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.

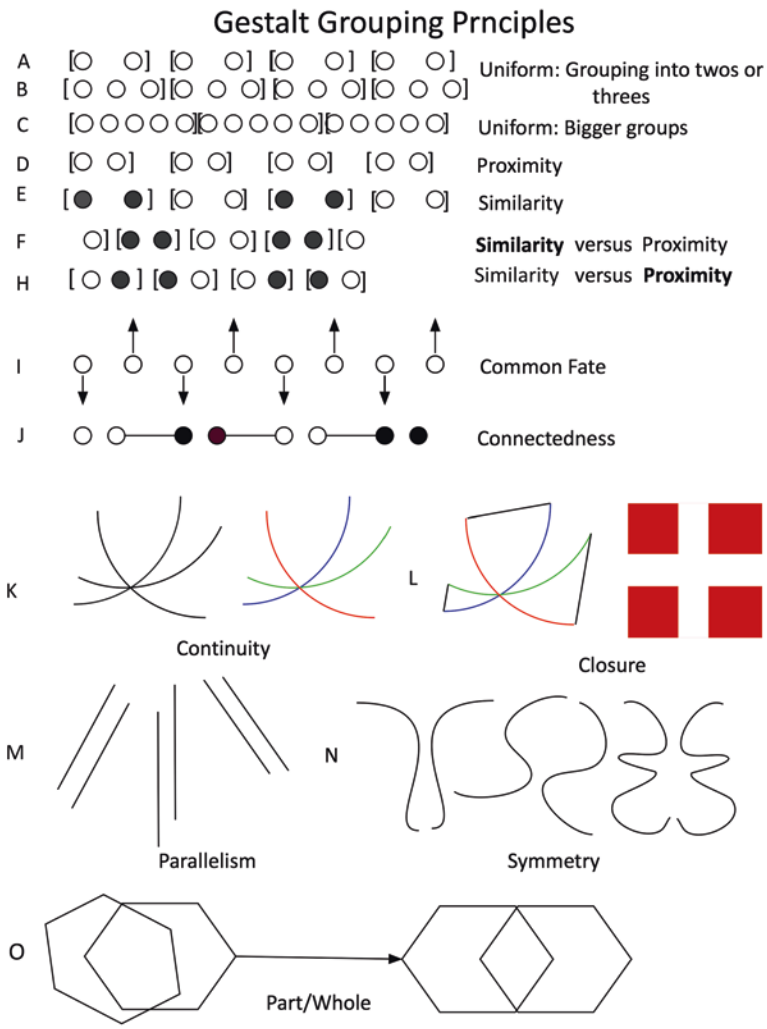

Fig. 2.1 Examples of the classical Gestalt grouping principles. It is easy to see how the groupings change as the proximity or similarity among elements is varied (F & H) or when extra elements are added or subtracted. Connectedness (J) can overcome the principles of similarity and proximity. The three 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

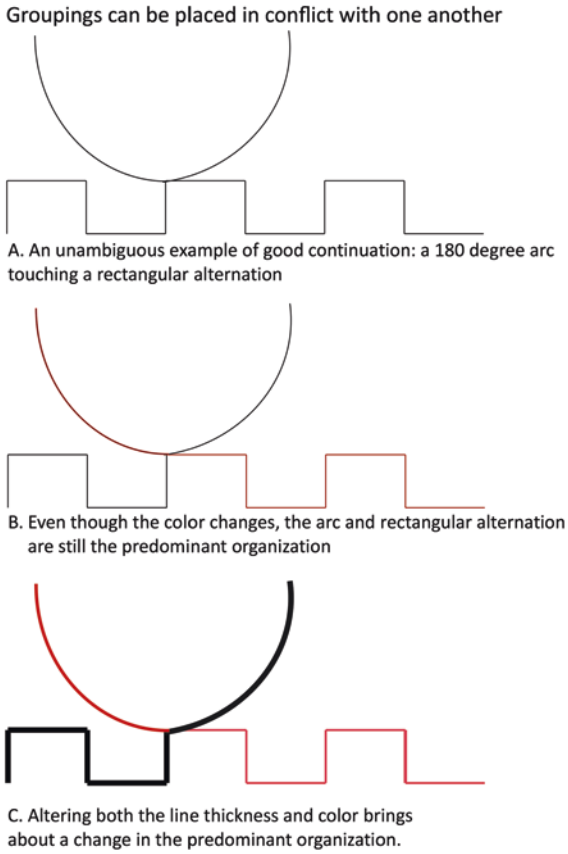

Fig. 2.2 The perceptual grouping is a reflection of the relative strengths of the grouping principles, which can be easily altered. Here, the shift is from continuity to color/line thickness similarity

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

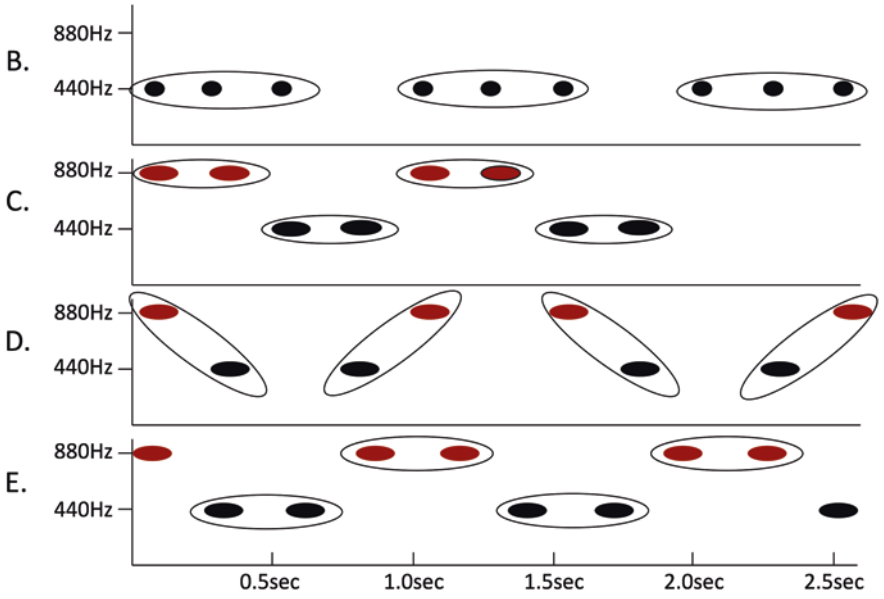

Fig. 2.3 Illustrations of sound files 2.3B–E

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.

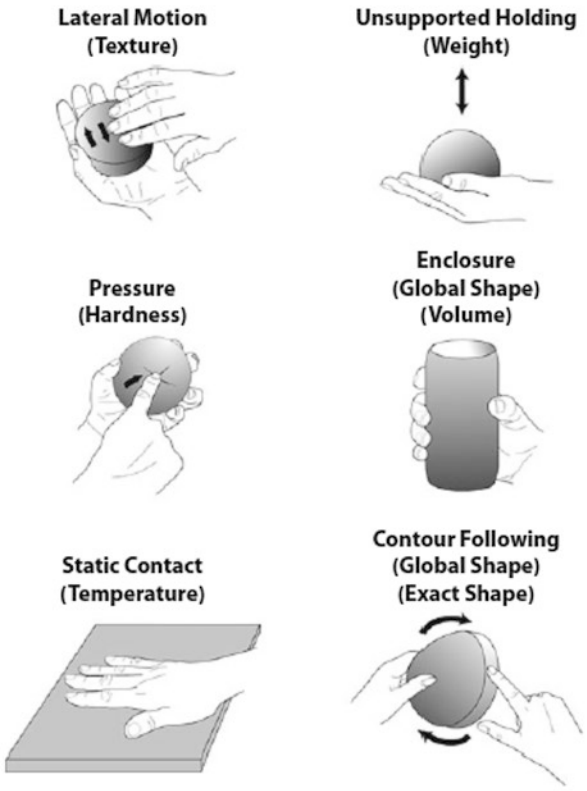

Fig. 2.4 The six exploratory procedures found by Lederman and Klatzky (1987). The "inside" region of the hand is critical. The ridges of the epidermis, which surprisingly act to reduce skin friction due to reduce surface contact, generate oscillations on the skin during sliding motions. Directly below 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

