CHAPTER 5

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

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

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

There are two issues here:

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

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

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

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

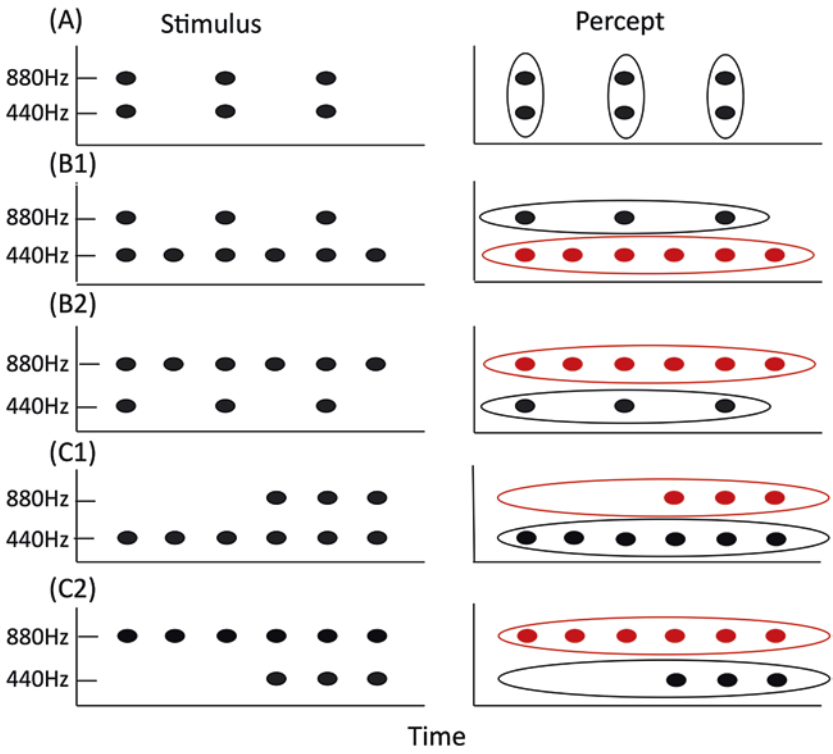

Fig. 5.1 If two sounds occur synchronously, the two sounds are combined as shown in (A). But if the sounds occur at different rates, then the sounds are heard separately (B1 & B2). Moreover, if one sound occurs before the other the sounds are heard separately even after they become synchronous (C1 &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

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

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

5.1 Color and Timbre

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

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

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

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

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

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

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

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

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

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

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

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

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

5.2.1 Ambiguity of Color and Timbre

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

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

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

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

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

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

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

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

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

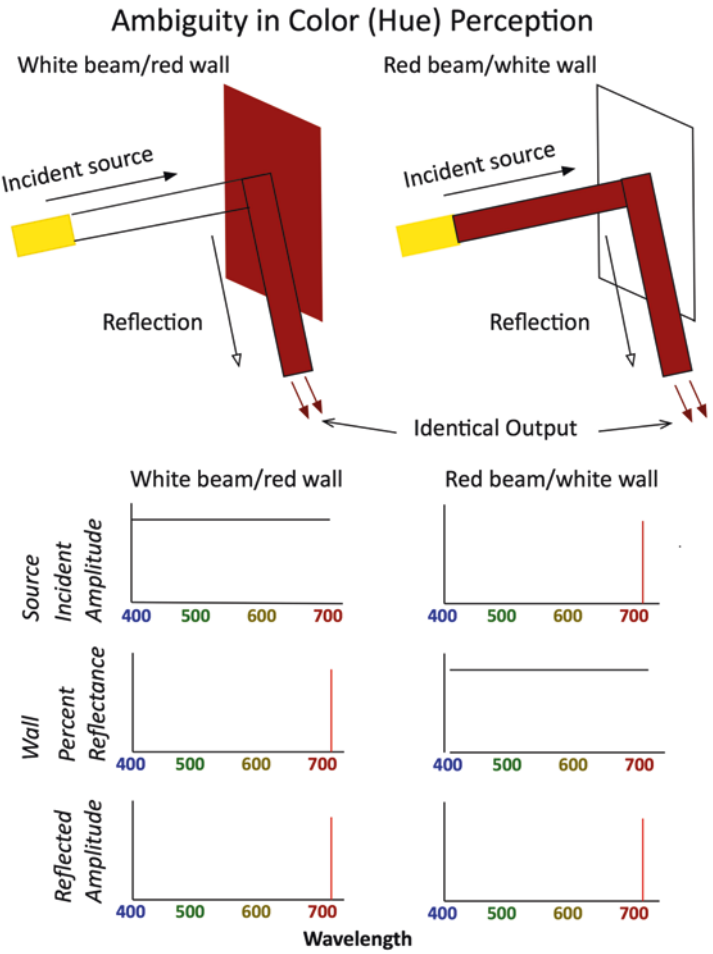

Fig. 5.2 A white beam source reflecting off a red wall yields the same sensation as a red beam reflected off a white wall. For the white beam/red wall, the source energy has equal energy at all wavelengths, but the red wall reflects only the light energy in the red region. The result is beam of red light.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

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

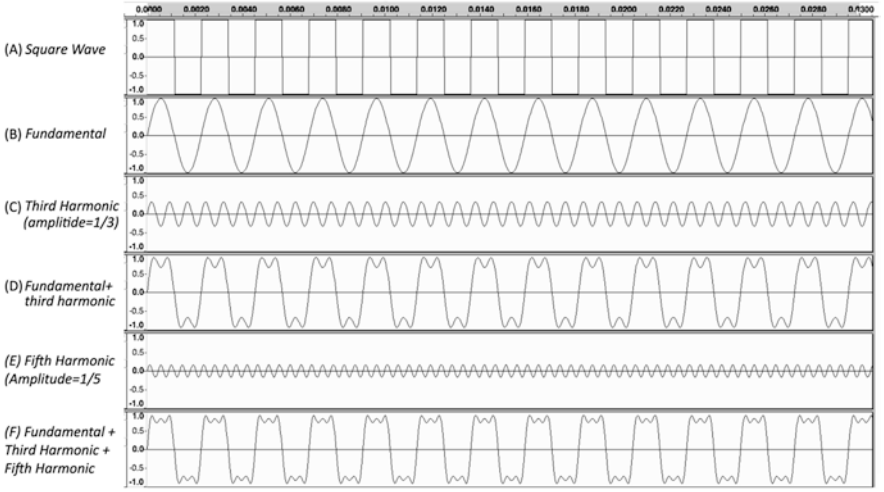

Fig. 5.3 The top panel shows the complete square wave. The second and third panels show the first components: a sine wave at F_0 and $3F_0$ with amplitudes of 1 and 1/3. The fourth illustrates the sum of F_0 and $3F_0$. The fifth panel shows $5F_0$ with amplitude 1/5 and the last panel illustrates the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

