The retina is composed of two kinds of cells, the rods and cones that serve different functions. There are 120 million rods and six million cones; rods are densely spread through the retina except in the fovea, which contains only cones. The rods are adapted for low-light conditions and are more sensitive than cones. Even one photon can trigger a rod. The rods are small, but many rods converge on higher-level neurons. This convergence maximizes their sensitivity to light and minimizes noise, but reduces the capacity to detect spatial changes in the array of the light (i.e., resolution) and the ability to detect motion. Cones function at higher intensities and are organized to maximize spatial resolution. They are densely packed in the fovea, at the center of the eye, to create a detailed sampling array that matches the incoming light. Furthermore,

each cone connects to several higher-level neurons possibly yielding several maps of the light array.

Each cone has a particular frequency at which it is maximally sensitive. For example, one type of cone is most sensitive to light in the region of 500 nm, which corresponds to the perception of green. Nonetheless, light rays ranging from 400 nm (blue) to 600 nm (yellow) can also stimulate those cells. The sensitivity at 400 nm and 600 nm is lower, however an increase in the intensity of the blue or yellow light can increase the firing rate of the green cell to match that at the most sensitive frequency. But, regardless of how those cells are stimulated, it still signals "green." This creates an inherent ambiguity since it is impossible to determine which color was presented just from the firing of that one type of cell. The identical neural response could be due to an intense blue or yellow light or a weaker green light.

In primates there are four kinds of retinal cells. Rods in the periphery of the eye are achromatic, and respond to brightness differences. The firing of the rods saturate in daylight so that they do not affect color perception. There are three kinds of cones responsible for color vision in the fovea, the central region of the retina: 1. short-wavelength blue cones (400 nm); 2. medium-wavelength green cones (500 nm); and 3. long-wavelength yellow red cones (575 nm). There are many more medium- and long-wavelength cones than short-wavelength cones.

The same ambiguity occurs for the 3500 inner hair cells in the cochlea in each ear. Each cell can be characterized by the frequency at which it is most sensitive. Like the retinal cells, each cochlear cell can be stimulated by nearby frequencies presented at higher intensities. But it is almost certain that regardless of the stimulating frequency, each cell in the cochlea signals its most sensitive frequency so that the identical ambiguity as to the stimulating frequency occurs.

We can understand this problem in reverse. Every light will stimulate more than one kind of visual receptor and every sound will stimulate more than one kind of auditory receptor. How then does the cortex decide which color and pitch occurred? The trick is to attend to the ratios of the firings of the different receptors rather than the magnitudes of the individual firings.

5.2.2.1 Color Receptors

As described above, if we had just one type of receptor, it would be possible to match the firing rate caused by one color to a second color by changing the intensity of the second color. But if there is more than one receptor this becomes impossible. Suppose that there are two receptors, green and yellow-red, and we present a 500 nm green light at the intensity I = 10 and that will stimulate the green receptors, for example, 10 units. But light at that frequency will also stimulate the yellow-red receptors (575 nm) to some degree (five units). Conversely, if we present a 575 nm light, the firing rates for the green and yellow-red receptors would be five and 10, respectively. If we double the intensity of the green or yellow-red light, then the firing rate of the green and yellow-red receptors would also double. Now, the firing rate of the green receptor to the green light at I = 20 (20 units), now equals the firing rate of the green receptor to the yellow-red light at I = 40, also 20 units. If we judged color only on the basis of the

individual firing rates, then we could not distinguish between the two colors. But, we don't. We judge colors on the basis of the ratios of the firings among the receptors. The ratio of green/yellow-red firings for green light is 2:1 and the ratio for yellow-red light is 1:2. Those ratios remain roughly constant across all levels of illumination and that would be true for all pairs of receptors: blue/green, blue/yellow-red, and green/yellow-red.

These ratios also act to uncorrelate the firing of the different receptors, particularly the medium and long wavelength cones. As mentioned above, every light will stimulate more than one receptor and a stronger light will stimulate each one even more. The ratios act to eliminate this "noise" so that each pair of opponent cells can accurately represent the relative strength of the color frequencies.

There are two problems, however. First, while the ratios remain invariant across differing light levels yielding unchanging colors, we lose information about the light intensity itself. Starting at the retina and continuing through the visual system, the visual system perceives color and intensity separately. Summing the outputs of all three receptors into a separate channel as shown in Fig. 5.6 carries the illumination information.

Second, the firings of the three classes of cones do not help us perceive the blobs and the fine details in the environment. To accomplish this, opponent process cells that respond to the firing ratios among the cones carry the color information. These cells are structured in concentric circles so that stimulation of the central region leads to increased firing, while stimulation of the sur-

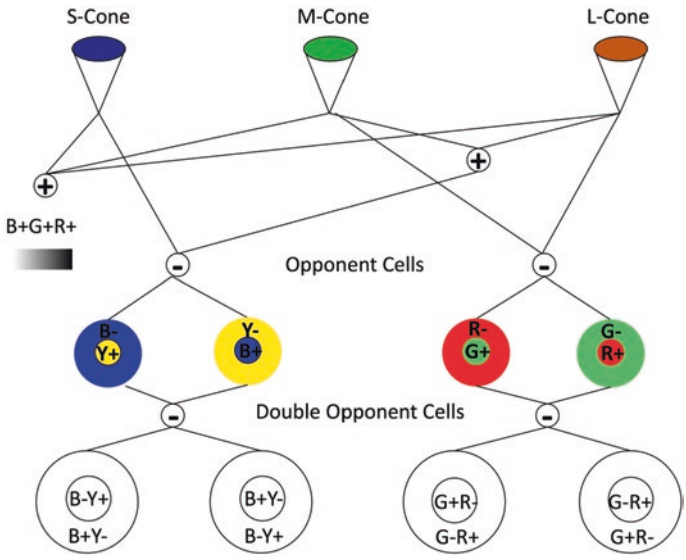


Fig. 5.6 The three cones in the fovea undergo two transformations. The first transformation creates two types of opponent cells, each with two variants, to encode color. In addition, one type of cells, B+R+G, adds the outputs of all three cones to encode lightness. The second transformation maximizes contrast by placing the two variants of each type of opponent cells in opposition

rounding region leads to decreased firing. The portrayal of opponent cells in Fig. 5.6 is that of its *receptive field*. The receptive field of an opponent cell is that retinal region where one color increases its firing rate and a different color in the surrounding retinal region reduces the firing rate. We identify the receptive field by flashing lights of different colors into the eye and measuring the change in response of the cell.

There are two types of opponent cells. The outputs of the long-wavelength (yellow-red labeled R) and medium-wavelength (green labeled G) retinal cones make up one opponent process cell, R/G, and outputs from the short-wavelength (blue labeled B) and the combination of the long- and medium-wavelength cones that yield yellow (Y) make up the second opponent process cell, B/R+G or B/Y.

Each type of opponent cell has two forms that reverse the colors in the excitatory central and inhibitory peripheral region. For example, in one form the R+/G- cells increase their firing rate when long-wavelength light falls on L-cones stimulating the central region and decease their firing when medium-wavelength light falls on M-cones stimulating the periphery. In the other form, the G+/R- cells increase their firing rate when medium-wavelength light falls on M-cones stimulating the central region and decease their firing when long-wavelength light falls on L-cones stimulating the periphery. In similar fashion, there are two forms for the B/R+G (Yellow) cells: B+/Y- or Y+/B-. This complicated process is illustrated in Fig. 5.6. The formation of the opponent cells explains complementary colors red/green and blue/yellow, and for the observation that blue, green, yellow, and red cannot be described as combinations of other colors.

Opponent cells seem best suited to identify large areas of one color, that is, blobs. (White regions would lead to a low firing rate as the excitatory and inhibitory regions cancel each other). But opponent cells would not respond to regions that undergo rapid color shifts. Suppose green leaves surrounded a red berry. The firing of the R/G opponent cells would be muted; the red berry would increase the firing due to the excitatory central region, while the green surround would decrease the firing due to the inhibitory surround. The output of the opponent cell would be just slightly above normal because the excitatory center is stronger than the inhibitory surround.

In contrast, double opponent cells shown in Fig. 5.6 combine the outputs of opponent cells and bring about contrast. The red berry increases the firing of the R+/G- center and the green leaves increase the firing of the G+/R- surround. It is a common supposition that the later evolution of the M- and L-cones in primates was due to the need to identify food in a leafy green background and the resulting double-opponent cells maximize the response to such hidden food.

5.2.2.2 Auditory Receptors

Even though the auditory system has many more receptors, the same sort of strategy, based on the pattern of firings, is probably how we derive the pitch of individual tones and the timbre of complex tones.

There are important differences, however. There are only four different retinal receptors: three cones tuned to different frequencies for daylight color

vision and one type of rod for nighttime vision. But, in the inner ear there are several thousand hair cells tuned to different frequencies. We can understand this difference in terms of the visual and auditory spectrum. Across the visual wavelengths, the most common visual sources (i.e., sunlight) are smooth and continuous (see Fig. 5.4). As the illumination varies, the relative amount of energy at each frequency changes relatively smoothly so that variation can be reflected by changes in the blue/yellow and red/green ratios. A small number of receptors integrating energy across a range of wavelengths can maintain an adequate representation of the light spectrum.

In contrast, auditory sources and filters have energies at discrete frequencies and in addition the source energy and the sound body resonances do not change in a simple way as the intensity changes. On a violin, both the amplitudes of a string and the resonances of the sound body change when bowing or plucking at different intensities at the same frequency. Even the way the violin is being held can change resonances. There are many unique sounds, and we need finer resolution achieved by the greater number of cochlea cells to discriminate among them.

5.2.2.3 Comparison of Visual and Auditory Receptors

These differences between the type and number of visual and auditory receptors have two important consequences:

1. For humans, any color, with the exception of blue, green, yellow, and red, can be matched by the sum of three other colors at normal illumination levels. Such a match is termed a *metamer*. Cornsweet (1970) elegantly explains how the number of colors necessary to match any single color is equal to the number of independent color receptors. For animals with just one receptor, two colors can be matched by adjusting the intensity of each one to yield the same firing rate; but for humans, the sum of three colors are necessary to match another one. Animals with four types of receptors will discriminate between a single color and a match of three colors that humans cannot tell apart. That is why color television uses three beams to create all the visible colors (for humans, but not chickens with four receptors).

   Following Cornsweet's (1970) argument, since any given sound source would stimulate many of our several thousand auditory receptors, we would need to sum several thousand individual frequencies to match it. Although this is theoretically possible, it is impractical and probably why sound synthesizers rather than attempting to create the sounds of instruments from individual frequencies store a recording of each note.

   As described above, metamers in human color vision are constructed by adding three colors to match the hue of an individual color. Metamers in human sound perception are of a different sort because we cannot add a 3000 Hz tone to a 1000 Hz tone to match a 2000 Hz tone. But we can produce loudness metamers by varying the amplitudes of tones at different frequencies. The human auditory system is maximally sensitive

in the 2000–3000 Hz frequency range so that we can match the loudness of a 1000 Hz tone to a 3000 Hz by increasing the loudness of the 1000 Hz tone. What is critical to understand is that metamers for any single property are specific to context. A color match at one illumination may not, and usually does not, match at different illuminations. In similar fashion, a loudness match at one amplitude will probably not match at different amplitudes. Furthermore, metamers for one property rarely create matches for any other property. Color mixtures that match a single hue will hardly ever match in saturation or lightness, and two sets of tones that match in loudness will rarely match in consonance or timbre.

2. We think that the colors and timbres in what we see and hear are real. It is only the illusions of color and timbre that convince us that the perception of color depends on the illumination as well as the color of surrounding objects, and that the perception of timbre depends on the frequency and intensity of the sound as well as sound energy of other objects in the surround. Perceived color is a second-order calculation based on the relative ratios of absorption in different parts of the visual field and is simultaneously influenced by our interpretation of the object's shape, depth, and orientation. In the same way, timbre is a second-order calculation based on the object's sound, and the interpretation of the distance and any background sound. Color and timbre perception should be understood as being part of the general problem of figure-ground organization that constructs objects.

   An extensive set of visual illusions is found at www.michaelbach.de/ot/index.com

5.3 Color Constancy

5.3.1 Reflections

The ability to figure out the "true" surface color and discount the source illumination is called *color constancy*. Color constancy is fundamental to survival; it allows us to detect ripe fruit and nuts, to avoid predators and poisonous plants regardless of the time of day or shading. As we shall discuss, constancy is not perfect but a compromise between the rays at the eye from the actual source illumination being reflected by the surface color of the object and the rays that would have occurred under a neutral source of illumination.

Consider the difficulties of color constancy. First, how do we recognize that the color of an object is the same under very different light sources? Even homogeneous surfaces of the same object can appear different depending on the position of the light source. As illustrated in Fig. 5.7, the surfaces of the glossy blue coffee cup would appear to be different colors if we examined them separately, but seem identical when viewed as a cup. The appearance of the surface of the cup comes from two kinds of reflection, specular and matte. Specular reflection occurs for smooth or mirror-like surfaces, where the angle of reflection equals

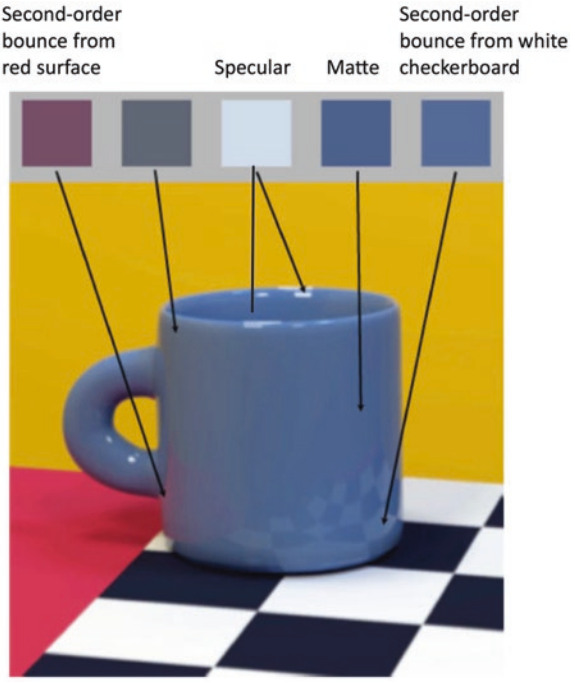


Fig. 5.7 A glossy cup creates several different reflections and yet is perceived as one cup under one illuminant. (Adapted from Brainard & Maloney, 2011. © Association for Research in Vision and Ophthalmology (ARVO))

the angle of incidence. We assume that the specular reflection has the same spectrum as the incident illumination. While it does not give information about the object color, specular reflection can allow the observer to isolate the spectrum of the illuminant. The specular reflection is shown as the middle color above the cup, so that white seems a good estimate of the illumination. Matte reflection, on the other hand, is due to embedded colorants in the surface. Parts of the spectrum are simply absorbed while other parts are reflected. Matte reflection occurs for rougher surfaces and is assumed to be equal in all direction. This is shown in the fourth location in the figure above the cup.