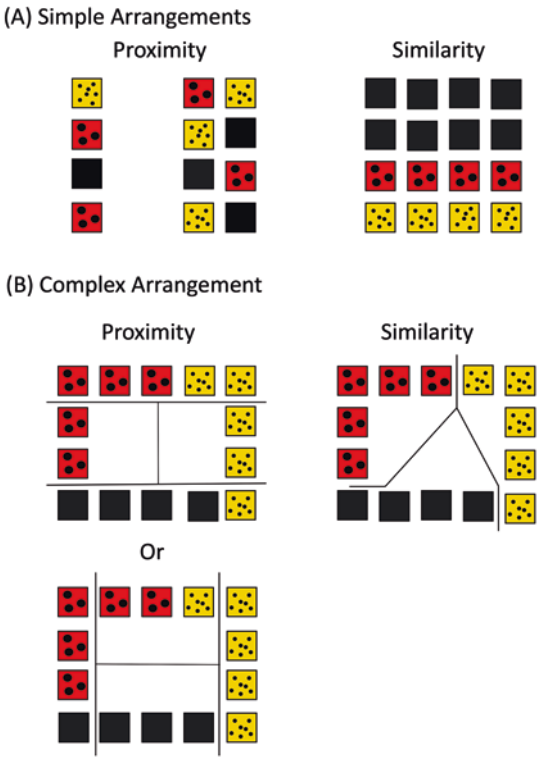

Fig. 2.5 Simple and complex arrangements for visual and tactual grouping. The 240-grit stimuli are represented by the yellow squares/small dots and the 40-grit stimuli by the red squares/larger dots and smooth stimuli by the black squares. These are hypothesized organizations based on similarity and 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

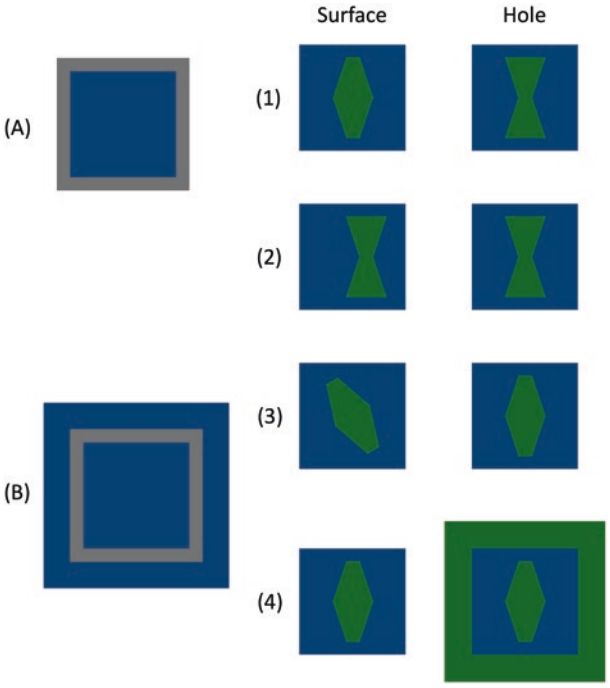

Fig. 2.6 In both (A) and (B), the blue regions can be seen either in front of or in back of the grey cut-out. Moreover, in (B), the blue regions can be seen as one or two surfaces. The perception of a green object sitting on the blue background is more likely if the object is concave (1), offset laterally (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

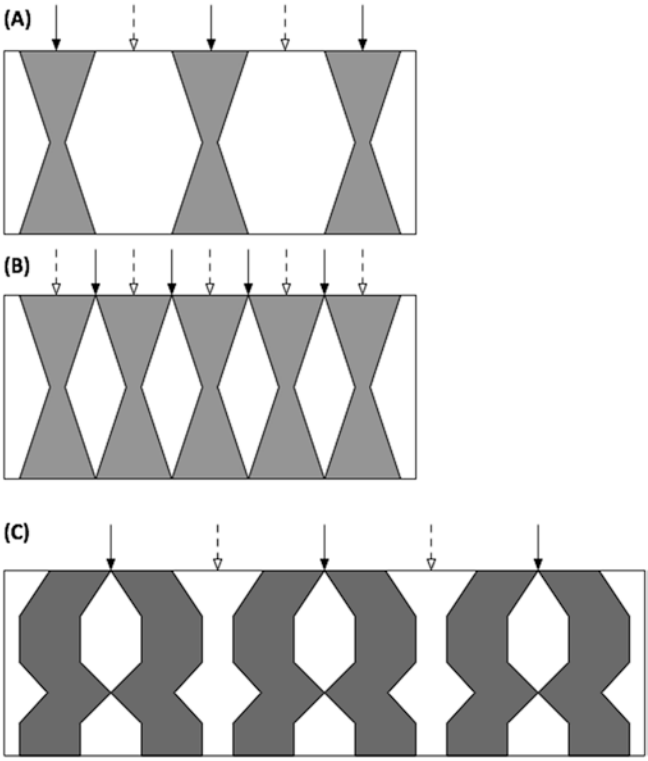

Fig. 2.7 Several factors influence the perception of the in-front figure and the behind ground. Comparison of (A) and (B) show the effect of size (and possibly convexity), and (C) shows the influence of convexity and parallelness. To me, the figure surfaces lie under the solid arrows, although it is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

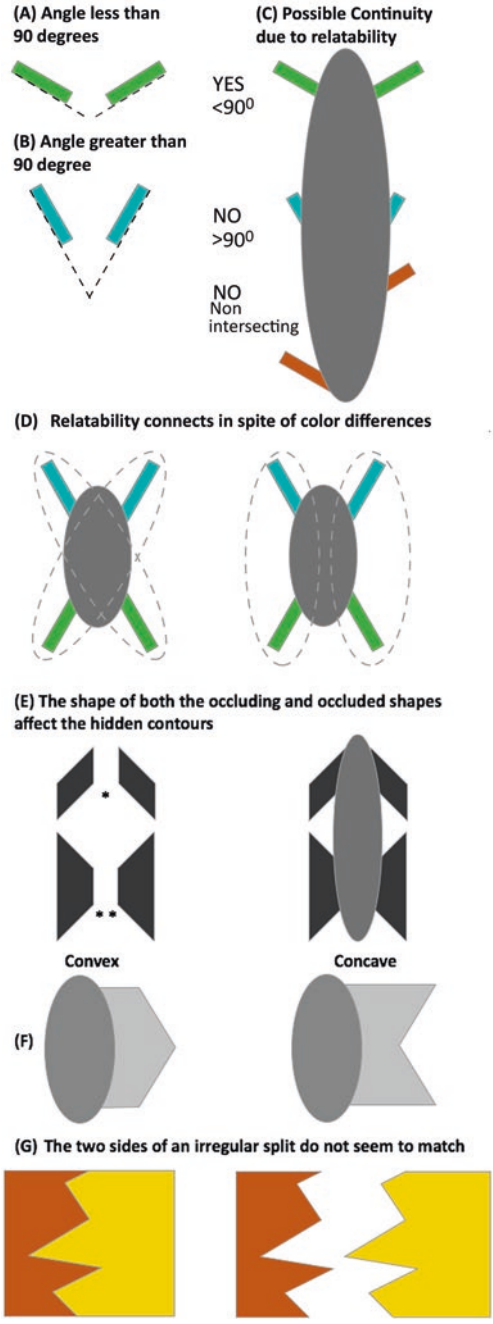

Fig. 2.8 In (A), the "turn" is less than 90° so that occluded parts would appear to be connected. But, in (B) the "turn" is greater than 90° so that the parts would not seem to connect. In (C), the different turn angles create the perception of connectedness only for the top green bars. In (D), relatable 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