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

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

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

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

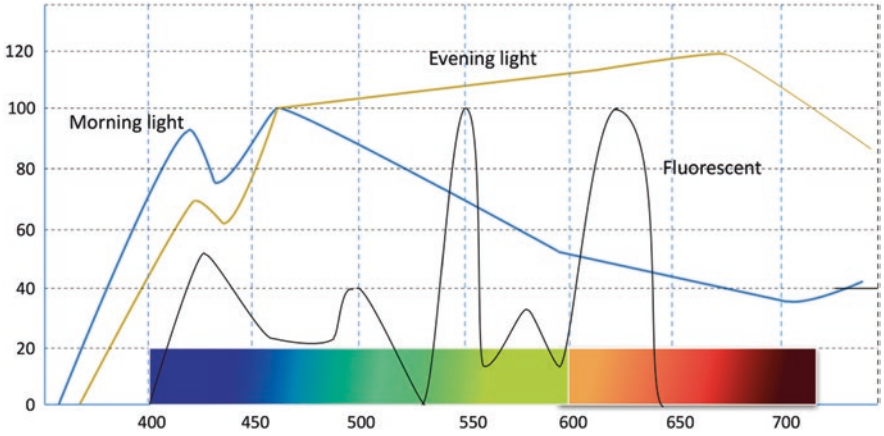

Fig. 5.4 The frequency spectra of morning daylight, evening light, and a typical fluorescent light

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

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

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

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

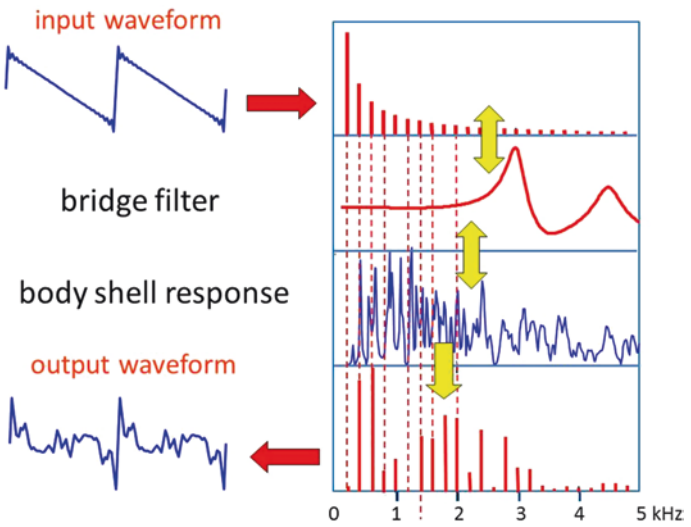

Fig. 5.5 The input waveform is modified by the bridge and body shell resonances to generate the radiated sound. The output is quite different than the input. The vertical dashed lines are the frequencies of the partials from the bowing. (Reprinted from Gough, 2016. Permission by author)

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

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

5.2.2 The General Strategy

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

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

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

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

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

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

5.2.2.1 Color Receptors

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

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

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

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

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

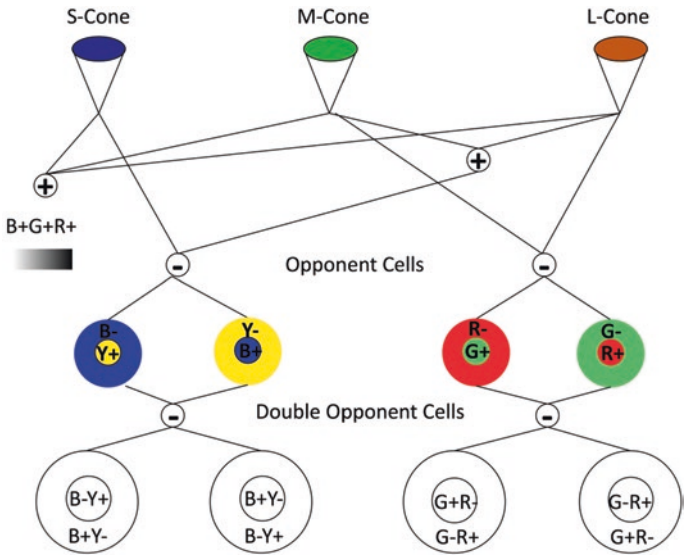

Fig. 5.6 The three cones in the fovea undergo two transformations. The first transformation creates two types of opponent cells, each with two variants, to encode color. In addition, one type of cells, B+R+G, adds the outputs of all three cones to encode lightness. The second transformation maximizes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

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

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

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

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

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

5.2.2.2 Auditory Receptors

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

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

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

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

5.2.2.3 Comparison of Visual and Auditory Receptors

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

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

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

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

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

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

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

5.3 Color Constancy

5.3.1 Reflections

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

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

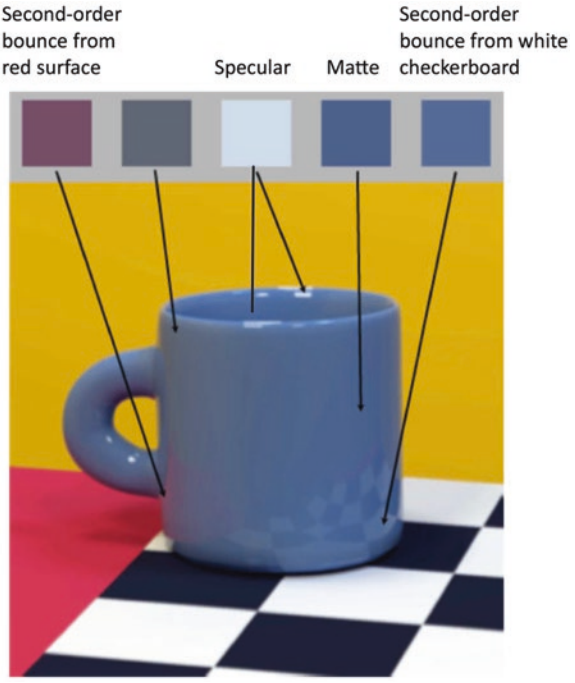

Fig. 5.7 A glossy cup creates several different reflections and yet is perceived as one cup under one illuminant. (Adapted from Brainard & Maloney, 2011. © Association for Research in Vision and Ophthalmology (ARVO))