Another aspect of the perceived object color is due to indirect or secondary bounce illuminations. The first color above the cup looks maroon due to the initial reflection off the red part of the surface that is subsequently reflected off the blue cup. Another example of the effect of the secondary illumination occurs for the fifth color that depicts the reflection off the checkerboard section of the surface. The indirect or secondary reflections can approach 15percent of the total reflected energy. Given the differences in the reflected waves, why do we see a solid cup rather than a set of disconnected surfaces? The answer is unclear; it may be due to the opponent cells combining connected (as described in Chap. 2) regions of similar color.

5.3.2 *Monge's Demonstrations*

Gaspard Monge, in 1789, was the first to argue that our estimate of color was based not only on the physical light reaching the eye, but also on the context, overall illumination, and previous experience. Observers looked at a white wall through a piece of red tinted glass where Monge had placed a piece of red paper. As the glass would transmit only red light, we might expect the red paper and white wall to appear saturated red. Were a snowflake known to be white placed on the wall, we would expect it, too, to appear red. But the observer "knows" that snowflakes are white, and that prior knowledge has the effect of almost completely desaturating the snowflake, red paper, and white wall even though all are perceived through the red glass. The observer has no way of separating the actual red region from the surrounding white wall due to the colored glass so that even the actual red region looks bleached. Monge's experiment is simulated in Fig. 5.8.

The rationale for this outcome, while torturous, is as follows:

(a) When the observer looks through the red glass, the snowflake looks red.
(b) But the observer "knows" that the snowflake is white (according to Monge, this effect requires a known white object).

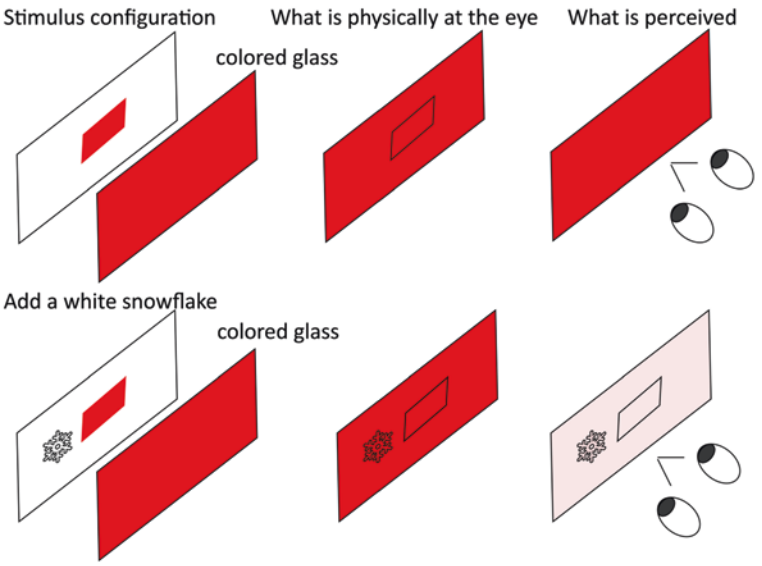


Fig. 5.8 In the first row, all that is seen is the region covered by the red glass and there is no desaturation or bleaching. The entire region looks bright red. In the bottom row, a white snowflake is placed on top of the background. The white snowflake should look red, the same color as the red paper and white background. But, because the snowflake is known to be white, all of the area seen through the glass is perceived to be white

(c) Because the snowflake looks red even though it is white, it must be illuminated by a red light.
(d) Because all the area looks red matching the snowflake, the entire area must be white.

An even more startling example of color constancy occurs when sunlight passes through a red-tinted transparent glass with an opening in the center onto a uniformly white surface. The observer is unaware of the hole. What should be seen is a white spot in the center of a red surround. But, what is seen is a green spot in the center of the red surround shown in Fig. 5.9.

Again, the logic is convoluted:

(a) If the viewer supposes the surface to be uniformly white and the transparent surface continuous, then the center of the surface ought to be red like the surround.
(b) But, the reflection at the center spot is white (from the surface). Since the viewer is unaware of the hole in the transparent surface, the center also ought to be red. To account for the white, that circular area needs to be a color such that the reflection from the supposed red light on that area yields white.
(c) Red and green are complementary colors so that their combination yields white due to the opponent process retinal cells (described above).
(d) To account for the center white hole when the rest of the surface is red, the center hole is perceived to be green.

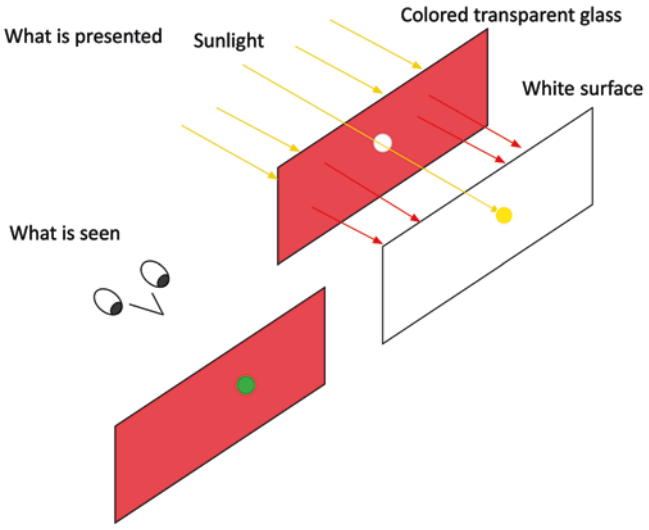


Fig. 5.9 An illustration of Monge's demonstration that context can bring about the illusion of a complementary color in the center of the surface. (The sunlight is shown as yellow for illustrative purposes)

Gelb did a similar demonstration (see excellent review by Gilchrist, 2015). Gelb illuminated a black disc with a white light and without any context the disc appeared white. After a sheet of white paper backed the disc, the disc reverted to black. Gilchrist (2015) makes the important point that the frame of reference strongly affects our perceptions. If the objects appear to be on the same plane or surface, the same illumination is likely, so differences are due to the reflectance. If they appear to be in different frames, on different surfaces, then their different appearance could be due to either the illumination or reflectance.

We are not arguing that an individual consciously goes through this logic. Instead, this is an example of what Helmholtz described as "unconscious inference." Originally, the thinking is conscious and modified by experience. After many such experiences, the thought process is condensed and automatic. Obviously, unconscious inferences can give rise to inconsistencies and misperceptions if inappropriately applied. Unfortunately as Kahneman (2011) points out, we are often unaware how these inferences (mis)shape our perceptions.

5.3.3 Asymmetric Matching

Recent work on the degree of color constancy has made use of techniques that involve asymmetric matching. Basically, a colored surface is illuminated under one source, and a set of test surfaces is illuminated under a different source. The observers must match the color of the original colored surface to one of the test surfaces. To do this successfully, the observer must abstract the surface reflectance of the original surface away from the first illuminant and then imagine what that surface would look like under the second illuminant.

The concept is illustrated in Fig. 5.10. In all configurations, a green square is placed in the middle of a white background. The matte surface of the green square will maximally reflect the middle wavelengths of the spectrum and the white background will reflect all the wavelengths. There are three illuminants: standard white (daylight), blue, and yellow. In all cases, the illuminant covers the square and background. Starting with the top configuration, the illuminant is made of equal amounts of energy at all wavelengths so that the background will appear white and the square (true) green. In the middle configuration, the illuminant is blue so that the background will appear bright blue and the green square will appear blue/green. In the bottom configuration, the illuminant is yellow, and the background yellow, and the green square will appear green/yellow.

Although we use a single color name to label the different surface reflectances and illuminants, neither are composed of a single wavelength. If that were the case, then the green square would look black when illuminated by the blue or yellow illuminant. Instead, the green square reflects a range of wavelengths spanning blue through yellow. Similarly, both illuminants also are made up of a range of wavelengths so that the green square can reflect the wavelengths common to the blue illuminant or common to the yellow illuminant.

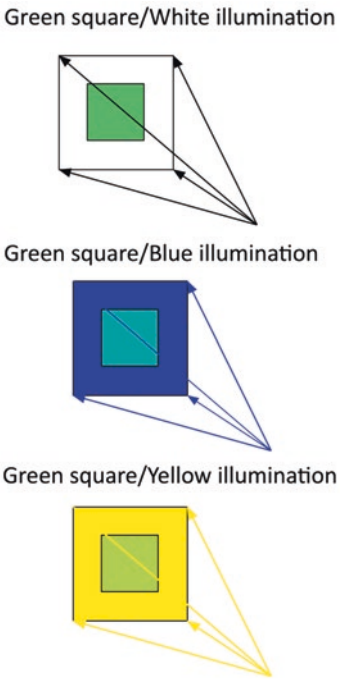


Fig. 5.10 Under a white illuminant, the green square appears green and the white background appears white. Under the blue and yellow illuminant, the background appears blue and yellow, respectively. The green square appears to be color between the green reflectance and the illuminant

In the experiment, a single square/single illuminant stimulus is paired with two squares/single illuminant and the observers have to choose which one of the two squares matches the single one in color. The first symmetric case tests whether the green square/white illuminant matches the green square/white illuminant or the green square/blue illuminant when both are presented on the white background. The choice here is a "no-brainer." Obviously, the match is between the two green squares under the white illumination. (This simple case is not included in the experiments). But the second case is more complex. Under the blue illuminant, the green square will not look green, but blue/green. To achieve color constancy, the observer must pair the green square under the daylight illumination to the blue/green square under the blue illuminant.

The second example in Fig. 5.11 follows the same logic. In the first case, not used in the actual experiment, the green square under the blue illuminant is compared to the green square under the blue or yellow illuminant against the blue background. Again, this is simple because the identical background makes the blue/green the match. But if the background is the yellow illuminant, then the correct match to the blue/green square is now the dull green/yellow square. To make the correct match, the observer must discount the effect of the blue illuminant and recognize that the center region is green, then realize that under a yellow illuminant that square would look dull green/yellow. Arrows show these responses.

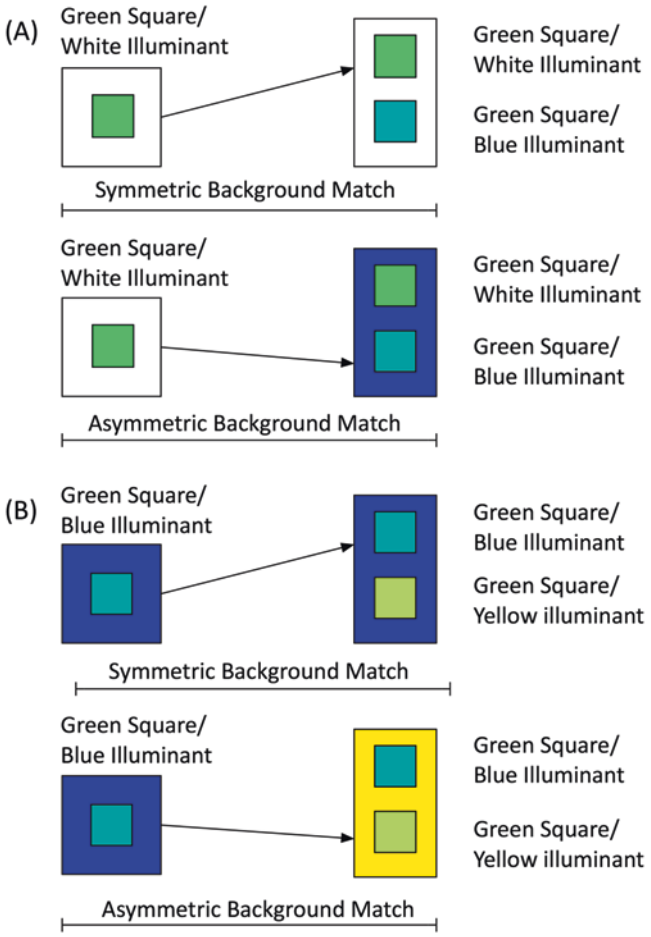


Fig. 5.11 For symmetric matches, the matching color in the test pair looks identical to the original. But for asymmetric matches, the matching color in the test pair does not look like the original, but the one that would occur if the green square were illuminated by a blue or yellow light. The hue of the illuminant can be determined from the background hue. Arrows indicate the correct matches

The procedures underlying asymmetric match experiments seem to maximize the difficulty of achieving color constancy. First, the changes in illumination, for example, blue versus yellow, are far greater than the naturally occurring differences between the morning bluish daylight and the evening yellowish daylight. Natural spectrums are continuous, and simply differ in tilt as shown in Fig. 5.4. Second, the experiments did not include specular or indirect reflectance. Third, observers were restricted to one view and unable to move about to obtain multiple views. Nonetheless, constancy was relatively high even though it is difficult to derive a statistical measure of it. If we imagine a set of

potential color matches between the green square/white illuminant and the green square/blue illuminant or green square/yellow illuminant, observers select a color match far closer to the color of the green square/blue illuminant or green square/yellow illuminant. Thus these results suggest a fair degree of constancy. However, there are large differences in the degree of constancy among observers, and observers with extensive experience may do better.

Radonjić, Gottaris, and Brainard (2015) attempted to make the color-matching task more realistic by increasing the complexity and contrast of the background. Increasing the complexity by placing the objects to be color-matched against checkerboard background did not improve color constancy, but making the background illumination inconsistent to the targets reduced constancy by 15 percent– to 20percent.

To summarize, many cues help us derive the object color. These include local contrast, global contrast, specular reflection, secondary reflections, shadows, and multiple views due to motion or changes in illumination. It seems reasonable that observers will switch their strategies depending on the situation and a cue worthless in one situation may be optimal in another. In all cases, however, the color belongs to the object and the observer's goal is to create a *coherent representation of the physical world*. If the objects seem to fall within one framework, then identical assumptions about the illumination on each object would seem appropriate. The coffee cup shown in Fig. 5.9 is assumed to be in one framework so that the differences in reflectance at different points on the surface will not be imagined to be due to changes across the surface itself or to different illuminants. If the objects seem to segregate into different frameworks such as being at different depths, then each object may have been illuminated differently.

One final point: We continually adapt to physiological changes. In general, the crystalline lens in our eyes becomes increasingly yellow as we age (primarily due to sunlight exposure), allowing less and less blue light to reach the retina. Nevertheless, we compensate for this yellowing so that white still seems to be white. After cataract surgery that removes the yellowish lens, the visual system, over the space of several months, recalibrates so that white returns to white.

5.3.4 "The Dress"

The picture of a dress originally published on Tumblr in 2015 went "viral" due the surprising split among viewers as to its colors. The photograph was taken on a cold winter day at dusk. In an informal Internet survey, 60 percent of the respondents reported it to be white and gold, 30 percent reported it to be blue and black, and 10 percent blue and gold (Wallisch, 2017). What the split implies is that what you see is not necessarily what others see and not necessarily what a color analysis with a photometer would measure. These outcomes support Monge's demonstrations illustrating that color perception is a second-order calculation based on the light reaching our eyes as well as our interpretation of the illumination of the scene.