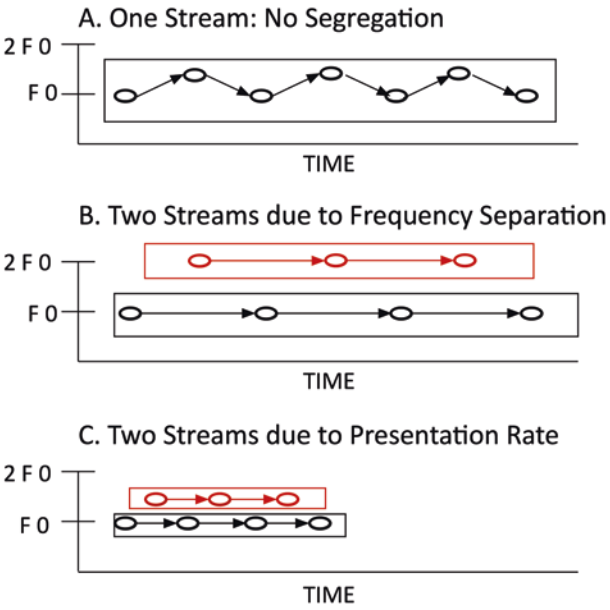

Fig. 2.9 Stream segregation arises if the frequency separation is increased (B) or the presentation rate is increased (C)

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

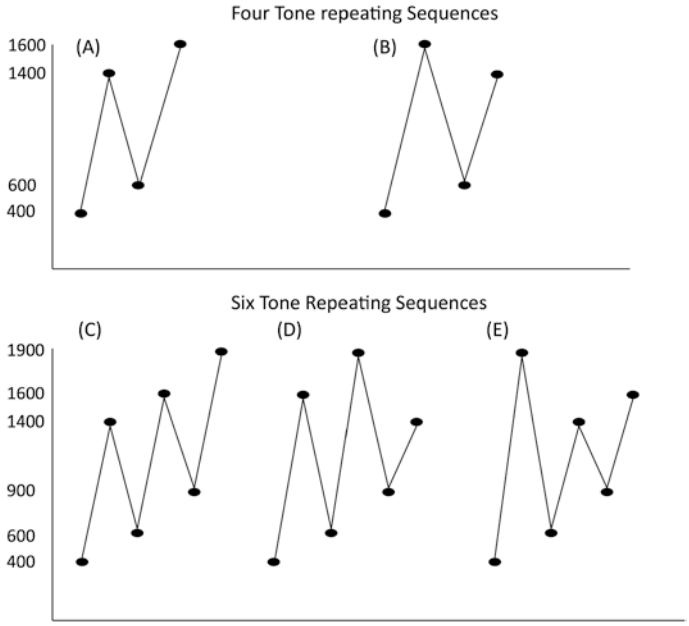

Fig. 2.10 The two versions of a four-tone repeating sequence composed of two low-pitch and two high-pitch tones are shown for two cycles. The order for (A) is 400 Hz, 1400 Hz, 600 Hz, 1600 Hz and the order for (B) is 400 Hz, 1600 Hz, 600 Hz, 1400 Hz. The three versions of a six-tone repeating sequence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

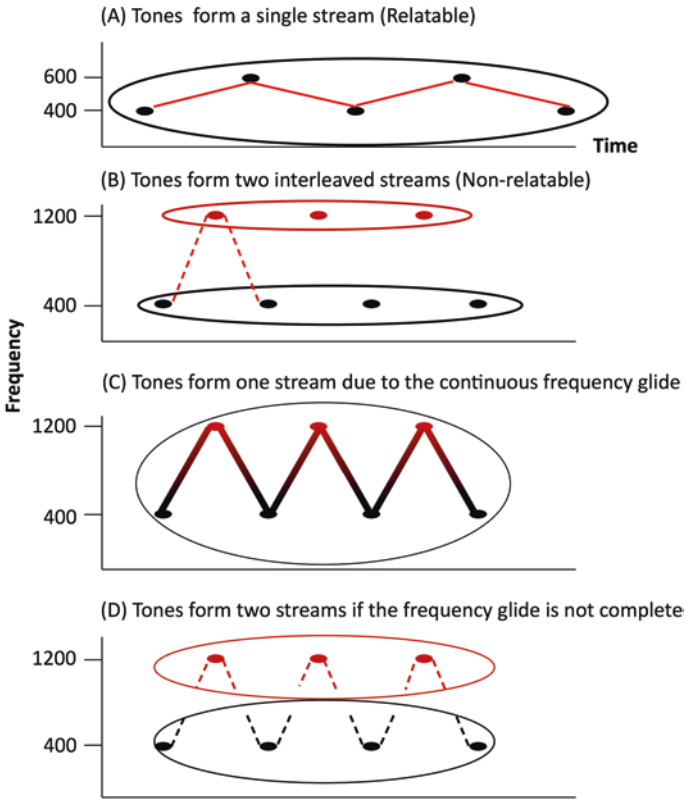

Fig. 2.11 (A) If the contour connecting the alternating tones is flat (i.e., relatable depicted by the solid red lines), the tones form one stream. (B) If the contour is sharp (i.e., non-relatable), the tones form two independent streams. (C) A frequency glide connecting the tones brings about one stream,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

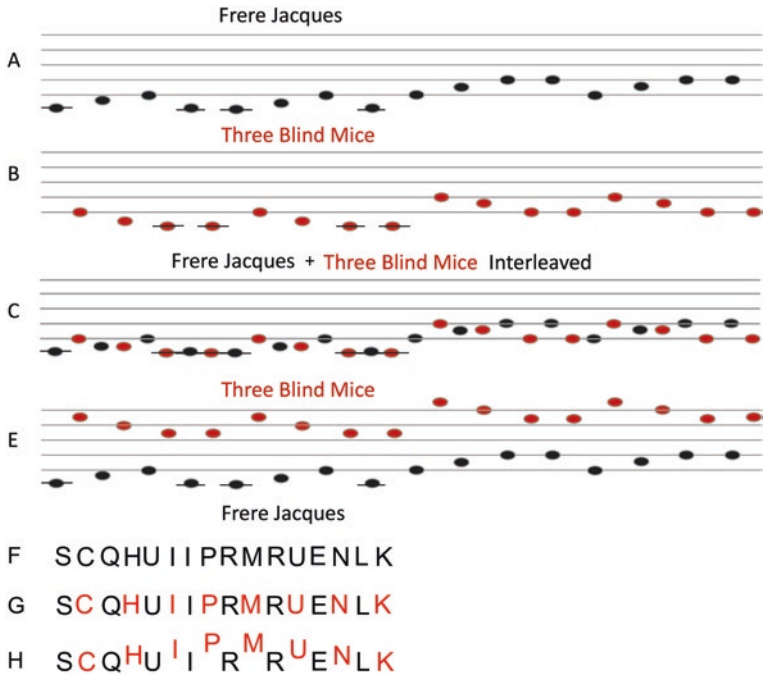

Fig. 2.12 "Frère Jacques" and "Three Blind Mice" are illustrated in (A) and (B). In (C), they are interleaved so that the contour has many simple repetitions and it is nearly impossible to pick out the two tunes. In (E), the notes of one tune ("Three Blind Mice" in red) are shifted by an octave; the 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