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

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

5.3.2 *Monge's Demonstrations*

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

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

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

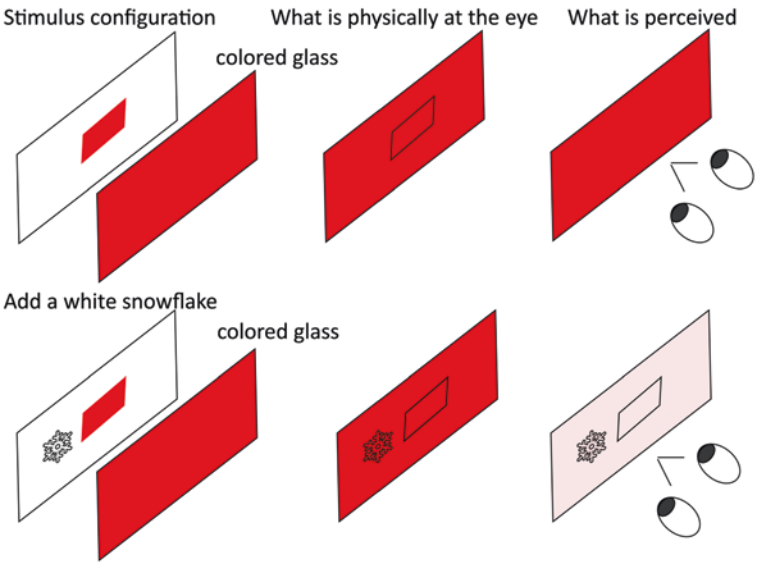

Fig. 5.8 In the first row, all that is seen is the region covered by the red glass and there is no desaturation or bleaching. The entire region looks bright red. In the bottom row, a white snowflake is placed on top of the background. The white snowflake should look red, the same color as the red paper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

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

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

Again, the logic is convoluted:

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

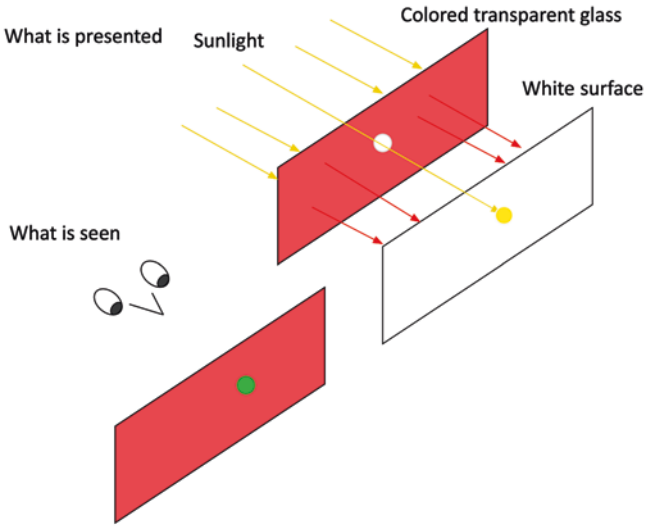

Fig. 5.9 An illustration of Monge's demonstration that context can bring about the illusion of a complementary color in the center of the surface. (The sunlight is shown as yellow for illustrative purposes)

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

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

5.3.3 Asymmetric Matching

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

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

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

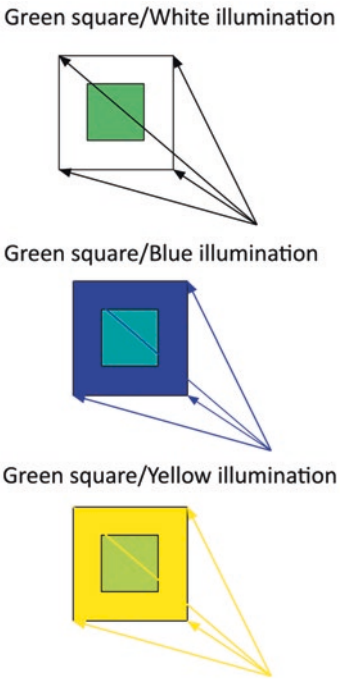

Fig. 5.10 Under a white illuminant, the green square appears green and the white background appears white. Under the blue and yellow illuminant, the background appears blue and yellow, respectively. The green square appears to be color between the green reflectance and the illuminant

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

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

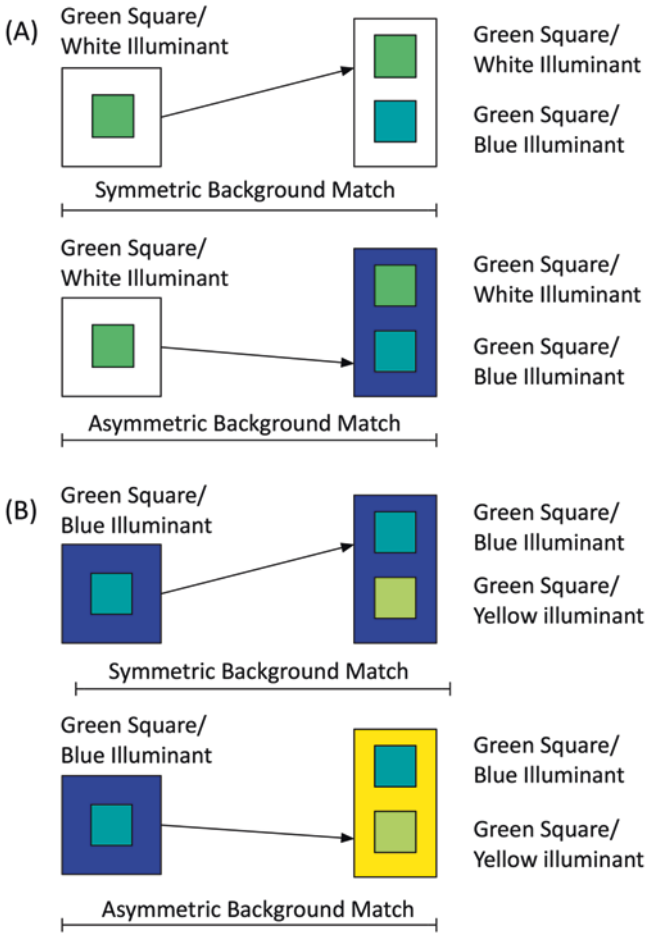

Fig. 5.11 For symmetric matches, the matching color in the test pair looks identical to the original. But for asymmetric matches, the matching color in the test pair does not look like the original, but the one that would occur if the green square were illuminated by a blue or yellow light. The hue of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