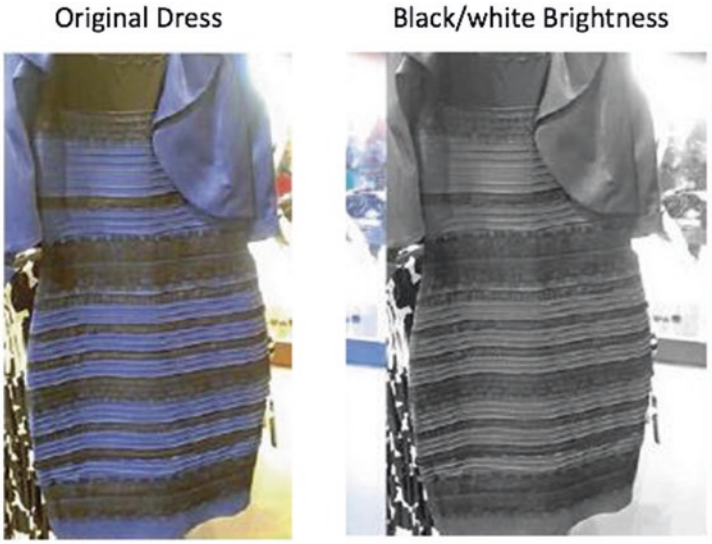


Fig. 5.12 The original colored dress and an achromatic white/black version. (Adapted and reproduced from Gegenfurtner, Bloj, & Toscari, 2015. By Permission, Elsevier)

The original photograph is show on the left in Fig. 5.12. To understand how the above split occurred, start with an achromatic version on the right. Following the logic used to explain Monge's demonstrations, this image could have come about in two ways. If the light source was in front, then the brightness of the image was probably due to a bright illuminant on a very dark dress. If the light source was in back so that the dress blocked most of the illumination, the brightness was probably due to a weak illuminant falling on light dress. Depending on the perceiver's guess about the position of the illuminant, overall the dress would look dark or light, respectively.

Using the same logic, the observer must deduce the "true" color of the dress by "neutralizing" the effect of the illumination. The apparent illumination seems to change from daylight at the top of the photo to shadows at the bottom. If one believed, on the one hand, that a "cool" blue light (i.e., morning sunlight) illuminated the dress, then any blueness in the photograph of the dress would be attributed to that light; neutralizing that illumination give rise to dress materials that appear to be white and gold. On the other hand, if one believed that "warm" yellowish light (i.e., afternoon sunlight) illuminated the dress, then neutralizing the yellowish illumination would give rise to dress materials that seem to be blue and black. (If the dress seemed to be illuminated artificially by a warm incandescent light, the dress would tend to look blue and black). These alternatives are illustrated in Fig. 5.13

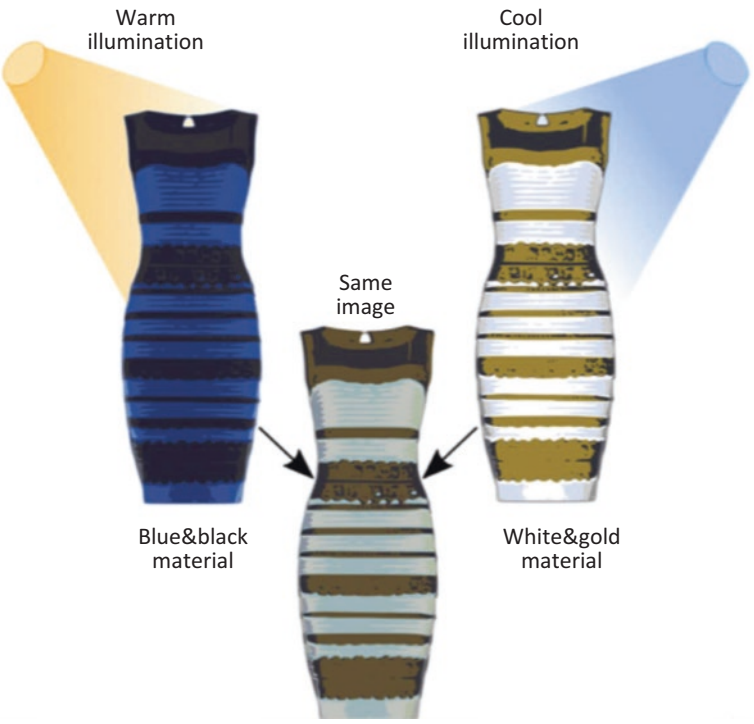


Fig. 5.13 The same image, in the center, could be due to blue/black dress illuminated by afternoon warm sunlight or a white/gold dress illuminated by cool morning sunlight. Different people, unconsciously compensating for their beliefs about the illumination will end up with a different percept. (Reproduced from Brainard & Hurlbert, 2015. By Permission, Elsevier)

To test this interpretation, several investigators have clipped copies of the dress and created scenes that seem to reflect different illuminations. For example, Witzel, Racey, and O'Reagan (2017) constructed one scene in which the light appears shadowed behind the model which overrepresents blue light and one in which the light appears to be shining directly on the model which overrepresents yellow light. When the light casts a bluish shadow, the dress looks lighter with gold stripes, but when the light is direct, dark blue with brown stripes. Furthermore, Wallisch (2017), based on an Internet survey, found that participants who believed that the dress was in shadows (i.e., bluish) were 20 percent to 40 percent more likely to describe the dress as white/gold than participants who believed the dress to be directly illuminated (75 percent to 45 percent).

The Bayesian explanation would be that the belief in one sort of illumination is based on previous experience, the "priors" in Bayesian terms. This has led to the hypothesis that "early risers," who more frequently experience bluish morning light, are more likely to see the dress as white/gold. In contrast, "late night birds" who more frequently experience the yellowish evening light are more likely to see the dress as blue/black. The evidence for this is weak, but I am certain that people have different priors that affect their perception of color of the dress (Wallisch, 2017).

In sum, the "dress" is another example that the light rays at the eyes, like sound waves at the ears, are inherently ambiguous. Our attempts to interpret them can lead to dramatically different beliefs about the distal objects. It is interesting to note that it is extremely difficult for people to shift from one color scheme to the other. This is not the case for multistable objects discussed in Chap. 3 where the shape shifts are inevitable and continuous. When multistable figures shift, the parts change their meaning. A curve that is part of a nose in one shape becomes part of an ear in another shape. But, for the "dress," it is always a dress and the stripes remain the same.

5.3.5 Does the Color of Objects Matter for Recognition?

Color, shape, motion, and other surface characteristics such as shading are the properties we use to break up the external scene into discrete objects. As described previously, infants are quite sensitive to color differences, and I suspect that stationary or moving colored blobs primarily segment the visual world for them. Adults seem to respond to shape more than color, but both affect the organization. Colored objects in the visual periphery aided localization, while colored objects in central vision aided object recognition (Nathmann & Malcolm, 2016)

The fundamental question is whether objects are first recognized by shape and then colored in, or whether shape and color are used conjointly to recognize objects. This question has led to experiments to investigate whether objects that are correctly colored are recognized faster than the same object if it is achromatic or even if colored incorrectly. For example, is a yellow banana more easily detected than a purple one?

The key distinction is the "diagnosticity" of the color for the object. A color is diagnostic for an object if we consistently identify it with that real-world object such as a banana; it is not diagnostic if the real world object could appear in any color, like a necktie. Obviously there are intermediate cases in which an object could occur in a limited set of colors. In many studies, diagnostic cues enhance object recognition, particularly when participants have to name the object (Bramao, Reis, Peterson, & Faisca, 2011). A gray seal, a yellow sun, a red stop sign, or a white snowflake is easier to recognize than a pink seal, a green sun, blue snowflake, or green stop sign. But, a gray car or sweatshirt is not easier to identify than a blue car or sweatshirt.

5.4 Countershading Camouflage

Another example of source-filter processes is countershading. Thayer, an American painter using decoys, was the first to illustrate how countershading could be used to avoid detection as dramatically shown in Fig. 5.14E. Our normal expectation is that sunlight comes from above (Fig. 5.14A). Hence, flat, horizontal, two-dimensional objects will reflect equal amounts of light at all points on their surface. Solid three-dimensional objects will self-shadow due to their shape. If the object is convex and bulges out, it will be brighter along the top and darker along the bottom (Fig. 5.14A) so that the perception becomes three-dimensional making the animal easier to detect. The brightness of a surface is the brightness of the incident light reaching the surface multiplied by the reflectance of the surface. On this basis, if the animal is colored so that its base is more reflective, that is, whiter, than its top surface (Fig. 5.14B), then the brightness of the animal's surface can be equalized (Fig. 5.14C). The three-dimensionality of the animal can be minimized by a gradation of shadow. "The animal now looks flat and Insubstantial" (Thayer, 1909). It is interesting that countershading is reversed for animals that hang from branches; the underside is darker. Rowland (2009) provides a review of the effectiveness of countershading.

5.5 Timbre

5.5.1 Source, Filter, and Resonance

The color section emphasized the problem of setting aside the effect of illumination to derive the "true" color of an object, the true color being what would be seen in the roughly uniform spectrum of daylight. This possibility was based on the assumption that object's reflectance spectrum is the same across all possible illuminations and does not change over time. If we gradually shifted the source illumination from blue to red, differences in the light reaching the eye would be due to the source, not changes in the reflectance of the object. If we could discount the variation caused by the source, we could picture the true object color.

This does not work for timbre because the resonances of the source and the resonances of the sound body filter occur at discrete frequencies. Suppose the sound body has resonances at 100 Hz, 400 Hz, and 700 Hz. As the source frequency is increased, it will induce the sound body to vibrate at those three frequencies with different movement patterns due to the material and shape of the object. (Between those frequencies the sound body vibrations will be very weak, if they exist at all). Moreover, the resonances of the filter change as a function of the way the source energy is applied. Vibrating a violin body at different position, intensities, and/or durations will change the frequencies of the resonances and thereby change the sound. Finally, all sounds occur in time so that the loudness and frequency change over time. Some source and sound

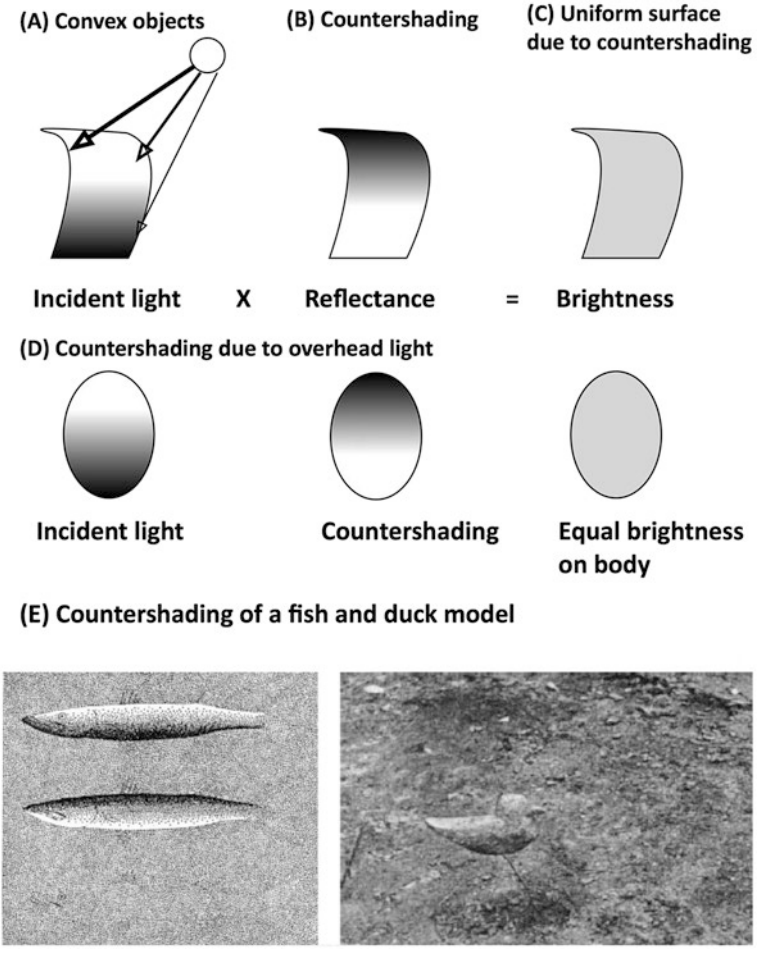


Fig. 5.14 (A) Sunlight reaching the top of a convex object will be stronger than that reaching the bottom (shown by the thickness of the arrows). (B) To compensate by countershading, the lower regions are more reflective than the top regions. Gradually changing the reflectance is most effective. (C) The energy of the incident light multiplied by the reflectance (i.e., the percentage of reflected light) yields the light reaching a predator. (D) Animals are often countershaded so that the top surface is darker than the underbelly. (E) An illustration by Cott (1940, page 37) and a photograph by Thayer (1909) demonstrate the effectiveness of countershading using a fish and duck model. For the fish, the overhead illumination (topmost drawing) is neutralized by the dark shading on the back of the fish (middle drawing). There really is a third fish at the bottom in the left panel and there really is a second duck model to the right in the right photo. (Thayer, 1909, Chapter 2, Page 24, Figure 4)

body resonances reach their maximum quickly and then decay rapidly while others reach their maximum and decay slowly. There is no "true" invariant timbre of an object.

For all of these reasons, descriptions of timbre are often ambiguous based on multiple properties of the frequency spectrum and temporal characteristics of the sound**.** When we hear a sound, we tend to identify the object and not its sound qualities. This contrasts with visual objects, which we often describe in terms of size, color, and shape. Nonetheless, it is those frequency and temporal characteristics that allow us to identify the object although there may not be a fixed set that work in real environments with overlapping sounds. A realistic goal may be to list a set of qualities and attempt to determine how they are used at different times.

To unravel timbre, we need to concentrate on the connections between the excitation source and filter that determine the evolution of the output spectrum over time. To do this we need to understand the coupling between the source and filter.

If we strike, vibrate, bend, twist, or blow across an object, the material may begin to vibrate at one or more of its resonance frequencies, termed vibration modes. Each vibration mode will have a distinct movement and can be characterized by the resonant frequency at which the motion is maximized and by its damping or quality factor. The material must possess enough stiffness or spring-like restoring property so that it snaps back and overshoots the initial position, for example, a violin or guitar string. The restoring force of the overshoot acts to reverse the original motion but the overshoot also overshoots and so on. The material begins to vibrate continuously in simple harmonic motion, like a violin string (see Fig. 5.3B for a picture of simple harmonic motion). To continue the resonant vibration, energy must be applied *in-phase* with the vibration. The motion of the excitation must match that of the material. If the excitation stops, the vibration of the material eventually dies out due to friction. It is important to note that the vibration frequency of the mode is that of the excitation even if that frequency is not the resonance frequency of the mode.