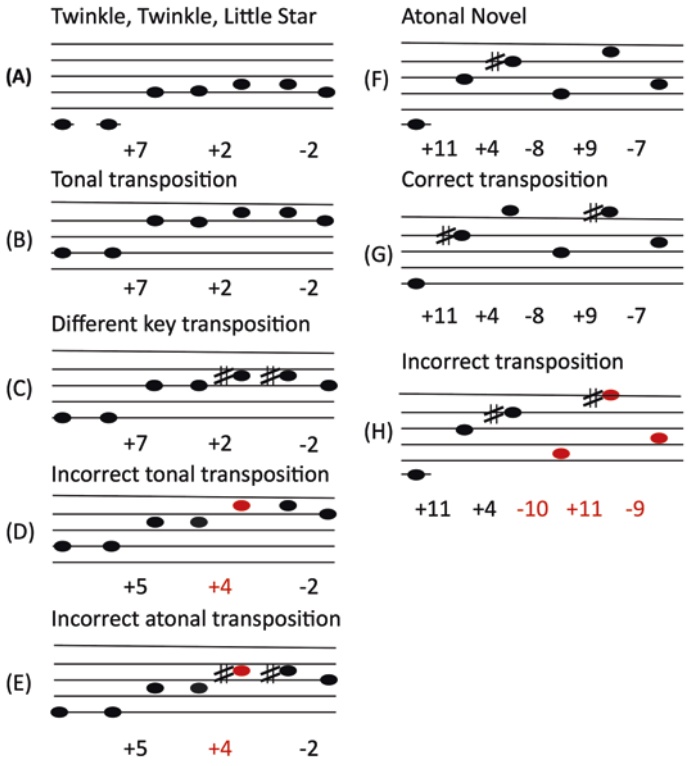

Fig. 2.13 The "target" melodies are (A) and (F). Listed beneath these short melodies are the numbers of semitone steps between the two surrounding notes. For "Twinkle, Twinkle, Little Star" (A), the correct transpositions (B) and (C) have the identical number of steps between notes. The incorrect transpositions (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

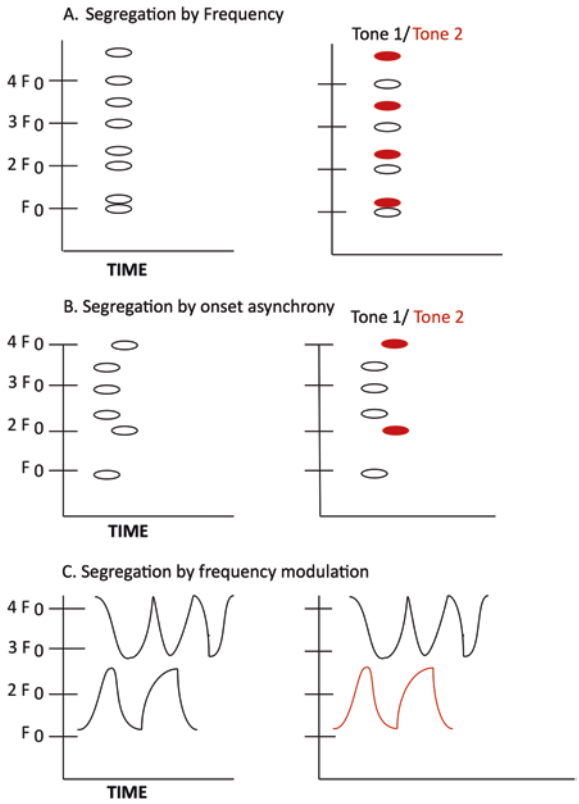

Fig. 2.14 In all three panels, the left side presents the auditory sound as a function of time along the horizontal axis, and frequency along the vertical axis. The right side represents the segregation into two parts. In (A), segregation is due to the harmonic relationships among the frequency components. 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.

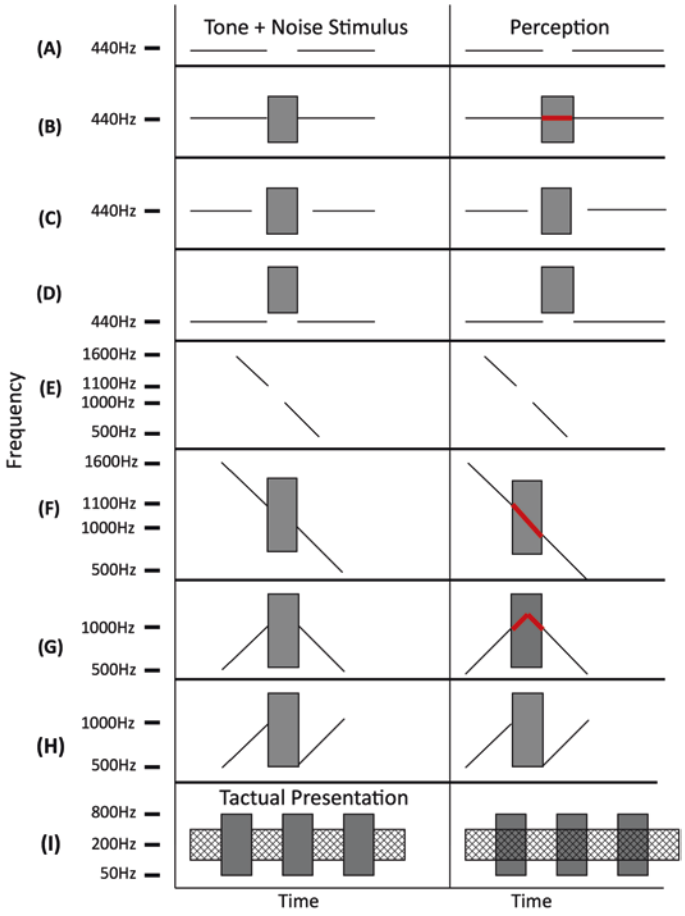

Fig. 2.15 Eight configurations are shown to illustrate the masking of a tone by a noise burst. A split tone is presented in (A) and a split glide is presented in (E). In both cases, the perception is veridical; two separate tones or glides are heard. In (B) and (F), a noise burst is inserted in the gap 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

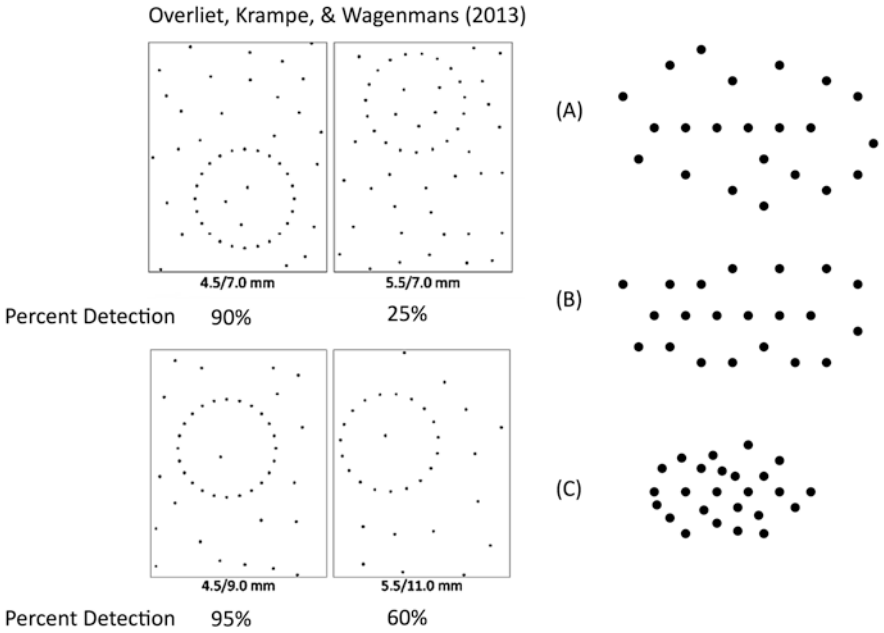

Fig. 2.17 Detection of raised contours. If the target dots are spaced closely, the circle target is easy to detect regardless of the spacing of the masking dots. But, if the target dots are spaced further apart (5.5 mm), then the spacing of the masking dots can completely obscure the target. It is relatively 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