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

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

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

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

One final point: We continually adapt to physiological changes. In general, the crystalline lens in our eyes becomes increasingly yellow as we age (primarily due to sunlight exposure), allowing less and less blue light to reach the retina. Nevertheless, we compensate for this yellowing so that white still seems to be white. After cataract surgery that removes the yellowish lens, the visual system, over the space of several months, recalibrates so that white returns to white.

5.3.4 "The Dress"

The picture of a dress originally published on Tumblr in 2015 went "viral" due the surprising split among viewers as to its colors. The photograph was taken on a cold winter day at dusk. In an informal Internet survey, 60 percent of the respondents reported it to be white and gold, 30 percent reported it to be blue and black, and 10 percent blue and gold (Wallisch, 2017). What the split implies is that what you see is not necessarily what others see and not necessarily what a color analysis with a photometer would measure. These outcomes support Monge's demonstrations illustrating that color perception is a second-order calculation based on the light reaching our eyes as well as our interpretation of the illumination of the scene.

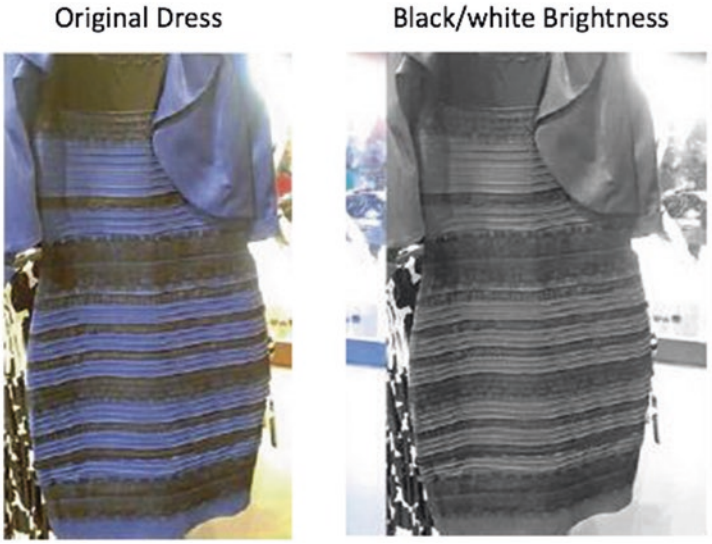

Fig. 5.12 The original colored dress and an achromatic white/black version. (Adapted and reproduced from Gegenfurtner, Bloj, & Toscari, 2015. By Permission, Elsevier)

The original photograph is show on the left in Fig. 5.12. To understand how the above split occurred, start with an achromatic version on the right. Following the logic used to explain Monge's demonstrations, this image could have come about in two ways. If the light source was in front, then the brightness of the image was probably due to a bright illuminant on a very dark dress. If the light source was in back so that the dress blocked most of the illumination, the brightness was probably due to a weak illuminant falling on light dress. Depending on the perceiver's guess about the position of the illuminant, overall the dress would look dark or light, respectively.

Using the same logic, the observer must deduce the "true" color of the dress by "neutralizing" the effect of the illumination. The apparent illumination seems to change from daylight at the top of the photo to shadows at the bottom. If one believed, on the one hand, that a "cool" blue light (i.e., morning sunlight) illuminated the dress, then any blueness in the photograph of the dress would be attributed to that light; neutralizing that illumination give rise to dress materials that appear to be white and gold. On the other hand, if one believed that "warm" yellowish light (i.e., afternoon sunlight) illuminated the dress, then neutralizing the yellowish illumination would give rise to dress materials that seem to be blue and black. (If the dress seemed to be illuminated artificially by a warm incandescent light, the dress would tend to look blue and black). These alternatives are illustrated in Fig. 5.13

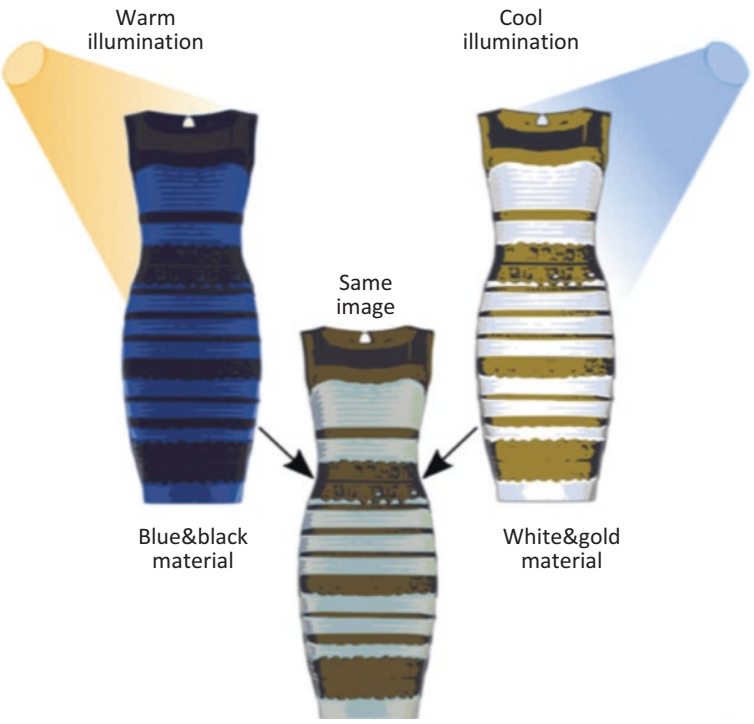

Fig. 5.13 The same image, in the center, could be due to blue/black dress illuminated by afternoon warm sunlight or a white/gold dress illuminated by cool morning sunlight. Different people, unconsciously compensating for their beliefs about the illumination will end up with a different percept. (Reproduced from Brainard & Hurlbert, 2015. By Permission, Elsevier)

To test this interpretation, several investigators have clipped copies of the dress and created scenes that seem to reflect different illuminations. For example, Witzel, Racey, and O'Reagan (2017) constructed one scene in which the light appears shadowed behind the model which overrepresents blue light and one in which the light appears to be shining directly on the model which overrepresents yellow light. When the light casts a bluish shadow, the dress looks lighter with gold stripes, but when the light is direct, dark blue with brown stripes. Furthermore, Wallisch (2017), based on an Internet survey, found that participants who believed that the dress was in shadows (i.e., bluish) were 20 percent to 40 percent more likely to describe the dress as white/gold than participants who believed the dress to be directly illuminated (75 percent to 45 percent).