The damping or quality factor controls the range of frequencies that can excite the mode and the rate at which the vibration mode begins to vibrate and to die out. For a highly damped mode with a low-quality factor, a wide range of frequencies can excite the vibration mode though the amplitude of the vibration is small and roughly equal across that range. The amplitude builds up to its maximum quickly, it increases and decreases in amplitude in synchrony with the excitation, and once the excitation is removed it dies out quickly. Conversely, for a lightly damped mode with a high-quality factor only a narrow range of frequencies can excite the vibration mode but within that range the amplitude may be high. The vibration reaches its maximum slowly, lags behind changes in the source excitation, and dies away slowly when the excitation stops. For a highly damped mode, the excitation can reach two-thirds of its maximum in one cycle and lose two-thirds of its amplitude in one cycle. In contrast, for a lightly damped mode, it might take 10 cycles to reach the same values.

The common example of a damped system is a swing. If the "pusher" is tightly coupled to the swing by means of a stick connected to the seat the movement will be the same as the arm motion; the amplitude of the swing will be identical at all arm frequencies and the movement will begin and end in synchrony with the arm motion. It is highly damped. In contrast, if the swing is connected to the arm motion only by a weak spring, the swing can reach a great height if the excitation occurs at the resonance frequency by pushing at the same high point in the motion, in phase. But, the swing will reach its maximum only after an extended period of pushing and die out after an extended period of not pushing. It is lightly damped. The goal for a hi-fi speaker is a flat frequency response, which requires all frequencies would be reproduced accurately with no delay in responding to the excitation. This would require a highly damped system. The disadvantage is the output will be weak, but a high-powered amplifier can overcome that.

Based on these differences, we would expect that the vibration modes of a violin would be lightly damped in order to play loudly. The wood top and bottom plates would be thin and free to vibrate to yield the maximum volume. If there were just a small number of modes at varying frequencies, then those musical notes that match the frequency of the vibration modes would be full volume, but the other notes would be muted. Fortunately, the complex wooden structure has many modes with adjacent frequency peaks, which evens out the frequency response.

It is the interaction between the excitation, source, and filter that makes the analysis of timbre so difficult. Every source vibration is composed of multiple frequencies, and every sound body filter has multiple vibration modes with different degrees of damping. Due to the combination of the various levels of damping, the timbre of the sound will change from onset, to steady state to offset as different frequencies reach their maximum and decay at different times. Each excitation will excite different source and filter modes, so that we should not expect to find a single acoustic property that can characterize an instrument, voice, or environmental event. The perceptual problem is therefore identical to that for color, where the problem is to recognize the same color under different illuminations; for timbre the problem is to recognize the same auditory objects and events under different excitations, frequencies, and amplitudes. We need transformations that can band together colors, sounds, objects, or events in different contexts.

5.5.2 Timbre of Instruments

Gaver (1993) has proposed organizing auditory events into three physical actions: (1) vibrations due to scraping, hitting, or plucking, which would include percussion and stringed instruments, breaking and bouncing objects, and footfalls; (2) aerodynamic sounds due to continuous excitation, which would include blown instruments, wind noise, and mechanical objects like jet engines; (3) liquid sounds due to dripping, splashing, or boiling.

Across a wide variety of experiments two factors have emerged as critical to the perception of timbre and ultimately to the perception of objects and events. The first is temporal characteristics, either the duration of the onset or attack of individual sounds or the timing between individual sounds (e.g., rattles, faucet drips). The second is the energy distribution of the frequency components. There are other factors, but they tend to vary from experiment to experiment.

We start with an experiment that investigates the timbre of instruments (McAdams, Winsberg, Donnadieu, De Soete, & Krimphoff, 1995). These instruments cut across Gaver's categorization: stretched strings (e.g., violin, piano), stretched membranes (snare drums, tympani), rigid materials that vibrate without tension (cymbals, xylophone) as well as aerodynamic sounds (woodwinds and brass instruments). There were 12 electronic instruments that simulate real ones and six electronic instruments that were hybrids of the real instruments. The participants were asked to judge the similarity between each possible pair of instruments. The instructions were made purposively vague to avoid influencing the judgments.

A statistical technique termed *multidimensional scaling* was used to place the instruments in a geometric space such that the distances between the instruments in the space were proportional to their judged similarities. A simple example of this process would occur with three sounds such that the similarities between A&B, B&C, and A&C were 1,1,2. In this case we could place the three sounds on a one-dimensional straight line A-B-C. If the judgments instead were 3,3,3 then the three sounds could be placed as an equilateral triangle in a two-dimensional space. In a more complicated example, if there were four sounds and the similarity judgments between each pair were identical, then the four sounds would form a pyramid in a three-dimensional space. The difficulty, of course, is to figure out what acoustic property each dimension represents. Given each instrument's position in the two- or three-dimensional space, the researchers correlate the positions on each dimension to their acoustic properties. It is not enough to say that all the woodwind instruments are on one side and brass instruments on the other side of a dimension. The goal is to identify the acoustic variables that underlie that separation and that could result from the mechanical properties of the instruments.

The distances between the instruments are mainly due to differences on two dimensions, as illustrated in Fig. 5.15. The position of the instruments along the first dimension correlates with the attack or rise time of the sounds. Plucked or struck instruments at one end of the dimension were characterized by short rise times, while wind instruments at the other end tend to have slow rise times. Participants are not actually judging the rise time; rather they are judging the perceptual consequences. Struck or plucked instruments are likely to have an initial noisy impact sound made up of a wide band of frequencies. All the vibration modes of the sound body start at the same time. In contrast, blown instruments are likely to have a noisy initiation that evolves into a stable blowing vibration. The damped vibration modes reach their maximum at different times to create an evolving sound.

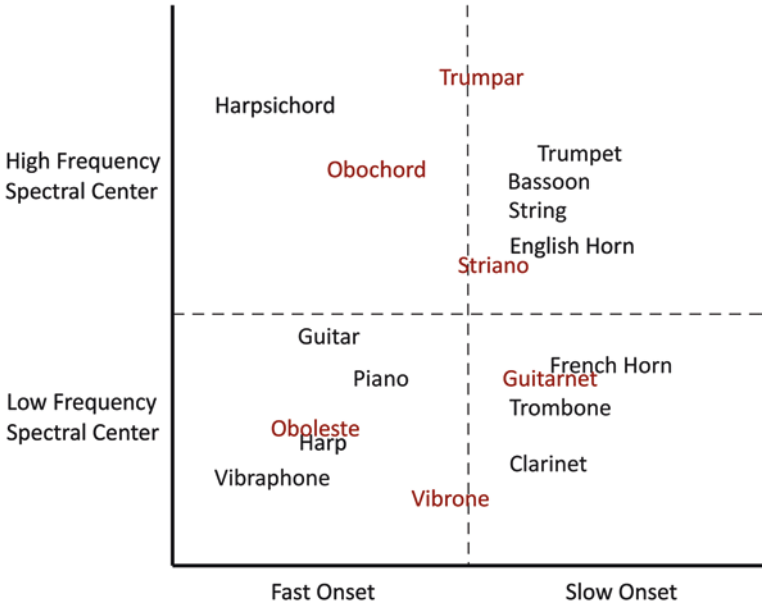


Fig. 5.15 The two-dimensional representation of the similarity judgments among pairs of instruments. The x-dimension represents differences in the onset speed of the sounds. The y-dimension represents differences in the spectral center of the sounds. Real instruments are black, hybrid instruments are red. (Adapted from McAdams et al., 1995)

Sound Files 5.15: Real and hybrid instrumental sounds shown in Fig. 5.15

The position of the sounds on the second dimension is correlated to the spectral centroid of the sounds, which is the frequency of the average energy. Suppose we had three harmonics at 100 Hz, 500 Hz, and 900 Hz; if the amplitude of the three harmonics were 2,3,1 then the centroid would be [2(100) + 3(500) + 1 (900)]/6 = 433 Hz and if the amplitudes were 1,1,4 then the centroid would be [1(100) + 1(500) + 4(900)]/6 = 700 Hz. Roughly, it is a measure of perceptual brightness or brilliance, the amount of energy at the higher frequencies.

In sum, the similarity judgments among musical instruments can be attributed to difference in the temporal and spectral characteristics of the radiated acoustic wave. Moreover, the spatial configurations tend to group the instruments into families based on their physical characteristics (Giordano & McAdams, 2010). Blown instruments are separated from impulsive ones.

The easiest way to illustrate that the timbre of a sound is caused from the joint action of timing and the spectrum is to reverse the sound and/or shape the amplitude of the harmonics, that is, change the spectral center. If we reverse the sound, the spectrum is identical, but the offset becomes the onset and vice versa. In Fig. 5.16A, the sound of a vibraphone, in reverse, sounds nothing like the original. We can also slow the onset, keeping the spectrum identical, and this also destroys the vibraphone sound.

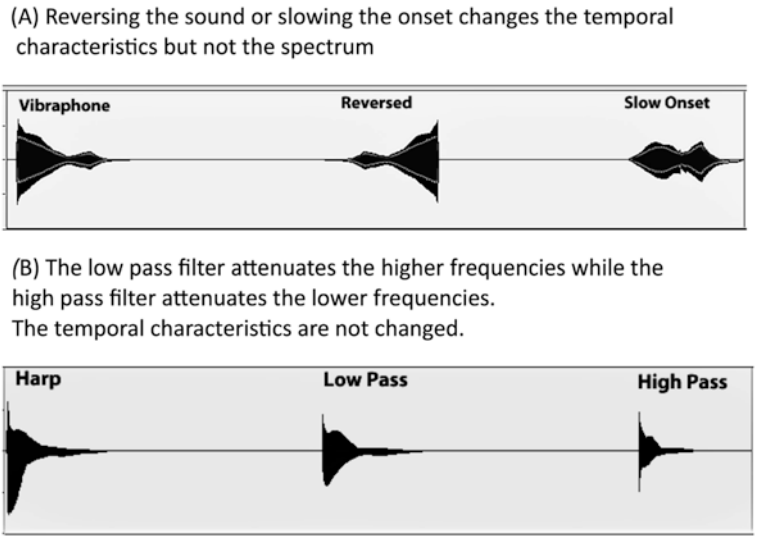


Fig. 5.16 (A) Waveforms of the original vibraphone sound, the reversed vibraphone, and the slow onset vibraphone. For all three sounds, the spectrum and spectral center are identical. (B) Waveforms of the original harp sound, the low-pass waveform, and high-pass waveform. The onset and duration are identical for all three sounds

Sound Files 5.16: Temporal and spectral changes that affect timbre illustrated in Fig. 5.16

We can also alter the spectrum (keeping the onset constant) shown in Fig 5.16B. For the harp, the amplitude peak occurs at the fundamental frequency of roughly 300 Hz with weaker peaks at the harmonics of 600 Hz, 900 Hz, and 1200 Hz as well as significant peaks at the harmonics out to 7000 Hz. The low-pass filter strongly reduces all peaks except for the 300 Hz fundamental and beyond 1000 Hz all disappear. Here the harp sounds hollow. The high-pass filter attenuates the amplitudes of the 300 Hz, 600 Hz, and 900 Hz peaks but does not affect the higher peaks. The harp now sounds tangy.

Nearly all the research of timbre is purely acoustical. Participants judge the similarity between two tones, judge individual attributes of the sound, or try to identify the instrument. What this misses is the feedback from the instrument to the musician and how that affects the perceived quality of the instrument. While playing, the actions of the musician create two sources of feedback. First is the acoustic feedback, the sound of the notes, due to the resonances of the instrument body. Second, but equally important, is the mechanical feedback due to the structure of the instrument that create vibrations on the fingers and hands holding the instrument or create contact forces if striking an instrument. Instruments have mass and are springy and their feedback to the roughly equal mass and springiness of the hand and arm determines how well musicians can adjust their playing technique to achieve an expressive acoustical goal. Playing is multisensory employing acoustical and haptic feedback (O'Modhrain & Gillespie, 2018).

Saitis, Järveläinen, and Fritz (2018) have reviewed research that makes clear that musicians judge the quality of violins based both on the acoustic qualities of the notes and the vibrotactile feedback. Three factors are important: (a) the strength of the felt vibrations by the left hand on the violin neck, by the shoulder, and jaw on the chin rest augment the perception of loudness and richness, conceptualized as *resonance* or *feel*; (b) the reactive force felt by speed and effort, felt through the bow by the right hand, conceptualized as *playability*; (c) the consistency of the vibratory feedback across the playing range, conceptualized as *balance*. The vibratory feedback both helps musicians control their playing style and expressiveness, but also contributes to their perception of the sound itself.

5.5.3 Timbre of Physical Actions

These results suggest that temporal and spectral characteristics of vibration sounds that are useful for analyzing the single sounds due to scraping, hitting, or plucking of percussion and stringed instruments could be extended to other kinds of acoustic events. Specifically, Gaver's analysis described above suggests that series of vibration sounds including objects breaking and bouncing, and footfalls could be understood in terms of the spectral characteristics of the individual sounds as well as the timing among those sounds, that is, the rhythm of the sounds. (Remember that the sounds of footsteps, even if they were not synchronous with the visual motion of the foot, influenced the perception of movement using lighted-dot figures in Chap. 2).

Hjoetkjaer and McAdams (2016) have investigated the perception of three types of materials, wood, metal, and glass, undergoing three types of actions, drop, strike, and rattle. If we isolate one type of action, say striking, then the difference between the materials would be limited to their resonant spectra. In similar fashion, if we isolate one type of material, say wood, then the differences between actions would be limited to the temporal pattern. The yields two interrelated questions: (1) when participants judge the similarity among the nine stimuli, are the two factors treated independently or are the stimuli perceived as a gestalt; (2) if all the materials are transformed to have identical spectra can participants identify the action, and conversely, if all materials undergo the same action can participants identify the material?

Figure 5.17 illustrates simplified temporal patterning for striking, dropping, and rattling. When these materials are struck, the initial sound burst decays due to its internal damping. If the material is dropped, the initial impact sound (that resembles the strike sound) is followed by two or three sounds due to bouncing. We can distinguish bouncing with its regular impacts, from breaking where multiple parts bounce independently and at random (Warren & Verbrugge, 1984). If the material is rattled, the sound is composed of sound bursts that vary in amplitude and timing.