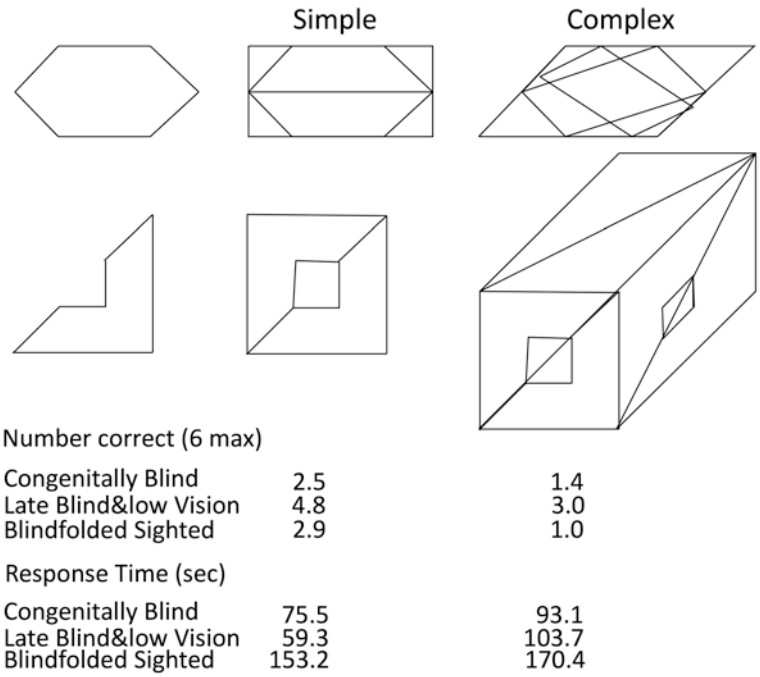

Fig. 2.18 Examples of simple and complex embedding of two geometric figures from Heller et al. (2003). The data are the averages across all the figures

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.

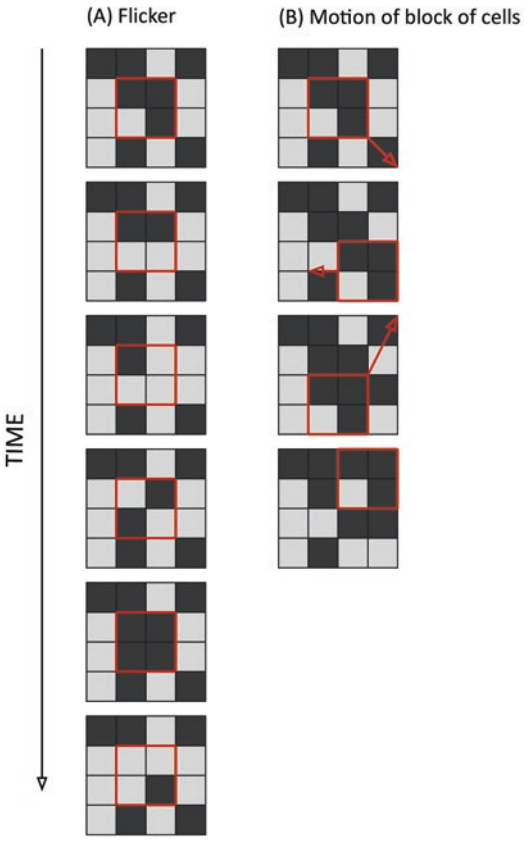

Fig. 2.19 (A) If the cells in a fixed region randomly shift brightness, that region appears to flicker. (B) If the cells in a fixed region shift position, those cells appear to move and float above the background cells

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

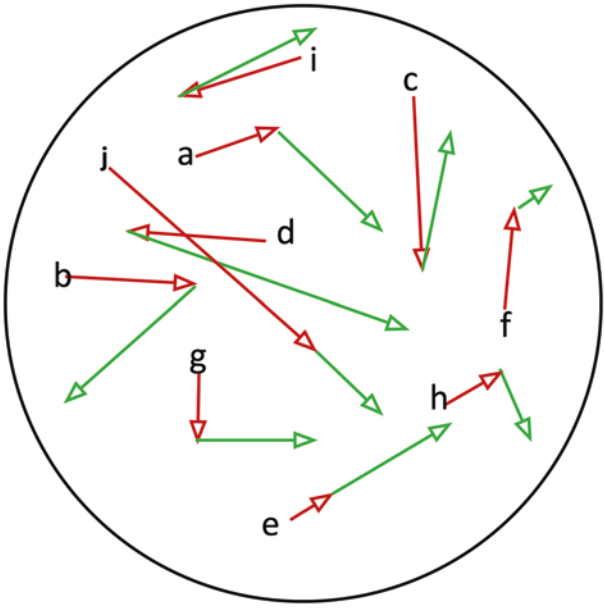

Fig. 2.20 The motions of 10 points are shown. The first step is drawn in red and the second in green. Only two points (e & j) move in the same direction on both steps. The common motion, although carried by different points (c,e,f,i), is up to the right

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.

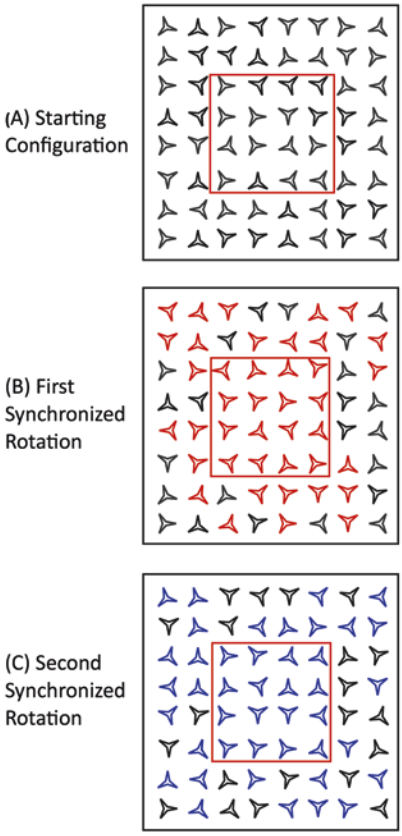

Fig. 2.21 In the starting configuration (A), the "windmills" are oriented randomly. In the first rotation (B), the windmills within the figure region rotate randomly in both direction and number of degrees (in red). Some of the windmills outside of the figure region also rotate, shown in red. In the 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.