The Bayesian explanation would be that the belief in one sort of illumination is based on previous experience, the "priors" in Bayesian terms. This has led to the hypothesis that "early risers," who more frequently experience bluish morning light, are more likely to see the dress as white/gold. In contrast, "late night birds" who more frequently experience the yellowish evening light are more likely to see the dress as blue/black. The evidence for this is weak, but I am certain that people have different priors that affect their perception of color of the dress (Wallisch, 2017).

In sum, the "dress" is another example that the light rays at the eyes, like sound waves at the ears, are inherently ambiguous. Our attempts to interpret them can lead to dramatically different beliefs about the distal objects. It is interesting to note that it is extremely difficult for people to shift from one color scheme to the other. This is not the case for multistable objects discussed in Chap. 3 where the shape shifts are inevitable and continuous. When multistable figures shift, the parts change their meaning. A curve that is part of a nose in one shape becomes part of an ear in another shape. But, for the "dress," it is always a dress and the stripes remain the same.

5.3.5 Does the Color of Objects Matter for Recognition?

Color, shape, motion, and other surface characteristics such as shading are the properties we use to break up the external scene into discrete objects. As described previously, infants are quite sensitive to color differences, and I suspect that stationary or moving colored blobs primarily segment the visual world for them. Adults seem to respond to shape more than color, but both affect the organization. Colored objects in the visual periphery aided localization, while colored objects in central vision aided object recognition (Nathmann & Malcolm, 2016)

The fundamental question is whether objects are first recognized by shape and then colored in, or whether shape and color are used conjointly to recognize objects. This question has led to experiments to investigate whether objects that are correctly colored are recognized faster than the same object if it is achromatic or even if colored incorrectly. For example, is a yellow banana more easily detected than a purple one?

The key distinction is the "diagnosticity" of the color for the object. A color is diagnostic for an object if we consistently identify it with that real-world object such as a banana; it is not diagnostic if the real world object could appear in any color, like a necktie. Obviously there are intermediate cases in which an object could occur in a limited set of colors. In many studies, diagnostic cues enhance object recognition, particularly when participants have to name the object (Bramao, Reis, Peterson, & Faisca, 2011). A gray seal, a yellow sun, a red stop sign, or a white snowflake is easier to recognize than a pink seal, a green sun, blue snowflake, or green stop sign. But, a gray car or sweatshirt is not easier to identify than a blue car or sweatshirt.

5.4 Countershading Camouflage

Another example of source-filter processes is countershading. Thayer, an American painter using decoys, was the first to illustrate how countershading could be used to avoid detection as dramatically shown in Fig. 5.14E. Our normal expectation is that sunlight comes from above (Fig. 5.14A). Hence, flat, horizontal, two-dimensional objects will reflect equal amounts of light at all points on their surface. Solid three-dimensional objects will self-shadow due to their shape. If the object is convex and bulges out, it will be brighter along the top and darker along the bottom (Fig. 5.14A) so that the perception becomes three-dimensional making the animal easier to detect. The brightness of a surface is the brightness of the incident light reaching the surface multiplied by the reflectance of the surface. On this basis, if the animal is colored so that its base is more reflective, that is, whiter, than its top surface (Fig. 5.14B), then the brightness of the animal's surface can be equalized (Fig. 5.14C). The three-dimensionality of the animal can be minimized by a gradation of shadow. "The animal now looks flat and Insubstantial" (Thayer, 1909). It is interesting that countershading is reversed for animals that hang from branches; the underside is darker. Rowland (2009) provides a review of the effectiveness of countershading.

5.5 Timbre

5.5.1 Source, Filter, and Resonance

The color section emphasized the problem of setting aside the effect of illumination to derive the "true" color of an object, the true color being what would be seen in the roughly uniform spectrum of daylight. This possibility was based on the assumption that object's reflectance spectrum is the same across all possible illuminations and does not change over time. If we gradually shifted the source illumination from blue to red, differences in the light reaching the eye would be due to the source, not changes in the reflectance of the object. If we could discount the variation caused by the source, we could picture the true object color.

This does not work for timbre because the resonances of the source and the resonances of the sound body filter occur at discrete frequencies. Suppose the sound body has resonances at 100 Hz, 400 Hz, and 700 Hz. As the source frequency is increased, it will induce the sound body to vibrate at those three frequencies with different movement patterns due to the material and shape of the object. (Between those frequencies the sound body vibrations will be very weak, if they exist at all). Moreover, the resonances of the filter change as a function of the way the source energy is applied. Vibrating a violin body at different position, intensities, and/or durations will change the frequencies of the resonances and thereby change the sound. Finally, all sounds occur in time so that the loudness and frequency change over time. Some source and sound