Figure 5.17 also shows simplified spectra of the materials used. The spectra of metal and glass increase in amplitude from low to middle frequencies and then have somewhat random peaks at higher frequencies. In contrast, the spectrum of wood is flat at the lower frequencies and then gradually decreases at the

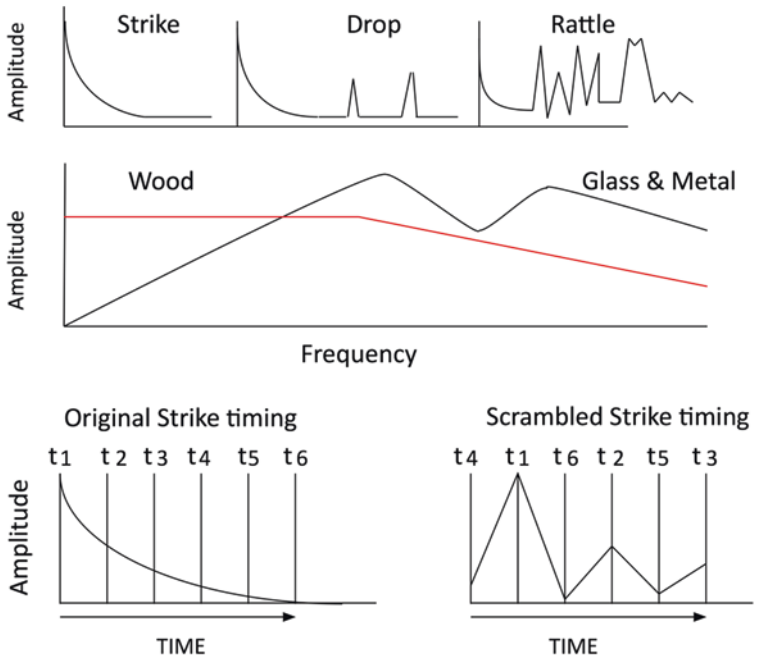


Fig. 5.17 Simplified representations of the temporal patterning of the three actions and the spectra of the three materials used by Hjoetkjaer and McAdams (2016). In the second experiment, the amplitudes of the strikes were scrambled across time to determine the effect of the temporal patterning, and a simplified version of the scrambled amplitudes is shown in the figure

Sound Files 5.17: The nine original sounds from three actions and three materials

higher frequencies. This suggests that while glass and metal are likely to be confused, they will be readily differentiated from wood.

In the initial experiment, participants judged the similarity among the nine types of stimuli. The results indicated that the similarity among the stimuli could be represented in two dimensions, as shown in Fig. 5.18. Wood was distinct from glass and metal probably due to the decline in high-frequency energy; striking tended to be distinct from dropping and rattling probably due to the absence of secondary sound bursts.

In a second experiment, Hjoetkjaer and McAdams (2016) attempted to assess the importance of the material and action separately by zeroing one out. To eliminate the effect of the spectrum, a constant broadband noise was presented using each of the three temporal patterns underlying the different actions. To eliminate the effect of the temporal pattern, the original pattern for each action was broken into time blocks that were then scrambled. The spectrum for each material was maintained. This process is illustrated in Fig. 5.17 for striking. Scrambling produces a pattern that resembles rattling.

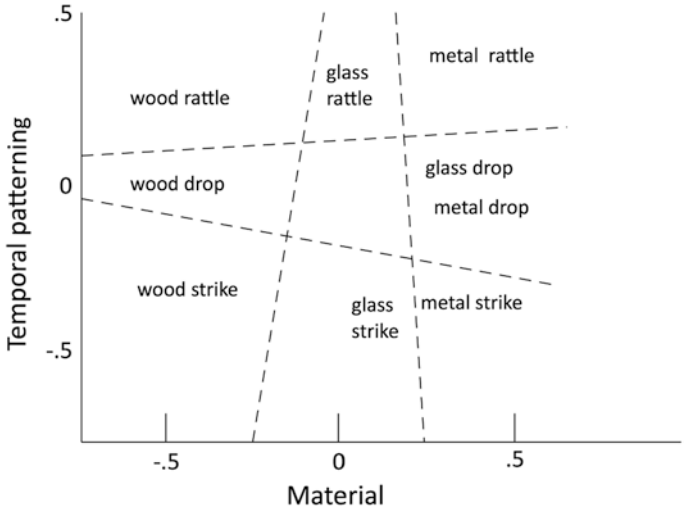


Fig. 5.18 The spatial representation of the similarity judgments among the nine sounds. These sounds roughly present the material and actions as independent factors supporting the results for instrumental sounds shown in Fig. 5.16. (Adapted from Hjoetkjaer & McAdams, 2016)

Sound Files 5.18: Original and modified wood drop and wood strike sounds

For the original sounds, the identification was nearly perfect for either the material (90 percent) or action (96 percent); any errors confused glass and metal. For the conditions that eliminated the spectral information, participants were unable to identify the material though roughly three-quarters of them judged the material to be metal, possibly due to the presence of a high-frequency energy "sizzled" sound**.** For the type of action, drops were nearly always identified correctly (88 percent), but strikes also were usually identified as drops (38 percent strikes to 60 percent drops) because the noise seems to eliminate the sharp onset created by the impact. Rattles were less affected, but somewhat likely to be judged as drops (68 percent rattles to 30 percent drops). In similar fashion, for the conditions that scrambled the temporal pattern, fully 96 percent of the responses identified the action as rattling, probably because the scrambled action had multiple onsets that resembled actual rattling. The percentage correct for material was very good (81 percent), although not quite equal to that for the original stimuli. As found for the original stimuli, wood was identified best, while metal and glass were confused about 25 percent of time.

5.5.4 Timbre of Environmental Sounds

Guyot et al. (2017) investigated Gaver's third category, the perception of liquid sounds. Water by itself does not make hardly any sounds; it is the popping of the air molecules entrained in the water that produces the sound. Smaller bubbles create higher-pitch sounds and aficionados think they can judge the

qualities of sparkling wines by the bubble sounds. The participants' task was to judge the sounds in terms of the physical event that caused the sound, not in terms of the actual sound itself. For example, water spray onto glass, water spray in a full tub, and splatty rain on a roof would all be classified as "jet" sounds even though they sound different. For our purposes, the most important finding was that the timing of the water sounds was a dominant feature. Short repetition sounds (i.e., drips) formed one category, while longer continuous sounds (boiling water) formed another category. The spectral characteristics did not distinguish among the sounds probably because the participants were instructed to judge on the basis of the physical causes of the sounds. As an aside, Katz (1925) has described "film" tactual perception of liquid or airflow as lacking an impression of an object, spatial orientation, or substance. The impression is simply that of immersion.

We can summarize these results in two ways. First, these outcomes illustrate how timbre is based on the evolution of spectral information over time. For struck instruments and events, the rapid, wide-frequency source impact onset causes the sound body to vibrate at its resonant frequencies, and these resonances die away at different rates. For bowed string and wind instruments, a noisy initiation period precedes a stable vibration, that is, after the bow fully engages the string. In turn, those source frequencies result in the vibrations of the sound body at the resonant frequencies. The onset and amplitude of the resonances are determined by the degree of coupling between the source and filter. For both kinds of instruments, the relative amplitudes of the sound body resonances yield the sense of brightness that has been correlated to the calculation of the spectral center.

Second, we can generalize the outcomes for single sounds to sequences. The spectrum and timing within each sound characterize the material and the timing between sounds yields the type of action. This is true for solids as well as liquids. Jumbling either the temporal or spectral properties kills the ability to identify those sounds.

Furthermore, the evolution of spectral information seems to be critical information for the identification for all sorts of sounds, although the relative importance of each possible cue depends on the sound and task. Ogg, Sleve, and Idsardi (2017) compared the categorical identification of musical instruments, speech, and human-environmental events (e.g., keys jangling, deforming newspaper) and found that listeners used temporal, spectral, noisiness, and spectral variability in various degrees to categorize this wide variety of sounds.

The descriptors for timbre often are cross-modal. Timbre is described in visual terms: sounds with higher spectral centers are bright, clear, active, or colorful while sounds with lower spectral centers are dark, dull, wooden, or colorless. But, timbre also is described in textural terms: rough or sharp versus smooth or delicate, warm versus cold, or heavy versus light. Thus, several experiments have compared the perception of the same object properties through visual, auditory, and tactual presentation individually and through multisensory presentation. Fujisaki, Soda, Motoyoshi, Komatsu, and Nishida (2014) paired the visual presentation of six materials (e.g., glass, ceramic, metal, stone, wood, and bark) with the sounds of eight materials (e.g., glass, ceramic, metal, stone,

wood, vegetable (pepper), plastic, and paper) being tapped with a wooden mallet. In some instances the visual and auditory presentations were congruent: visual presentation of a piece of glass with the sound of the glass being tapped. In other instances the presentations were incongruent: visual presentation of a piece of glass with the sound of a piece of wood being tapped. The results for the congruent presentations were straightforward; participants rated the correct material category highest. If the presentation was incongruent, for example, visual glass with auditory wood, participants compromised and chose the material that may have been a second or third choice for each modality, but was plausible for the individual visual glass and auditory wood presentations. Specifically, visual glass was sometimes categorized as plastic. Similarly, auditory wood was sometimes categorized as plastic. When presented together, the preferred description was therefore plastic. If participants were asked to judge the visual properties (e.g., uniform surface versus rough surface, or opaque versus transparent) or the auditory properties (e.g., dampened versus ringing sound or low pitch versus high pitch) for congruent and incongruent pairings they simply judged according to the relevant modality. If asked to judge other properties such as cold versus warm or hollow versus solid that were not modality specific, participants tended to average the judgments from each modality. The authors argue that these results can be understood in terms of Bayesian outcomes, in which the most reliable modality for the task is most heavily weighted.

The sensations from viewing wood samples, listening to the sound after tapping the samples, or running a finger along those surfaces is obviously different. To test whether the perceptions are equivalent and independent of the modality, Fujisaki, Tokita, and Kariya (2015) asked participants to judge the material and affective properties of the wood samples after viewing, listening, or rubbing them. The first dimension for each modality reflected its intrinsic properties: light/heavy, sparse/dense, and fragile/study for visual presentation; dull sound/sharp sound, mixed sound/pure sound, and damped sound/ringing sound for auditory presentation; rough/smooth, matte surface/gloss surface, and dull sound/sharp sound for touch. The second and third factors reflected the evaluative properties and these were similar for each modality. The second factor grouped terms like expensiveness, rareness, sophistication, and interestingness while the third factor grouped terms like relaxed feeling, pleasantness, and liked. Thus the ratings on the affective or emotional terms were the same across the three kinds of modalities in spite of the large differences in sensations.

5.6 Timbre Constancy

We have argued that timbre adheres to an object much like color. The fundamental problem for color perception is to discount any changes in illumination and recover the reflectance to discover what the color would be under daylight. But the problem for timbre perception is not comparable because the timing and amplitude of the resonances changes at different frequencies. There is nothing that resembles a stable reflectance.

5.6.1 *Independence of Spectral Center and Frequency*

Yet, it is possible to imagine questions and experiments that resemble those for asymmetric color matching. First, the spectral center is based on the relative strength of the harmonics of the fundamental frequency. The spectral center can be raised by increasing the fundamental frequency or by increasing the amplitudes of the higher harmonics. Can participants distinguish the two? If the spectral center is increased, do listeners mistake that for an increase in pitch? Second, the timbre of instruments and singers change at different notes. Can experienced and inexperienced listeners recognize if instruments or singers are identical at different notes?

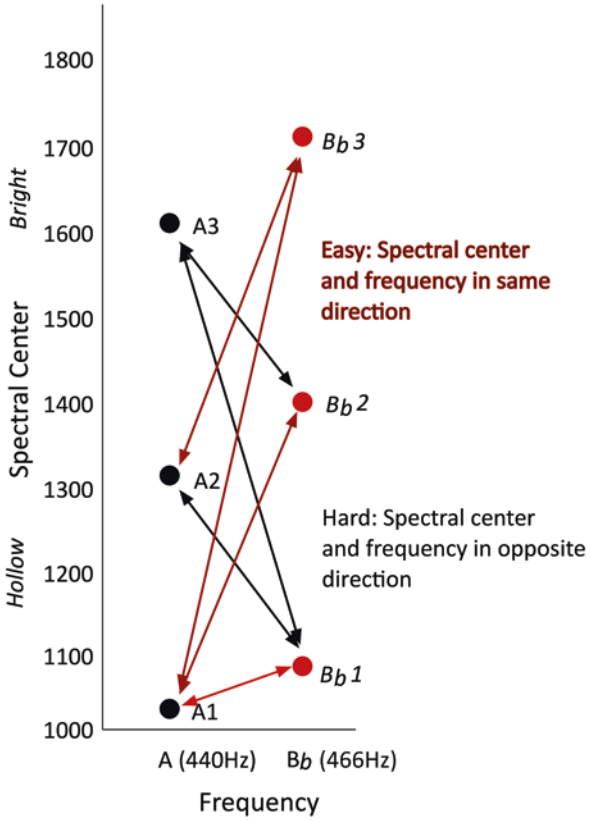


Fig. 5.19 If both the spectral center and the frequency increased for one tone, it was easy to judge which tone had the higher pitch. But, if one tone had the higher spectral center, but the other tone had the higher frequency, it was difficult to judge which tone had the higher pitch. The first five harmonics were used to calculate the spectral center: 440, 880, 1320, 1760, and 2200 Hz for A, 466, 932, 1398, 1864, and 2330 Hz for B_b. The relative amplitudes for spectral center 1 were 0.5, 0.4, 0.3, 0.2, and 0.1; the relative amplitudes for 2 were 0.3, 0.3, 0.3, 0.3, and 0.3; the relative amplitudes for 3 were 0.1, 0.2, 0.3, 0.4, and 0.5

Sound Files 5.19: Comparisons among sounds that vary in fundamental frequency and spectral center

The basic experiment to determine if individuals can judge timbre independently of frequency can be visualized in Fig. 5.19. There would be six sounds: two adjacent frequencies and three levels of timbre based on the spectral center. Differences in the spectrum bring about different perceptions of brightness that may be confused with differences in pitch. On each trial, the participant would be presented two of the sounds and identify the one with the higher pitch. This would be easy when the spectral center is similar, A1-B_b1, A2-B_b2, and A3-B_b3, or when the spectral center and frequency both increase, shown in red. The difficulties arise when the spectrum and frequency differences are in the opposite direction. For example, consider the comparison between A3 and B_b1. Although the actual frequency of A3 is lower than B_b1, if timbre and frequency are not independent the higher spectral center of B_b1 might lead participants to judge A3 as being the higher pitch. In similar fashion, confusions might occur between A2 and B_b1 and between A3 and B_b2.

Experiments (e.g., Allen & Oxenham, 2014) have used a similar procedure. The results indicated that timbre differences based on changes in the spectral center are often confused with independent changes in the fundamental frequency, particularly for the incongruent conditions in which increases in spectral brightness conflicted with decreases in frequency. The greatest number of confusions occurred between similar sounds; A2 versus B_b1 and A3 versus B_b2 were more likely to be confused than A3 versus B_b1. Probably the latter two sounds were so different that confusions did not occur. Surprisingly, performance was identical for musicians and non-musicians. What this means is that timbre and pitch can be judged separately, but that changes in either the height of the spectral center or frequency can be confused with each other. It seems unlikely that changes in timbre due to variation in the onset time would be confused with changes in frequency.