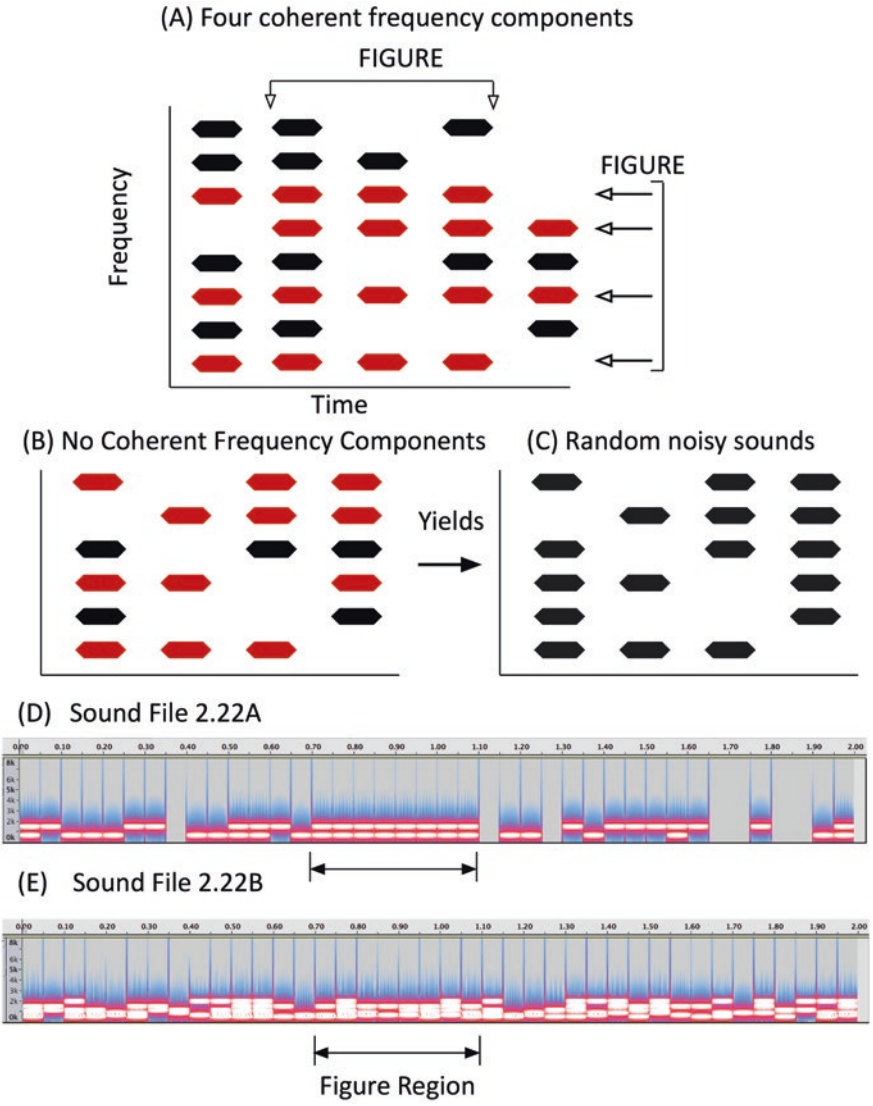

Fig. 2.22 In (A), the four coherent frequency components (in red) in the middle three sounds form a figure, that is, a sound that seems to occur in each of the three sounds in spite of the other overlapping components. In (B), even though each (red) component occurs in three of the four sounds, none 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-

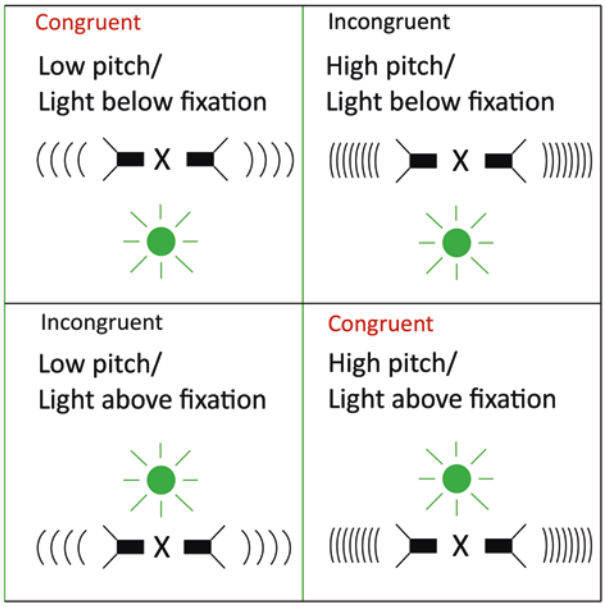

Fig. 2.23 Congruent and incongruent stimuli used to study cross-modal correspondence. The "x" is the fixation point, and the sound is often presented by two speakers equilateral from the fixation point placed behind a screen

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

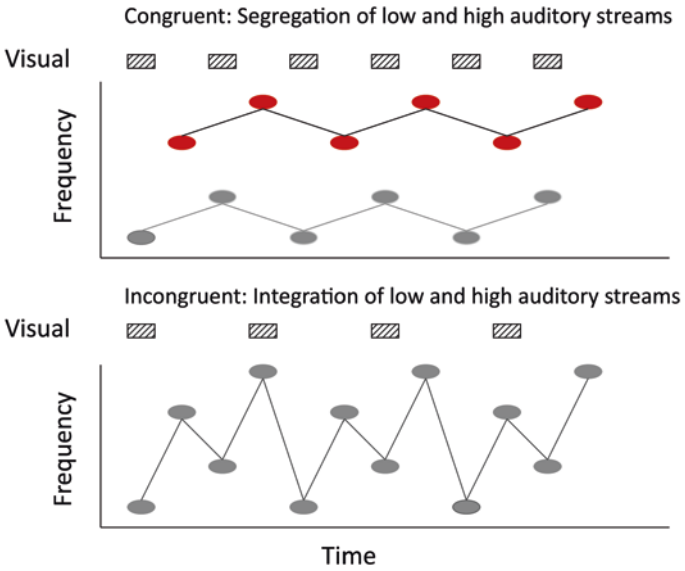

Fig. 2.24 Visual stimuli were presented with the low-pitch tones in the congruent presentation condition and that promoted segregation. In contrast, the visual stimuli were presented with every fourth tone in the incongruent condition. Here, the light occurs equally often with each high and low tone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

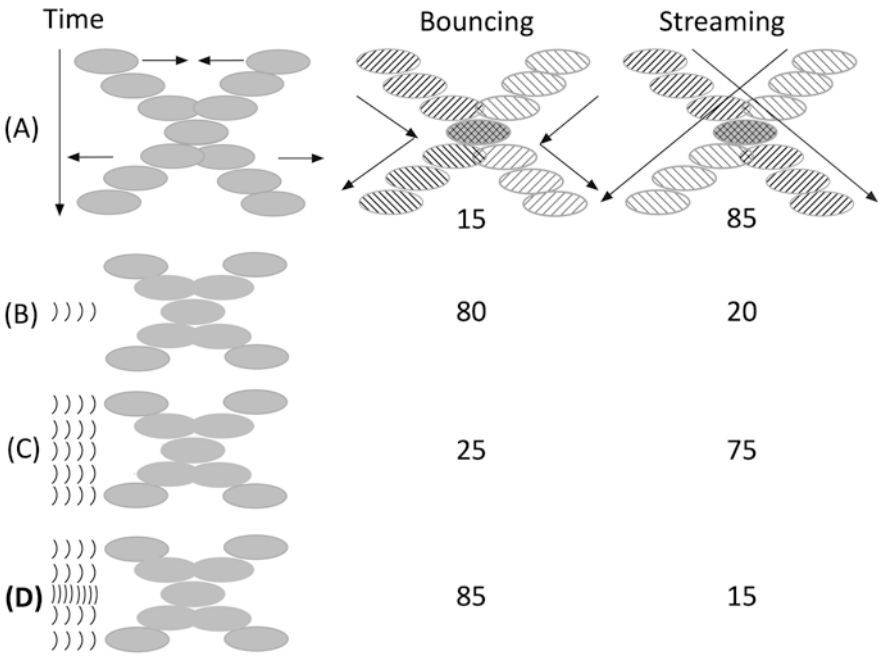

Fig. 2.25 When two lighted circles approach each other and then separate, two percepts are possible. The circles could appear to cross each other and continue on their way or they could appear to bounce off each other and return to their original locations. The normal perception is streaming (A). If 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