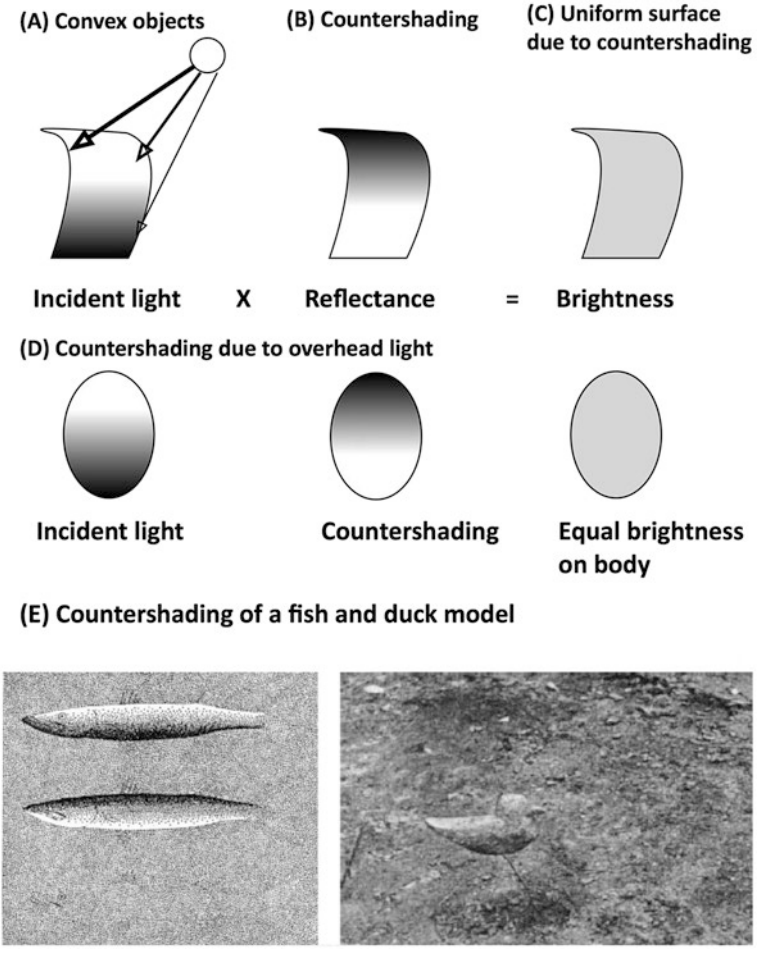

Fig. 5.14 (A) Sunlight reaching the top of a convex object will be stronger than that reaching the bottom (shown by the thickness of the arrows). (B) To compensate by countershading, the lower regions are more reflective than the top regions. Gradually changing the reflectance is most effective. (C) The energy of the incident light multiplied by the reflectance (i.e., the percentage of reflected light) yields the light reaching a predator. (D) Animals are often countershaded so that the top surface is darker than the underbelly. (E) An illustration by Cott (1940, page 37) and a photograph by Thayer (1909) demonstrate the effectiveness of countershading using a fish and duck model. For the fish, the overhead illumination (topmost drawing) is neutralized by the dark shading on the back of the fish (middle drawing). There really is a third fish at the bottom in the left panel and there really is a second duck model to the right in the right photo. (Thayer, 1909, Chapter 2, Page 24, Figure 4)

body resonances reach their maximum quickly and then decay rapidly while others reach their maximum and decay slowly. There is no "true" invariant timbre of an object.

For all of these reasons, descriptions of timbre are often ambiguous based on multiple properties of the frequency spectrum and temporal characteristics of the sound**.** When we hear a sound, we tend to identify the object and not its sound qualities. This contrasts with visual objects, which we often describe in terms of size, color, and shape. Nonetheless, it is those frequency and temporal characteristics that allow us to identify the object although there may not be a fixed set that work in real environments with overlapping sounds. A realistic goal may be to list a set of qualities and attempt to determine how they are used at different times.

To unravel timbre, we need to concentrate on the connections between the excitation source and filter that determine the evolution of the output spectrum over time. To do this we need to understand the coupling between the source and filter.

If we strike, vibrate, bend, twist, or blow across an object, the material may begin to vibrate at one or more of its resonance frequencies, termed vibration modes. Each vibration mode will have a distinct movement and can be characterized by the resonant frequency at which the motion is maximized and by its damping or quality factor. The material must possess enough stiffness or spring-like restoring property so that it snaps back and overshoots the initial position, for example, a violin or guitar string. The restoring force of the overshoot acts to reverse the original motion but the overshoot also overshoots and so on. The material begins to vibrate continuously in simple harmonic motion, like a violin string (see Fig. 5.3B for a picture of simple harmonic motion). To continue the resonant vibration, energy must be applied *in-phase* with the vibration. The motion of the excitation must match that of the material. If the excitation stops, the vibration of the material eventually dies out due to friction. It is important to note that the vibration frequency of the mode is that of the excitation even if that frequency is not the resonance frequency of the mode.

The damping or quality factor controls the range of frequencies that can excite the mode and the rate at which the vibration mode begins to vibrate and to die out. For a highly damped mode with a low-quality factor, a wide range of frequencies can excite the vibration mode though the amplitude of the vibration is small and roughly equal across that range. The amplitude builds up to its maximum quickly, it increases and decreases in amplitude in synchrony with the excitation, and once the excitation is removed it dies out quickly. Conversely, for a lightly damped mode with a high-quality factor only a narrow range of frequencies can excite the vibration mode but within that range the amplitude may be high. The vibration reaches its maximum slowly, lags behind changes in the source excitation, and dies away slowly when the excitation stops. For a highly damped mode, the excitation can reach two-thirds of its maximum in one cycle and lose two-thirds of its amplitude in one cycle. In contrast, for a lightly damped mode, it might take 10 cycles to reach the same values.

The common example of a damped system is a swing. If the "pusher" is tightly coupled to the swing by means of a stick connected to the seat the movement will be the same as the arm motion; the amplitude of the swing will be identical at all arm frequencies and the movement will begin and end in synchrony with the arm motion. It is highly damped. In contrast, if the swing is connected to the arm motion only by a weak spring, the swing can reach a great height if the excitation occurs at the resonance frequency by pushing at the same high point in the motion, in phase. But, the swing will reach its maximum only after an extended period of pushing and die out after an extended period of not pushing. It is lightly damped. The goal for a hi-fi speaker is a flat frequency response, which requires all frequencies would be reproduced accurately with no delay in responding to the excitation. This would require a highly damped system. The disadvantage is the output will be weak, but a high-powered amplifier can overcome that.

Based on these differences, we would expect that the vibration modes of a violin would be lightly damped in order to play loudly. The wood top and bottom plates would be thin and free to vibrate to yield the maximum volume. If there were just a small number of modes at varying frequencies, then those musical notes that match the frequency of the vibration modes would be full volume, but the other notes would be muted. Fortunately, the complex wooden structure has many modes with adjacent frequency peaks, which evens out the frequency response.