5.6.2 Timbre of Sources at Different Frequencies

5.6.2.1 Instruments

Figure 5.5 illustrates that the sound body filters of most objects contain multiple resonance modes. Hence, we might expect that the sound quality of objects, that is, timbre, will vary as the source frequencies change. A compelling demonstration of the change in timbre across frequencies is found in the Acoustical Society of America Auditory Demonstrations CD. On this track, first you will hear the actual notes of a bassoon across three octaves. It does sound like a bassoon throughout. Then the spectrum of the highest note was determined, and all the other notes were synthesized based on this spectrum. Suppose the relative amplitudes of the first four harmonics of the highest note were 2, 4, 1, and 3. For a 800 Hz tone, the amplitudes of the first harmonic would be 2(800 Hz), the second harmonic would be 4(1600 Hz), the third harmonic would be 1(2400 Hz), and the fourth harmonic would be 3(3200 Hz). For a 1200 Hz tone, the first four harmonics would have the same relative amplitudes: 2(1200 Hz), 4(2400 Hz), 1(3600 Hz), and 3(4800 Hz). The spectrum, that is, the relative amplitude of the harmonics, remains constant throughout the playing range. The synthesized notes should sound like the actual notes of the

bassoon, but that obviously does not occur. At the different scale notes, different resonances of the bassoon body emerge, changing the timbre.

Sound File 5.20: The original bassoon notes and the simulated notes created by keeping the spectrum constant. By permission, Acoustical Society of America

Given this variability in the sound reaching the listener, it is no surprise that the ability to recognize instruments, objects, and events is relatively poor. While it is possible to imagine that colors possess a reference under daylight, there are no such references for timbre. In addition to the inherent acoustic variability, there are environmental differences (enclosed rooms versus outdoors), memory limitations, individual differences, and expectations that also act to depress recognition.

The simplest experiments present two wind instrumental sounds at different pitches (Handel & Erickson, 2001). In some trials, the instruments are identical, (e.g., Clarinet G_3/Clarinet G_4) and in others the instruments and pitches differ (Clarinet G_3/Trombone G_4). The participant's task was simply to judge whether the two sounds were from the same instrument or not. For non-musicians, this is a relatively difficult judgment, and if the pitch difference was an octave or more, all pairs were judged as being different instruments. At intervals less than one octave, participants tended to judge two instruments within the same type (French horn, clarinet, and trombone or English horn and trumpet) as identical. Sound examples are found in Sound Files 5.21A–D.

Sound Files 5.21: Comparisons of instrumental notes that differ by less than one octave and by more than one octave

These outcomes should be treated with caution, however. All were wind instruments, and we certainly would not expect confusion between a wind and a stringed instrument. Moreover, the participants were inexperienced and musically trained participants would be expected to do better. Steele and Williams (2006) found that experienced musicians were able to correctly discriminate between a bassoon and French horn at separations of two octaves, while non-musicians were unable to make that discrimination beyond one octave. It is interesting to note that while musicians and non-musicians make exactly the same similarity judgments between instruments, musicians are better at timbre and pitch discrimination.

The argument made here is that is that because timbre from one source changes from note to note due to the different resonances, one needs to create a transformation that would connect the sounds. From this perspective, only two sounds are presented in the above experiments and that would preclude the formation of such a transformation. Musicians, having experienced many more notes from instruments are more likely to have created such internal transformations that enable them to extrapolate and interpolate what other notes would sound like.

5.6.2.2 The Oddball Task: Instruments and Singer's Voices

To investigate whether a richer set of notes can improve the ability to recognize instruments and singers at different pitches, Erickson and co-workers made use of an oddball procedure. Three or six notes were presented sequentially, the

same instrument or singer presenting all but one of the notes. The participant's task was to identify the odd note. Compared to the experiments that used but two notes, the three-note sequence should provide a slightly stronger sense of the transformations across pitch, while the six-note sequence should generate a much stronger sense of the change in timbre across the notes. This should lead to the paradoxical outcome that the six-note sequence would produce better performance, although it is a more complex task.

To provide a baseline, the three-note sequences were presented using one instrument or voice. In all but one condition more than 80 percent of the responses chose the most dissimilar note as the oddball. If there actually was an oddball instrument or voice, the results were more complex. If the two instruments were both woodwinds, then the same errors occurred; participants nearly always judged the most dissimilar pitch as the oddball. However, if one instrument was a woodwind and the other a brass, then the ability to identify the oddball when it was not the most dissimilar pitch increased dramatically. In sum, if the two instruments were the same type, then participants were unable to differentiate their extrapolations and pitch dominated the judgments. If the

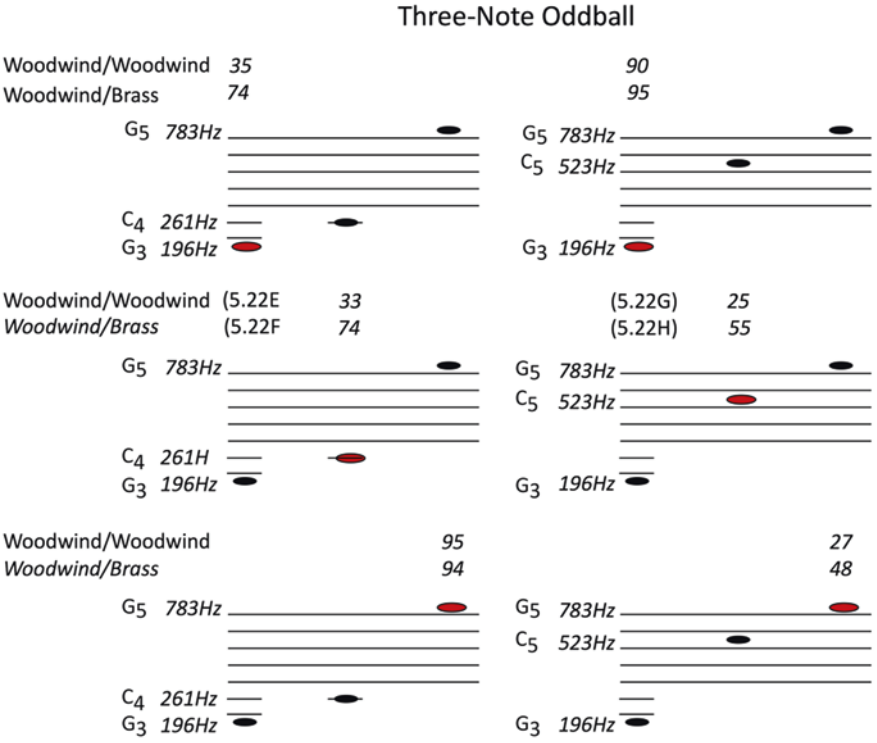


Fig. 5.22 The percentages correct from the three-note oddball task. The oddball note played by the second instrument is portrayed in red. The oddball note was usually picked as being the extreme pitch particularly if both instruments were in the same class. If an instrument in another class played the oddball note identification improved

Sound Files 5.22: The oddball comparisons shown in Fig. 5.22E–5.22H (Middle row)

instruments were from different types, participants were able to perceive the differences in timbre even across dissimilar pitches. These outcomes can be seen in Fig. 5.22.

The results from the three-note oddball task are a mixed bag. There is some evidence that the third note aided identification, but the performance within the same class of instruments just slightly improved. To investigate whether a richer set of six notes would truly increase the identification of the oddball, Erickson and Perry (2003) constructed three-note and six-note sequences sung by sopranos and mezzo-sopranos. For soprano/soprano sequences, one soprano sang two of the three or five or the six notes and the other soprano sang the remaining note. The same procedure was used for the mezzo-soprano/mezzo-soprano and soprano/mezzo-soprano sequences.

The identification results for the three-note oddball experiments mirrored those for instruments. Participants were unable to identify the oddball singer. Surprisingly, the performance for the six-note sequences was better absolutely, even though chance performance was 16 percent versus 33 percent for the three-note sequences. What is more important was that the errors were scattered through the six notes instead of bunching at the lowest and highest pitches. The errors reflected idiosyncratic variation in the singers, further supporting the conclusion that participants were constructing a transformation that enabled them to link the different sung notes. These transformations created a frame of reference that allowed the idiosyncratic variation to be isolated (and possibly misjudged as the oddball).

All of this occurred in a controlled laboratory setting, but in our daily lives nearly every sound could have come from many different sources. For example, a *click* could be due to a ballpoint pen, light switch, computer keyboard, or a simulated camera snap on a smart phone. Ballas (1993) presented everyday sounds and asked participants to identify them. The responses to some sounds were almost unanimous and were easy to identify in later experiments. The responses to other sounds were quite diverse, and more difficult to identify.

What this all means is that there is no single cue for identifying the source of sounds. Ballas (1993) concludes that there are many domains that contribute to identification. First, there are temporal properties ranging from the timing of multiple impacts to the length of the onset transient. Second, there are spectral properties such as the spectral center arising from the resonances of the source. Third, there are small variations in the resonances that yield the sense of change. Fourth, there is the familiarity with sounds (e.g., musicians do better than non-musicians) that arises from the frequency of occurrence in the environment. Finally, we should not forget expertise that arises from extensive study of types of sounds. Bird-watchers, railroad buffs, or car enthusiasts can make fine distinctions that novices cannot. (This may be identical to chess experts being able to detect small differences among the placement of pieces).

Perhaps there is “no smoking gun” for any sort of perception. There is always the inverse problem for color and timbre and doubly so for the objects in the world.

5.7 Echolocation

Echolocation is the process of emitting sounds in order to determine from these reflections and echoes the location and material properties of objects. Both sighted and blind individuals make use of the information gained through echolocation. Builders and carpenters tap walls to determine the location of supporting studs; boat builders and physicians tap enclosed spaces to determine if the spaces are hollow or filled with fluid. Blind individuals snap fingers, make clicking sounds with tongues, stomp feet, and tap canes to locate vertical material surfaces. The source is the produced sound, for example, the rapping against the wall or the tongue click, and the filter is the reflecting surface or hollow that changes the frequency spectrum from the source to the echo. As detailed below, accurate perception depends on the comparison of the source and the echo. The reflection by itself is usually insufficient. Moreover, as found previously, all perceiving is based on the combination of cues evolving over time.

The current interest can be traced to the discovery by Donald Griffin in 1938 of the ultrasonic calls of bats used to navigate and to hunt insects. The comparable use of ultrasonic calls by dolphins and other toothed whales was discovered later. It is not surprising that the ultrasonic calls of bats and dolphins differ in such dissimilar surrounds. In air, the speed of sound is slow and the higher frequencies rapidly dissipate. Narrow frequency sounds can be sustained for the best detection of prey while the frequency of repeating short broadband sounds best for localization can be increased as the bat approaches prey without creating confusion between the outgoing sound and the echo. Bats have time-delay neurons that respond best to signals separated by a fixed delay. (I have often wondered how bats when moving in a large group avoid being confused by the calls of other bats. It seems that the bats vary the frequency of their calls to reduce acoustic interference. This seems analogous to blind individuals moving their head direction when echolocating and individuals changing their speaking voice in crowded noisy rooms). In water, the speed of sound is roughly five times faster than in air so that the calls of dolphins tend to be shorter wide broadband signals at slower rates. In contrast to bats, dolphins reduce the amplitude and rate of clicks as they approach prey. Otherwise, the signals and echoes would overlap and interfere.

The ability of humans to echolocate has been known for hundreds of years. But the fact that it is based on acoustic information has been discovered only in the last 60–70 years because the blind could not identify the cues they used. Many attributed their ability to heightened facial or tactile sensations such as shadows or pressures across the eyes. Studies begun in the 1940s showed that auditory information was primary. Blind and deaf people were unable to judge when to stop walking in front of a wall, but blind and hearing individuals could. If blind people wore hearing protectors or were placed in noisy rooms, they increased the intensity of their foot shuffling or made other noises to compensate. Facial masks made from cotton did not affect localization, undercutting the facial vision theory (Ammons, Worchel, & Dallenbach, 1953).

The vocal sounds humans use for echolocation are usually short wide band clicks produced by tongues or snapping fingers, with maximum energy in the range between 2000–4000 Hz with higher frequencies from 6000 Hz to 10,000 Hz. The duration ranges from 3 ms to 8 ms and the clicks repeat from 1.5 to 2 times per second. The outgoing click is beam-like; the amplitude is symmetrical and relatively constant out to 60^0 both horizontally and vertically (Thaler et al., 2017). Given the beam characteristics, it would be difficult for a walking person to use echolocation to investigate ground level objects, and that accounts for the use of canes. Although a hissing continuous noise would make an effective source, it has not been employed in the experimental tasks. Based on these sounds, we can identify potential auditory cues that could be used to determine the surface, location, and orientation of objects (Kolarik, Cirstea, Pardhan, & Moore, 2014).

5.7.1 Acoustic Cues

A. The most obvious cue for distance is the loudness of the returning echo. In general, loudness decreases as a function of the distance squared ($1/d^2$) between the source and the listener. The sound moves out to the distant object and back again as the echo, so that as the distance doubles from the source, the distance the sound travels increases fourfold and the level drops to 1/16th (if there is no absorption at the target).

But loudness of the returning echo is inherently ambiguous. If we were simply trying to estimate the distance of an external source, any given loudness would be due to the inherent loudness of the source and the size, distance, or orientation of the source. A low-intensity sound could have come from a nearby soft source or a loud distant one. Echoes too are ambiguous. A soft returning echo could come from a distant object, but also could be due to other properties of the object; a smaller as opposed to a larger reflecting surface will return a smaller percentage of the outgoing source energy; an offset or angled surface also will return just a fraction of the energy because much of the echo is reflected away from the source; or a soft textured surface might absorb much of the incoming energy so that little of the energy is returned. To untangle the ambiguity, the listener could make use of prior knowledge about the size, orientation, and surface of the objects in the environment, or some of the other available cues discussed below.

B. The second obvious cue is the time delay between the outgoing click and the returning echo. The speed of sound is roughly 1100 ft./sec or 340 meters/sec. For an object 1 ft. away, the echo delay would be about 2 msec (2/1100) and for an object 3 ft. away the echo delay would be about 6 msec.

The difficulty in making use of the echo delay is due to the *precedence or Haas* effect. In general, any external sound is likely to reach the listener multiple

times after being reflected by the surfaces in our cluttered environment. The paths will be different lengths bombarding the listener with the more or less identical sounds, one after the other. If we analyzed each sound separately, each one would seem to have come from a different direction. But we do not hear all of these echoes; instead we hear the sound as coming from the direction of the first arriving sound. The later arriving echoes reinforce the direct sound and lead to a sense of spaciousness, but they do not affect the perceived location. A neat example occurs if several hi-fi speakers are arrayed in a line in front of a listener. As the listener walks back and forth the sound seems to travel along because the sound from the speaker in front has the shortest path and arrives first.

The precedence effect has two consequences. First, the emitted click, perceived acoustically or through bone conduction, can merge with and dominate the returning echo suppressing it so that both the emitted click and echo seem to be in front of the face. Second, more distant surfaces can be hidden. If we have two surfaces 10° either side of the midline, the echo from the nearer surface will reach the listener first and the echo from the further surface is perceived to be the same direction as the first echo. Another consequence of the precedence effect is to minimize the effectiveness of time differences between echoes reaching each ear as described below. The act of vocalizing the outgoing source may act to minimize the suppression of the echo due to the precedence effect (Wallmeier, GeBele, & Wiegrebe, 2013).