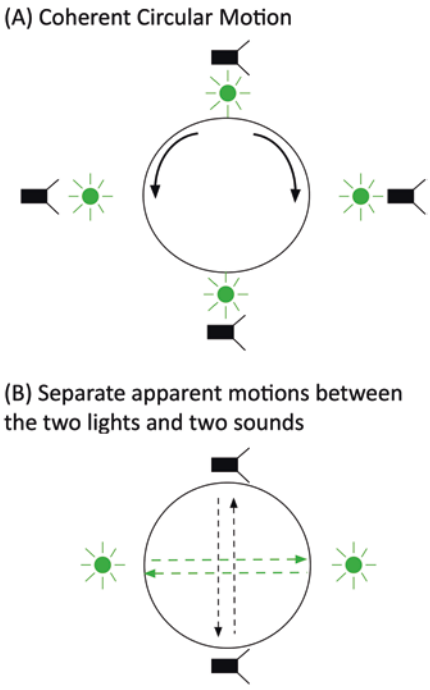

Fig. 2.26 Coherent circular motion is perceived using either the four lights or four tones, and participants can judge the direction of rotation (A). However, the perception of rotation disappears if the participants have to integrate the positions of two lights and two tones. Instead, the tones are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

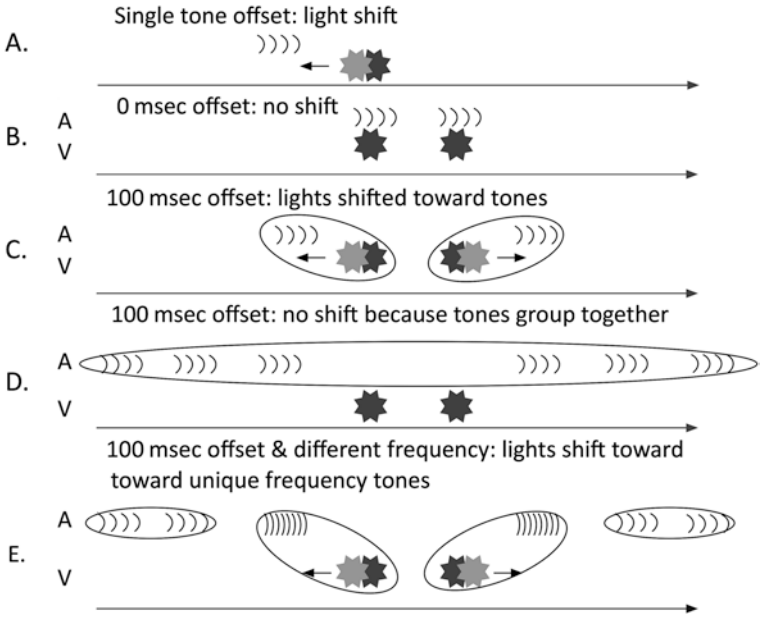

Fig. 2.27 Temporal ventriloquism: Presentation of tones can affect the perceived timing of visual stimuli. The darker stars represent the original timing of the lights; the gray stars and arrows indicate the change in temporal position due to the tone presentation

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

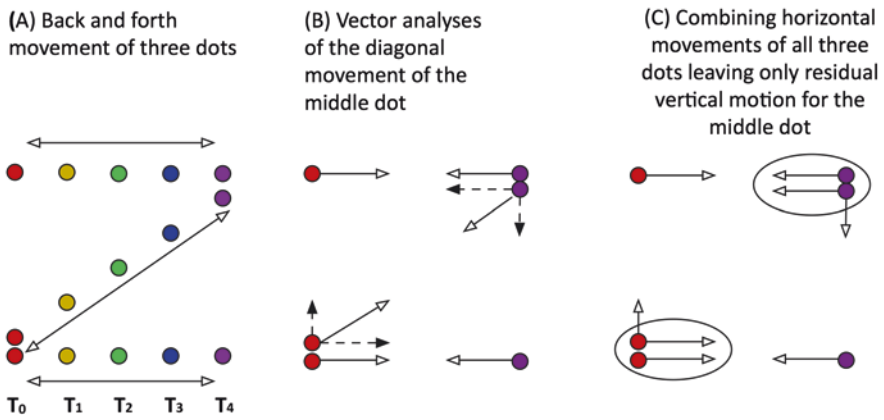

Fig. 2.28 In (A), the movement of each dot is shown at five time points (each colored differently). In (B), the diagonal movement of the middle dot (open arrow) is split into its horizontal and vertical motion components (closed arrows). In (C), the horizontal components are bound to the horizontal motion 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.

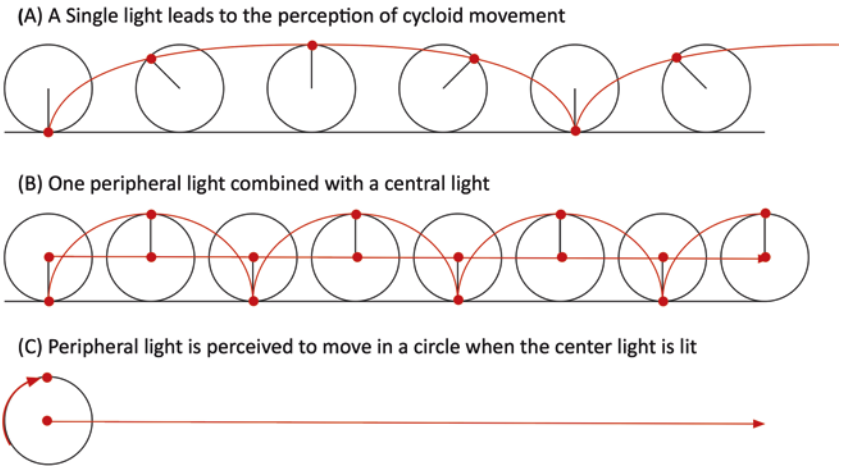

Fig. 2.29 (A) A single light mounted on the periphery of a rolling wheel generates the perception of cycloid motion. (B) If a single central light is added to the peripheral light, viewers report seeing a single light circling around a moving wheel (C). A demonstration is shown in the YouTube video "Motion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

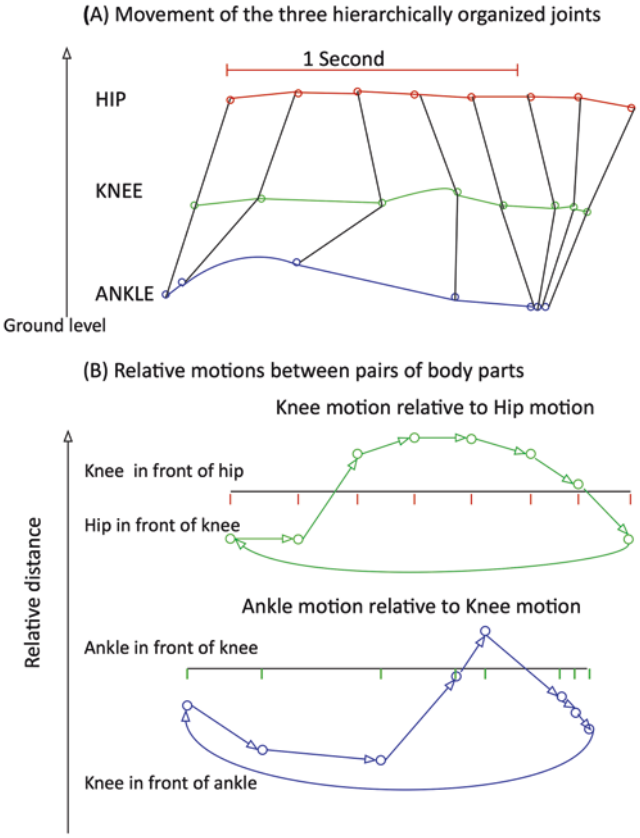

Fig. 2.30 (A) The vertical movement of the hip, knee, and ankle is shown for slightly more than 1 sec. Each motion is an entirely ambiguous, slow vertical oscillation. (B) For each pair, hip to knee and knee to ankle, the relative motion between the lower faster body part (knee and ankle) and the connected 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