It is the interaction between the excitation, source, and filter that makes the analysis of timbre so difficult. Every source vibration is composed of multiple frequencies, and every sound body filter has multiple vibration modes with different degrees of damping. Due to the combination of the various levels of damping, the timbre of the sound will change from onset, to steady state to offset as different frequencies reach their maximum and decay at different times. Each excitation will excite different source and filter modes, so that we should not expect to find a single acoustic property that can characterize an instrument, voice, or environmental event. The perceptual problem is therefore identical to that for color, where the problem is to recognize the same color under different illuminations; for timbre the problem is to recognize the same auditory objects and events under different excitations, frequencies, and amplitudes. We need transformations that can band together colors, sounds, objects, or events in different contexts.

5.5.2 Timbre of Instruments

Gaver (1993) has proposed organizing auditory events into three physical actions: (1) vibrations due to scraping, hitting, or plucking, which would include percussion and stringed instruments, breaking and bouncing objects, and footfalls; (2) aerodynamic sounds due to continuous excitation, which would include blown instruments, wind noise, and mechanical objects like jet engines; (3) liquid sounds due to dripping, splashing, or boiling.

Across a wide variety of experiments two factors have emerged as critical to the perception of timbre and ultimately to the perception of objects and events. The first is temporal characteristics, either the duration of the onset or attack of individual sounds or the timing between individual sounds (e.g., rattles, faucet drips). The second is the energy distribution of the frequency components. There are other factors, but they tend to vary from experiment to experiment.

We start with an experiment that investigates the timbre of instruments (McAdams, Winsberg, Donnadieu, De Soete, & Krimphoff, 1995). These instruments cut across Gaver's categorization: stretched strings (e.g., violin, piano), stretched membranes (snare drums, tympani), rigid materials that vibrate without tension (cymbals, xylophone) as well as aerodynamic sounds (woodwinds and brass instruments). There were 12 electronic instruments that simulate real ones and six electronic instruments that were hybrids of the real instruments. The participants were asked to judge the similarity between each possible pair of instruments. The instructions were made purposively vague to avoid influencing the judgments.

A statistical technique termed *multidimensional scaling* was used to place the instruments in a geometric space such that the distances between the instruments in the space were proportional to their judged similarities. A simple example of this process would occur with three sounds such that the similarities between A&B, B&C, and A&C were 1,1,2. In this case we could place the three sounds on a one-dimensional straight line A-B-C. If the judgments instead were 3,3,3 then the three sounds could be placed as an equilateral triangle in a two-dimensional space. In a more complicated example, if there were four sounds and the similarity judgments between each pair were identical, then the four sounds would form a pyramid in a three-dimensional space. The difficulty, of course, is to figure out what acoustic property each dimension represents. Given each instrument's position in the two- or three-dimensional space, the researchers correlate the positions on each dimension to their acoustic properties. It is not enough to say that all the woodwind instruments are on one side and brass instruments on the other side of a dimension. The goal is to identify the acoustic variables that underlie that separation and that could result from the mechanical properties of the instruments.

The distances between the instruments are mainly due to differences on two dimensions, as illustrated in Fig. 5.15. The position of the instruments along the first dimension correlates with the attack or rise time of the sounds. Plucked or struck instruments at one end of the dimension were characterized by short rise times, while wind instruments at the other end tend to have slow rise times. Participants are not actually judging the rise time; rather they are judging the perceptual consequences. Struck or plucked instruments are likely to have an initial noisy impact sound made up of a wide band of frequencies. All the vibration modes of the sound body start at the same time. In contrast, blown instruments are likely to have a noisy initiation that evolves into a stable blowing vibration. The damped vibration modes reach their maximum at different times to create an evolving sound.

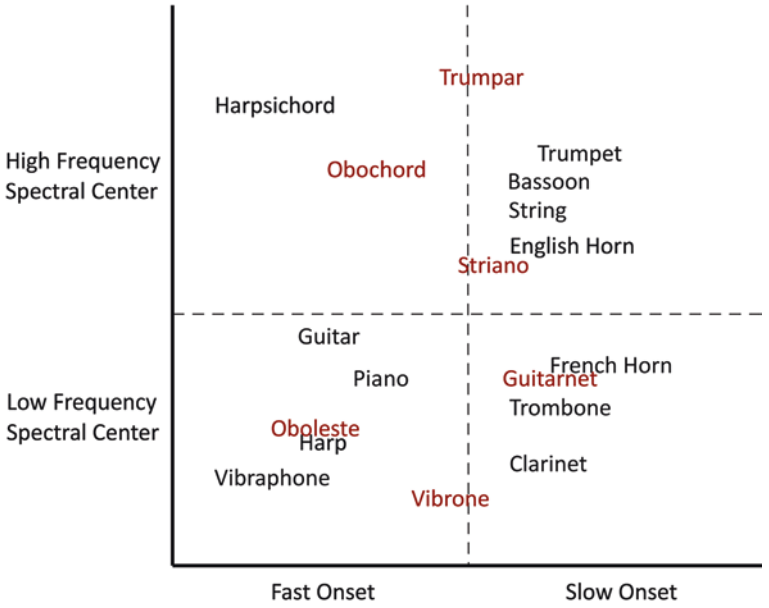

Fig. 5.15 The two-dimensional representation of the similarity judgments among pairs of instruments. The x-dimension represents differences in the onset speed of the sounds. The y-dimension represents differences in the spectral center of the sounds. Real instruments are black, hybrid instruments are red. (Adapted from McAdams et al., 1995)

Sound Files 5.15: Real and hybrid instrumental sounds shown in Fig. 5.15

The position of the sounds on the second dimension is correlated to the spectral centroid of the sounds, which is the frequency of the average energy. Suppose we had three harmonics at 100 Hz, 500 Hz, and 900 Hz; if the amplitude of the three harmonics were 2,3,1 then the centroid would be [2(100) + 3(500) + 1 (900)]/6 = 433 Hz and if the amplitudes were 1,1,4 then the centroid would be [1(100) + 1(500) + 4(900)]/6 = 700 Hz. Roughly, it is a measure of perceptual brightness or brilliance, the amount of energy at the higher frequencies.

In sum, the similarity judgments among musical instruments can be attributed to difference in the temporal and spectral characteristics of the radiated acoustic wave. Moreover, the spatial configurations tend to group the instruments into families based on their physical characteristics (Giordano & McAdams, 2010). Blown instruments are separated from impulsive ones.