C. The third cue is the ratio of direct to reflected sound. A sound in an enclosed environment will reach the listener from many directions. The direct sound from the object is the loudest and arrives first so that due to the precedence effect it defines the direction. Following that are sounds that have bounced off the walls and ceiling once or multiple times along many paths of differing lengths so that they reach the listener from many angles with varying delays. The reflected waves are weaker due simply to air friction and at each reflection, the sound loses energy; fuzzy materials absorb more energy, hard non-porous material absorb less. Still, overall, the energy of the reflected waves is roughly 10 times that of the direct wave. But, if the energy of direct wave is quickly swamped by the reflected waves, the direction and clarity of the source is lost. Nonetheless, while the reflected waves reduce the perception of direction, it seems to have little effect on the perception of distance (Wallmeier & Wiegrebe, 2014)

If the sound continues until the level is uniform and then stopped, the reverberation time is defined as the duration of the sound until it is reduced to one millionth of the original level. The energy lost in any given time interval is proportional to the remaining energy. In general, the reverberation time for smaller rooms is shorter because it will have more reflections in a given time interval and therefore lose energy more quickly.

The intensity of the reflected echo decreases with increasing distance according to $1/d^2$ but the intensity of the reflected sound off the surround decreases only slowly as the room size increases. Thus, the ratio of direct to reverberant energy is smaller for larger rooms, which becomes an important cue for judging the source distance. However, for ongoing sounds it is unlikely that listeners can distinguish between the direct and reflected sound. They probably make use of the amount of reverberant energy, the correspondence between the reverberant energy in each ear, and/or the overall level. The best strategy of course is to make use of the optimal cue for each listening situation (Kopco & Shinn-Cunningham, 2011).

Sound Files 5.23: Simulated Reverberations in differently sized spaces

D. Proximity resonance. As a sound wave moves toward and bounces off a surface the energy at that surface increases due to constructive addition, where the amplitudes of the incoming wave and reflected wave combine so that the amplitude at the wall is twice the initial amplitude. Moving away from the wall, the amplitude oscillates; at some points higher-pressure regions overlap, but at other points the higher-pressure regions overlap the lower pressure regions producing "destructive addition." The pressure variation decreases further away from the wall because the returning wave has lost energy.

The important point for echolocation is that the change in amplitude due to reflecting wave is a function of wavelength. Simply put, the buildup in amplitude near the wall is greater for low frequencies. For a tone of 110 Hz, the amplitude starts to increase 15″ (0.38 m) from the wall (wavelength/8), but for an 1100 Hz tone the increase begins just 1.5″ (0.038 m) from the wall. Thus it seems likely that blind individuals would attend to the low-frequency variation.

Ashmead and Wall (1999) investigated the ability of sighted people to detect the distance to a frontal wall based on the spectrum of the reflected sound wave. To do so, the authors simulated the spectrum of sound at varying distances from a wall. Sounds that represented those close to a wall had more low-frequency energy than those that represented sounds further out. Overall, participants were able to detect a wall at a distance of about 1.5′ (0.5 m).

Surprisingly, simulating movement toward the wall did not improve detection beyond simply presenting the spectrum of the closest point. The 1.5′ (0.5 m) distance matches measurements when blind children walk along a hallway. If the children walked along a narrow hallway less than 6′ (about 2 m) wide, they were able to maintain a smooth trajectory. If they veered slightly, they would be able to make use of the spectral change perceivable near the opposite wall and correct their course.

E. Spectrum of returning echo. Due to friction, high frequencies decay more rapidly than low frequencies. Listeners judge the distance of sounds with high frequency as being closer than sounds without high frequency. Furthermore, the material of the reflecting surface will affect the spectrum of the returning echo. Softer materials such as carpet absorb a higher per-

centage of higher-frequency energy than harder materials such as metal or rigid plastic so that comparing the spectrum of the source to that of the echo can be used to make a somewhat rough guess as to the surface.

Experts can discriminate among shapes, such as a square versus a triangle. For two- and three-dimensional objects, the reflected echoes may arrive at different times and therefore interact with each other creating spectral changes. There may also be specific spatial reflections at the edges of objects that change the spectrum of the returning echoes. Although the cues used to make these object discriminations are not known, it is likely that spectral changes and loudness differences contribute.

F. Prior knowledge. Experience with events and objects can yield cues to distance. We have heard thunder, ambulances, and fire engines at close hand, have learned that whispering does not include voicing, and that shouts are characterized by high-frequency energy. That knowledge allows us to disregard the overall level in judging distance and allows us to make estimates based on the level and spectrum of those sounds.
G. Repetition pitch. Experts often comment that changes in pitch are important cues for estimating distance. As suggested above, this may be due to enhancement of different frequencies due to proximity resonance. But a sense of pitch can arise as a consequence of the overlap of the source and echo or the delay between the direct echo and the later echoes reflected off a wall.

Imagine standing in the middle of a room with a sound source directed at the front wall. The direct echo will arrive first followed by reflections off the floor, the sidewalls, or the back wall. Suppose the sound bouncing off the floor is delayed by 10 msec, that is, the reflected sound has travelled an additional 11′. Assume that the direct echo and the floor bounce echo are identical. What one hears is the combination of the two sounds, the direct echo plus the floor echo delayed by 10 msec away (Fig. 5.24). The same outcome would occur near the sidewall of a concert hall. You would hear the sum of the direct sound plus the sound delayed by 10 msec after bouncing off the sidewall.

The correlation between adjacent elements in the direct echo and floor echo is zero because they are a sequence of random numbers. But, the correlation between adjacent elements in the sum has increased to some degree because adjacent values share the same amplitude. For example, 8+3 and 4+3 share 3, 8+3 and 8+6 share 8, and so on. Part of the sum is identical between pairs. This correlation yields the faint pitch based on the 10 msec delay or 100 Hz. If the delay increases to 20 msec (the reflected echo traveled an addition 22′), the perceived pitch would become 50 Hz, or if the decay decreased to 5 msec (the reflected echo traveled an addition 5.5′) the perceived pitch would increase to 200 Hz. On this basis, the pitch from the repetition delay could bring about the perception of distance. As the pitch increases, surfaces would appear closer and conversely as the pitch decreases surfaces would appear further.

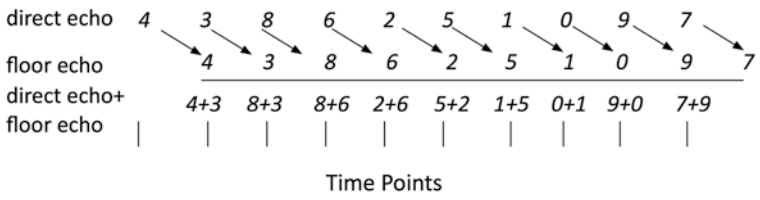


Fig. 5.24 The amplitudes, represented by the random numbers, of the direct and floor bounce echo are identical. The sense of pitch caused by the delay+add of the two echoes is created by the partial correspondence between the amplitudes in adjacent points, for example, (4+3) & (8+3) and (8+3) & (8+6). Notice that there is no correspondence in amplitude between time points two steps apart

Sound Files 5.24: Examples of repetition pitch at different delays and number of echoes

While repetition pitch may seem an unlikely phenomenon, it occurs in nearly all environments. Even in an open field, for example, the delayed sound bouncing off the ground leads to repetition pitch. In concert halls, seats near sidewalls can be susceptible to repetition pitch from those delayed sound waves. It is better to sit in the middle. (In general though, the reflections occur at different delays that would eliminate the repetition pitch). If a sound source is in front of a staircase, each step can yield one reflection, and equally spaced steps will result in a series of equally spaced echoes that result in a strong sense of pitch. The only sure way of avoiding the effects of repetition noise is to float in outer space. But, then again sound pressure will not travel in a vacuum.

H. Binaural cues. Due to the circular head and outer ear (i.e., pinna), lateral sources give rise to interaural time and interaural level differences. An informal observation shows the advantage of binaural hearing. If one approaches a wall sideways so the sound in each ear differs, it is far easier to detect the wall than if one approaches the wall straight on.

The maximum interaural time and level difference would occur when the source is directly opposite one ear. Even at +/−90°, the time difference is only 1 msec but does not change as the distance from the source varies. Even though we can easily make discriminations at this short time interval, we might expect that in normal environments this cue could not be used to determine distance due to the precedence effect. The interaural level difference changes as a function of the frequency of the source and its distance. Low frequencies at the near side are enhanced at close distances due to proximity resonance (i.e., sound waves bouncing off the head). But at greater distances the low frequencies bend around the head so that there is little difference in intensity between the near and far ears. (Sounds will bend around the head when the wavelength of the head matches the diameter of the head, roughly 1500 Hz). For higher frequencies, the head creates a shadow such that the sound at the closer ear is more intense than the farther one that does not vary with the distance of the source. Thus, the relative intensities of the low and high frequencies can provide a cue to distance.

Multiple studies provide strong evidence that expert echo locators make very accurate judgments about distance, location, and surface qualities and that sighted people can learn to make the same discriminations after a relatively short training period (e.g., Donelli, Brayda, & Gori, 2016). (It is true that the performance of such sighted individuals does not match those of experienced echolocators, but this is probably due to the short training). It is frustrating given the strong performance and the wide variety of cues available for distance, orientation, shape, and surface texture that there is very little agreement about how judgments are made.

The lack of consensus is due to many factors. The initial studies were mainly demonstrations that echolocation by humans was in fact possible so that potential acoustic cues were not measured nor correlated to performance. Recent studies have imposed more control to determine which acoustic cues were available and which correlate to accuracy (Kopco & Shinn-Cunningham, 2011; Papadopoulos, Edwards, Rowan, & Allen, 2011). To do so, researchers have first recorded the source and echoes from a surface at the head of a manikin in an anechoic or in a reverberant enclosure. The sounds measured at the manikin's ears (sometimes termed *phantom* sounds) are altered in various ways to accentuate or eliminate acoustic variables (e.g., removing low frequencies), and then presenting the higher frequencies to participants by means of headphones. The importance of the cues can be evaluated from the accuracy of the judgments. The downside is that the participants become passive, and unable either to vary their source output or move their heads. Even so, it is unclear how movement affects the accuracy of echolocation (Wallmeier & Wiegrebe, 2014). In general, we do not know how this research generalizes to real-life judgments.

Probably the biggest stumbling block is the assumption that there is a fixed set of cues. Depending on context, individuals will make use of any cue that works. For example, Kopco and Shinn-Cunningham (2011) suggest that in highly reverberant rooms, participants use the direct to reverberant energy for low frequencies particularly for nearby surfaces, but in low reverberant rooms use the interaural level differences. The highly reverberant room will tend to mask the level difference.

Perceiving should be opportunistic. Blind people can use self-generated sound to navigate, but can also make use of their knowledge of environmental sounds. We have mentioned individual sounds such as thunder and fire engines, but there are also "walls" of sound such as rivers or a lines of traffic, and the trajectories of individual cars or black flies. It seems impossible to construct a "neutral" experiment, all perceiving occurs in a context and the best we can do is discover the optimal cues in each such context.

5.7.2 Physiological Mechanisms

Recent experiments have investigated the neural underpinnings of echolocation for early blind, late blind, and sighted. Several studies have found increased activation in secondary visual areas during echolocation for the early blind but not for the late blind or the sighted participants. Bauer et al. (2017) found

anatomical changes in the early blind; in some brain areas the density of their connections increased, while in other areas the density decreased. As described above, the early-blind experts outperform the others in various tasks and that leads to the notion that the recruitment of traditional visual areas gives rise to the superior discrimination.

I am skeptical about this interpretation.

1. In reality, we do not know if the activation actually causes the improved performance or simply follows the auditory "calculation." All we know is that there is more blood flow which is presumed to indicate increased neural activity. The translation of that activity into distance or surface judgments is completely opaque.
2. Early-blind experts learn to correlate auditory cues to aspects of the environment and that would necessarily require some way to calibrate those cues. The obvious way to do this is reaching out to judge distance and orientation, or running a hand around a surface to judge shape. The motor and tactual systems connect the auditory cues to the environment. It is interesting to note that it is harder to make distance judgments for overhead surfaces that cannot be reached than frontal or lateral surfaces that can be touched. It would seem that echolocation in the early blind would be accompanied by increased activation in the tactual and motor cortical areas, rather than in the visual areas. In contrast, the late blind and sighted would probably calibrate the acoustic cues to objects and surfaces through their visual memory so that it seems reasonable that visual areas would be activated. But, that is not the case.
3. Spatial information from echolocation and from direct visual input is fundamentally different. The spatial information underlying echolocation is temporal, involving the interval between the tongue click and the returning echo, or the intervals between successive echoes in reverberant enclosures. The visual information is spatial, the extension and adjacency of extended surfaces, or the connected movement of surfaces. We do not have any understanding how visual brain circuits presumably organized for spatial relationships can encode temporal relationships to derive the absent spatial relationships. It could be that echolocation occurs by means of rhythmic timing circuits, which occur across cortical regions.

5.7.3 Echolocation Summary

Clearly we have strayed from the technical definition of echolocation to include material on passive acoustic information useful for orientation, navigation, and exploring objects. The source-filter model is most relevant to understand active echolocation, but a more generous definition of the filter, to include the acoustics of the general environment, can provide insights about how blind and

sighted can maneuver and make sense of the external world. What is common is the convergence of multiple cues that may be substitutable in some instances and whose usefulness may vary in other instances.

5.8 Overall Summary

Visual and auditory stimulation results from excitation impinging upon and being modified by the properties of the filter, and then being propagated to the perceiver. Source-filter models explain many of the physical properties of the energy reaching the eyes and ears and give insights into why perceiving is so difficult. The same object can look and sound so different in different contexts because the proximal stimulation can arise from many different combinations of excitation and filter. The perceiver is forced to interpret the proximal stimulation in its particular context. Depending on that interpretation, people can see or hear different objects, as found for "the dress." Given the potential for alternative percepts, it is amazing to me that so few interfere with our actions or cause bodily harm.

References

Allen, E. J., & Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *Journal of the Acoustical Society of America, 135*, 1371–1379. https://doi.org/10.1121/1.4863269

American National Standards Institute. (1973). *Psychoacoustics terminology*. S3.20. New York, NY: American National Standards Institute.

Ammons, C. H., Worchel, P., & Dallenbach, K. M. (1953). "Facial vision": The perception of obstacles out of doors by blindfolded and blindfolded-deafened subjects. *American Journal of Psychology, 56*, 519–553.

Ashmead, D. H., & Wall, R. S. (1999). Auditory perception of walls via spectral variations in the ambient sound field. *Journal of Rehabilitation Research and Development, 36*, 313–322.

Ballas, J. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 250–267.