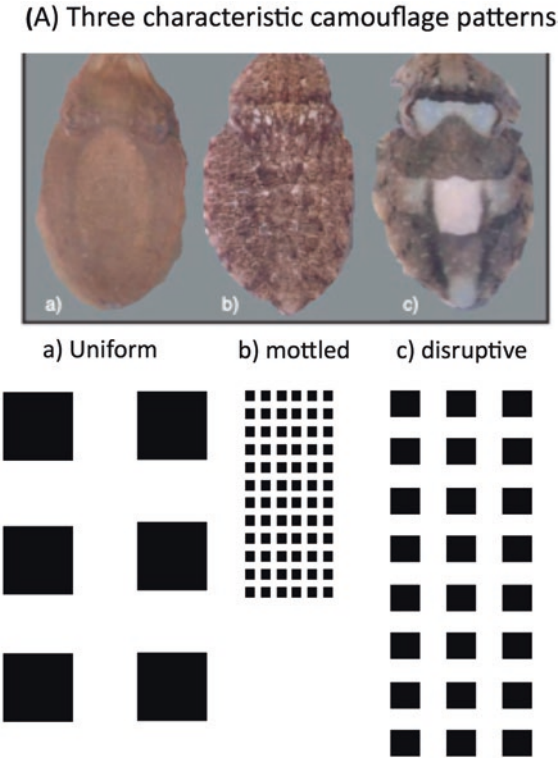

Fig. 2.31 In (A), the three characteristic kinds of camouflage of the cuttlefish are shown. The use of each kind depends on the size of the checker squares of the background. (Adapted from Barbosa et al., 2007)

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

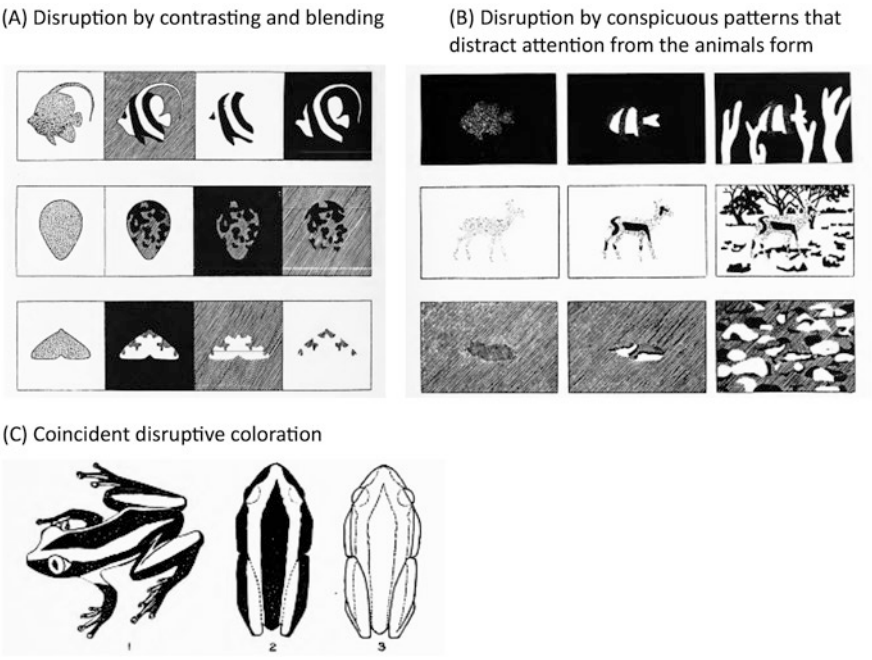

Fig. 2.32 Illustrations from Cott (1940) showing disruptive markings that lead predators to focus on distracting internal patterns and not on body shape (A and B). Coincident disruptive coloration split the frog's body into three separate parts (C). The rightmost drawing shows the frog colorless; the 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.

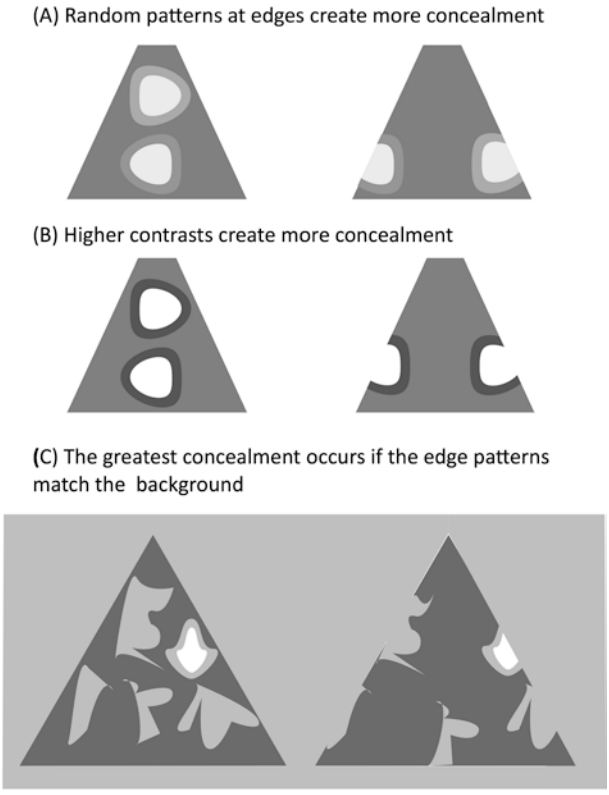

Fig. 2.33 (A) Random patterns at the edges of objects conceal the shape of those objects and bring about a higher survival rate. (B) This advantage is increased when the random patterns are higher contrast such as the "eyespots" on the wings of moths. (C) Concealment is maximized if the edge spots match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

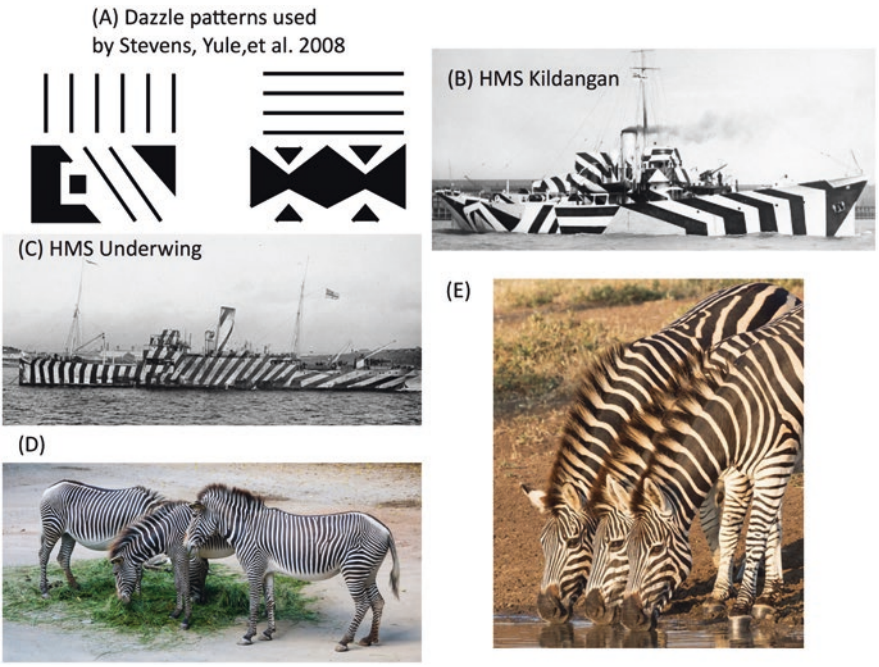

Fig. 2.34 (A) Four different kinds of dazzle markings that resulted in reduced predation. (B & C) Dazzle markings were used extensively during World War I to conceal ships. (B) HMS Kildangan. Photograph Q 43387 in collections of Imperial War Museum (collection no 2500-06); (C) HMS Underwing by Surgeon 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