Bauer, C. M., Hirsh, G. V., Zajac, L., Koo, B.-B., Collignon, O., & Merabet, L. B. (2017). Multimodal MRI-imaging reveals large-scale structural and functional connectivity changes in profound early blindness. *PLoS One, 12*, 1–26. https://doi.org/10.1371/journal.pone.0173064

Brainard, D. H., & Hurlbert, A. C. (2015). Colour vision: Understanding #thedress. *Current Biology, 25*, R549–R568. https://doi.org/10.1016/j.cub.2015.05.020

Brainard, D. H., & Maloney, L. T. (2011). Surface color perception and equivalent illumination models. *Journal of Vision, 11*(5), 1–18. https://doi.org/10.1167/11.5.1

Bramao, I., Reis, A., Peterson, K. M., & Faisca, L. (2011). The role of color information on object recognition: A review and meta-analysis. *Acta Psychologica, 136*, 244–253. https://doi.org/10.1016/j.actpsy.2011.06.010

Cornsweet, T. (1970). *Visual perception*. New York, NY: Academic Press.

Cott, H. B. (1940). *Adaptive coloration in animals*. London, UK: Methuen.

Dedrick, D. (2015). Some philosophical questions about color. In A. J. Elliot, M. D. Fairchild, & A. Franklin (Eds.), *Handbook of color psychology* (pp. 131–145). Cambridge, UK: Cambridge University Press.

Donelli, A., Brayda, L., & Gori, M. (2016). Depth echolocation learnt by novice sighted people. *PLoS One, 11*, e0156654. https://doi.org/10.1167/journal.pone.0156654

Erickson, M. L., & Perry, S. R. (2003). Can listeners hear who is singing? A comparison of three-note and six-note discrimination tasks. *Journal of Voice, 17*, 353–369. https://doi.org/10.1067/S0892-19970903000021-3

Fujisaki, W., Soda, N., Motoyoshi, I., Komatsu, H., & Nishida, S. (2014). Audiovisual integration in the human perception of materials. *Journal of Vision, 14*, 1–20. https://doi.org/10.1167/14.4.12

Fujisaki, W., Tokita, M., & Kariya, K. (2015). Perception of the material properties of wood based on vision, audition and touch. *Vision Research, 109*, 185–200. https://doi.org/10.1016/j.visres.2014.11.020

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology, 5*, 1–29.

Gegenfurtner, K. R., Bloj, M., & Toscari, M. (2015). The many colours of "the dress". *Current Biology, 25*, R523–R548. https://doi.org/10.1016/j.cub.2015.04.043

Gilchrist, A. (2015). Perceptual organization in lightness. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 391–412). Oxford, UK: Oxford University Press.

Giordano, B. L., & McAdams, S. (2010). Sound source mechanics and musical timbre: Evidence from previous studies. *Music Perception, 28*, 155–168. https://doi.org/10.1525/mp.2010.28.2.155

Gough, C. E. (2016). Violin acoustics. *Acoustics Today, 12*(2), 22–30.

Guyot, P., Houix, O., Misdaris, N., Susini, P., Pinquier, J., & Andre-Obrecht, R. (2017). Identification of categories of liquid sounds. *Journal of the Acoustical Society of America, 142*, 878–889. https://doi.org/10.1121/1.4996124

Handel, S., & Erickson, M. L. (2001). A rule of thumb: The bandwidth for timbre invariance is an octave. *Music Perception, 19*, 121–0126. https://doi.org/10.1525/mp.2001.19.1.121

Haywood, N. R., & Roberts, B. (2013). Build-up of auditory stream segregation induced by tone sequences of constant or alternating frequency and the resetting effects of single deviants. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 1652–1666. https://doi.org/10.1037/a0032562

Hjoetkjaer, J., & McAdams, S. (2016). Spectral and temporal cues for perception of material and action categories in impacted sound sources. *Journal of the Acoustical Society of America, 140*, 409–420. https://doi.org/10.1121/1.4955181

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kaiser, P. K., & Boynton, R. M. (1996). *Human color vision* (2nd ed.). Washington, DC: Optical Society of America.

Katz, D. (1925). *Der Aufbauder Tastwelt (The World of Touch)* (L. E. Krueger, trans. & Ed.). Hillsdale, NJ: LEA Associates.

Kolarik, A. J., Cirstea, S., Pardhan, S., & Moore, B. C. J. (2014). A summary of research investigating echolocation abilities of blind and sighted humans. *Hearing Research, 310*, 60–68. https://doi.org/10.1016/j.heares.2014.01.010

Kopco, N., & Shinn-Cunningham, B. G. (2011). Effect of stimulus spectrum on distance perception for nearby sources. *Journal of the Acoustical Society of America, 130*, 1530–1541. https://doi.org/10.1121/1.3613705

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbre: Common dimensions, specificities, and latent subject classes. *Psychological Research, 58*, 177–192. https://doi.org/10.1007/BF00419633

Nathmann, A., & Malcolm, G. L. (2016). Eye guidance during real-world scene search: The role color plays in central and peripheral vision. *Journal of Vision, 16*, 1–16. https://doi.org/10.1167/16.2.3

O'Modhrain, S., & Gillespie, R. B. (2018). One more, with feeling: Revisiting the role of touch in performer-instrument interaction. In S. Papetti & C. Saitis (Eds.), *Musical haptics* (Vol. 18, pp. 11–27). Springer.

Ogg, l., Sleve, R., & Idsardi, J. (2017). The time course of sound category identification: Insights into acoustic features. *Journal of the Acoustical Society of America, 142*, 3459–3473. https://doi.org/10.1121/1.5014057

Papadopoulos, T., Edwards, D. S., Rowan, D., & Allen, R. (2011). Identification of auditory cues utilized in human echolocation-objective measurement results. *Biomedical Signal Processing and Control, 6*, 280–290. https://doi.org/10.1016/jbspc.2011.03.005

Radonjić, A., Gottaris, N. P., & Brainard, D. H. (2015). Color constancy in a naturalistic, goal-directed task. *Journal of Vision, 15*, 1–21. https://doi.org/10.1167/15.13.3

Rowland, H. M. (2009). From Abbot Thayer to the present day: What have we learned about the function of countershading? *Philosophical Transactions of the Royal Society, Section B, 364*, 519–527. https://doi.org/10.1098/rstb.2008.0261

Saitis, C., Järveläinen, H., & Fritz, C. (2018). The role of haptic cues in musical instrument quality perception. In S. Papetti & C. Saitis (Eds.), *Musical haptics* (pp. 73–93). Springer.

Steele, K. M., & Williams, A. K. (2006). Is the bandwidth for timbre invariance only one octave? *Music Perception, 23*, 215–220. https://doi.org/10.1525/mp.2006.23.215

Thaler, L., Reich, G. W., Zhang, X., Wang, D., Smith, G. E., Tao, Z., … Antonio, M. (2017). Mouth-clicks used by blind expert human echolocators – Signal description and model based signal synthesis. *PLoS Computational Biology, 13*, e1005670. https://doi.org/10.1371/journal.pchi.1005670

Thayer, A. H. (1909). *Concealing-coloration in the animal kingdom: An exposition of the laws of disguise through color and pattern: Being a summary of abbot H. Thayer's discoveries.* New York, NY: Macmillan.

Wallisch, P. (2017). Illumination assumptions account for individual differences in the perceptual interpretation of a profoundly ambiguous stimulus in the color domain: "The dress". *Journal of Vision, 17*, 5. https://doi.org/10.1167/17.4.5

Wallmeier, L., GeBele, N., & Wiegrebe, L. (2013). Echolocation versus echo suppression in humans. *Proceedings of the Royal Society, B, 280*, 2013.1428. https://doi.org/10.1098/rspb.2013.1428

Wallmeier, L., & Wiegrebe, L. (2014). Ranging in human sonar: Effects of additional early reflections and exploratory head movements. *PLoS One, 9*, e15363. https://doi.org/10.1371/journal.pone.0115363

Warren, W. H. J., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 704–712.

Witzel, C., Racey, C., & O'Reagan, J. K. (2017). The most reasonable explanation of the "dress": Implicit assumptions about illumination. *Journal of Vision, 17*, 1–19. https://doi.org/10.1016/17.2.1

CHAPTER 6

Summary

Throughout, I have made the argument that we feel that perceiving is effortless and accurate regardless of context. But, it is difficult to understand how "getting it right" is accomplished. Perceiving things depends on the simultaneous integration of many physical attributes and the transformation of the resulting neural signals in the central nervous system. In fact, I often wonder how it is possible to perceive at all. In the end, it is a "best estimate" based on the input sensory signals that are modified by top-down neural control processes, context, and past experience. Instead of a static word like perception, we should use a dynamic word like perceiving to emphasize that we are constantly adapting our actions and knowledge to a changing world.

In this short summary, we return to the basic problem of constructing objects and sources in extended space and time. To perceive those objects and sources, we first need to identify what parts of visual, auditory, and tactual sensations go with each object or source, that is, which are the connected points. Edges and gaps, both physically present and inferred, bound objects and sources and separate them; edges belong to the objects and sources, and place figures in front of grounds. Temporal synchrony and rhythmic grouping bind sounds into sources and allow for the hearing of simultaneous sources even when the sensations from each are mashed together. But as the apparent movement and Ternus displays show, changes in either spatial arrangements or in the timing between the displays alter the other dimension. Local processes must be integrated into global ones.

Not only do we have to construct the objects, surfaces, and sources, we have to track them over time as they transform and emerge in different forms and backgrounds. In some cases we can solve the correspondence problem by picking up the predictable changes arising from physical properties. On the whole, physical changes are correlated with one another, allowing us to distinguish among alternative possibilities. If the vertical and horizontal dimensions of a square change equally, for example, that suggests back or forth movement. But

S. Handel, *Perceptual Organization*,
https://doi.org/10.1007/978-3-319-96337-2_6

if only one dimension changes, that suggests rotation. Correlated or redundant changes reduce errors, since one change can lead to the correct percept even in the absence of all others.

In other cases, there is no simple solution to the correspondence problem. Consider face recognition, all faces have the same parts (e.g., foreheads, eyes, noses, lips, etc.) and the configuration is the same, eyes above noses, noses above lips, and so on. The problem then, is to identify specific individuals amidst many similar others (Behrmann, Richler, Avidan, & Kimchi, 2015). The prevalent naïve view is that people are very good at it in spite of the numerous cases of false eyewitness identifications. What is actually true is more complex, and it bears on all the issues discussed here. The task is simple: people view one photograph of an individual (even for extended periods of time), and then are asked to identify that individual in a different pose (i.e., different haircut, shadows, facial expression, gaze direction) among a set of photographs of different individuals. People are very good at identifying familiar faces, but are quite poor at identifying unfamiliar faces. In fact, for most people even extensive training does not improve performance for novel faces. Young and Burton (2017) suggest that having seen a familiar face in many different poses, we have built up a model of how the face will look in still other poses. In other words, for that face we have learned to separate the stable characteristics from the variable ones and have learned the transformations that connect all the possible poses. But the variability and resulting transformations will differ for each face, so that without extensive experience with each new face, people are unable to match those poses.

Since the variability associated with each face is idiosyncratic it is impossible to create a general transformation that will allow us to identify any face in a new pose. We cannot solve the correspondence problem without observing the poses of each face separately. This outcome is not unique to faces but is true for the discrimination of tonal and atonal melodies discussed in Chap. 2, the discrimination of instruments at different pitches discussed in Chap. 5, the identification of singers at different notes discussed in Chap. 5, and the detection of human movements discussed in Chap. 2. A good parallel is the identification of instruments at different pitches taken up in Chap. 5. Musicians who are experienced with one instrument are able to identify that instrument at different pitches, but they are not able to do so for unfamiliar instruments. Each instrument changes timbre in idiosyncratic ways across their playing range, so experience with one instrument does not transfer to other instruments.

Perceiving occurs at the same time across multiple temporal and spatial levels that are interlocked. Small visual spatial regions are grouped into larger more encompassing areas, small uniform surface regions become part of larger varying surfaces, and individual sounds and beats turn into rhythmic meters that organize time. This ability to see, touch, and hear a sequence at different time intervals and spatial distances yields a hierarchical representation in which the levels are not independent, but are constrained by each other. They are not

like Russian Dolls in which independent, fully formed smaller dolls fit inside independent, fully formed larger ones.

Rhythms create multilayered figures in time and space. One can zoom in and out. When listening it is possible to react to every beat or to beats widely separated in time. Without the faster beats, the slower beats are simply recurring accents, and without the slower beats, the faster ones also are simply recurring accents. Meter is the mental construct that fuses the beats at different levels together and organizes time. In similar fashion, when looking it is possible to view a narrow spatial field or a wide one. Viewers construct layers to group the visual regions into larger, more encompassing areas containing overlapping objects. But, the auditory meter and the visual space do not simply match the auditory or visual sensory features. The meter and space emerge from the interaction of those sensory features and the perceptual acts.

Another compelling example of how these interacting multiple layers create our perceptual world is demonstrated by our ability to abstract the movements of parts of objects moving in different trajectories and at differing speeds as shown in the videos by Johansson (1973) in Chap. 2 and Toiviainen, Luck, and Thompson (2010) in Chap. 4. The slower components act as frames of reference for the faster ones, so that it is possible to partition the movement of the faster components into two parts: One part common to the slower component that "stays" with the slower one, and one part that is unique to the faster component and is seen separately. The perception of auditory meter based on beats at differing rates, and the perception of visual motion based on movements at different speeds, is therefore based on the same concept of hierarchical layers. I am strongly drawn to this conceptualization. It makes theoretical questions such as whether perception is bottom-up or top-down largely irrelevant, and it also seems to make distinctions about brain localization unlikely to be useful in explaining what we see, feel, or hear.

Ultimately, there is no single answer to the basic questions of "why do we see, hear, and touch what we see, hear, and touch" or "why do objects and sources look, sound, and feel the way they do." Furthermore, the sensory and perceptual processes are not linear or unidirectional: it is not the case the A leads to B, B leads to C, and so on. The processes always interact, the feedback from later stages modify the outputs from earlier ones. The receptors at the eye, ear, and hand transform the sensory energy into neural signals. Those signals are further modified at higher brain regions before reaching the cortex. The neural information is underdetermined; many external objects could have generated the same sensory energy and the transformed neural signal. The cortical firings are interpreted in terms of their coherence and organization, and the person's expectations (the prior probabilities in Bayesian terms), to give rise to the percept. None of this is simple or straightforward; that's what makes perceiving so surprising and interesting.

References

Behrmann, M., Richler, J., Avidan, G., & Kimchi, R. (2015). Holistic face perception. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 758–774). Oxford, UK: Oxford University Press.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*, 201–211.

Toiviainen, P., Luck, G., & Thompson, M. R. (2010). Embodied meter: Hierarchical eigenmodes in music induced movement. *Music Perception, 28*(1), 59–70. https://doi.org/10.1525/mp.2010.28.1.59

Young, A. W., & Burton, A. M. (2017). Recognizing faces. *Psychological Science, 26*, 212–217. https://doi.org/10.1177/0963721416688114