The easiest way to illustrate that the timbre of a sound is caused from the joint action of timing and the spectrum is to reverse the sound and/or shape the amplitude of the harmonics, that is, change the spectral center. If we reverse the sound, the spectrum is identical, but the offset becomes the onset and vice versa. In Fig. 5.16A, the sound of a vibraphone, in reverse, sounds nothing like the original. We can also slow the onset, keeping the spectrum identical, and this also destroys the vibraphone sound.

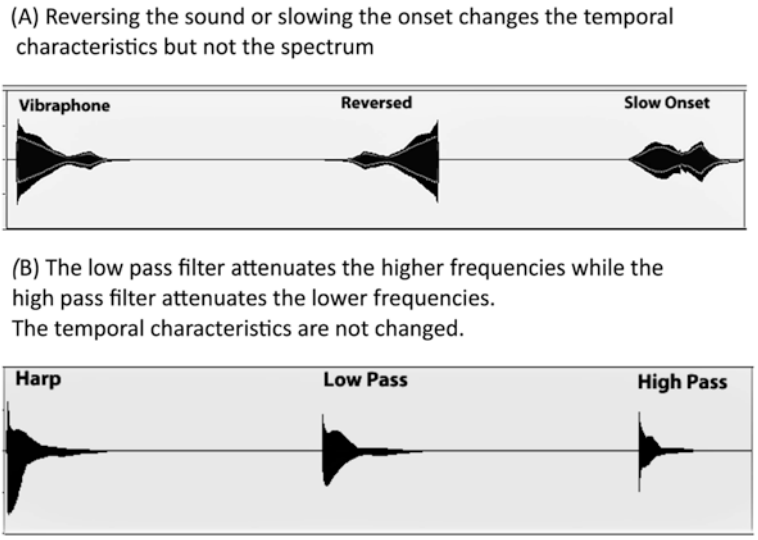

Fig. 5.16 (A) Waveforms of the original vibraphone sound, the reversed vibraphone, and the slow onset vibraphone. For all three sounds, the spectrum and spectral center are identical. (B) Waveforms of the original harp sound, the low-pass waveform, and high-pass waveform. The onset and duration are identical for all three sounds

Sound Files 5.16: Temporal and spectral changes that affect timbre illustrated in Fig. 5.16

We can also alter the spectrum (keeping the onset constant) shown in Fig 5.16B. For the harp, the amplitude peak occurs at the fundamental frequency of roughly 300 Hz with weaker peaks at the harmonics of 600 Hz, 900 Hz, and 1200 Hz as well as significant peaks at the harmonics out to 7000 Hz. The low-pass filter strongly reduces all peaks except for the 300 Hz fundamental and beyond 1000 Hz all disappear. Here the harp sounds hollow. The high-pass filter attenuates the amplitudes of the 300 Hz, 600 Hz, and 900 Hz peaks but does not affect the higher peaks. The harp now sounds tangy.

Nearly all the research of timbre is purely acoustical. Participants judge the similarity between two tones, judge individual attributes of the sound, or try to identify the instrument. What this misses is the feedback from the instrument to the musician and how that affects the perceived quality of the instrument. While playing, the actions of the musician create two sources of feedback. First is the acoustic feedback, the sound of the notes, due to the resonances of the instrument body. Second, but equally important, is the mechanical feedback due to the structure of the instrument that create vibrations on the fingers and hands holding the instrument or create contact forces if striking an instrument. Instruments have mass and are springy and their feedback to the roughly equal mass and springiness of the hand and arm determines how well musicians can adjust their playing technique to achieve an expressive acoustical goal. Playing is multisensory employing acoustical and haptic feedback (O'Modhrain & Gillespie, 2018).

Saitis, Järveläinen, and Fritz (2018) have reviewed research that makes clear that musicians judge the quality of violins based both on the acoustic qualities of the notes and the vibrotactile feedback. Three factors are important: (a) the strength of the felt vibrations by the left hand on the violin neck, by the shoulder, and jaw on the chin rest augment the perception of loudness and richness, conceptualized as *resonance* or *feel*; (b) the reactive force felt by speed and effort, felt through the bow by the right hand, conceptualized as *playability*; (c) the consistency of the vibratory feedback across the playing range, conceptualized as *balance*. The vibratory feedback both helps musicians control their playing style and expressiveness, but also contributes to their perception of the sound itself.

5.5.3 Timbre of Physical Actions

These results suggest that temporal and spectral characteristics of vibration sounds that are useful for analyzing the single sounds due to scraping, hitting, or plucking of percussion and stringed instruments could be extended to other kinds of acoustic events. Specifically, Gaver's analysis described above suggests that series of vibration sounds including objects breaking and bouncing, and footfalls could be understood in terms of the spectral characteristics of the individual sounds as well as the timing among those sounds, that is, the rhythm of the sounds. (Remember that the sounds of footsteps, even if they were not synchronous with the visual motion of the foot, influenced the perception of movement using lighted-dot figures in Chap. 2).

Hjoetkjaer and McAdams (2016) have investigated the perception of three types of materials, wood, metal, and glass, undergoing three types of actions, drop, strike, and rattle. If we isolate one type of action, say striking, then the difference between the materials would be limited to their resonant spectra. In similar fashion, if we isolate one type of material, say wood, then the differences between actions would be limited to the temporal pattern. The yields two interrelated questions: (1) when participants judge the similarity among the nine stimuli, are the two factors treated independently or are the stimuli perceived as a gestalt; (2) if all the materials are transformed to have identical spectra can participants identify the action, and conversely, if all materials undergo the same action can participants identify the material?

Figure 5.17 illustrates simplified temporal patterning for striking, dropping, and rattling. When these materials are struck, the initial sound burst decays due to its internal damping. If the material is dropped, the initial impact sound (that resembles the strike sound) is followed by two or three sounds due to bouncing. We can distinguish bouncing with its regular impacts, from breaking where multiple parts bounce independently and at random (Warren & Verbrugge, 1984). If the material is rattled, the sound is composed of sound bursts that vary in amplitude and timing.

Figure 5.17 also shows simplified spectra of the materials used. The spectra of metal and glass increase in amplitude from low to middle frequencies and then have somewhat random peaks at higher frequencies. In contrast, the spectrum of wood is flat at the lower frequencies and then gradually decreases at the

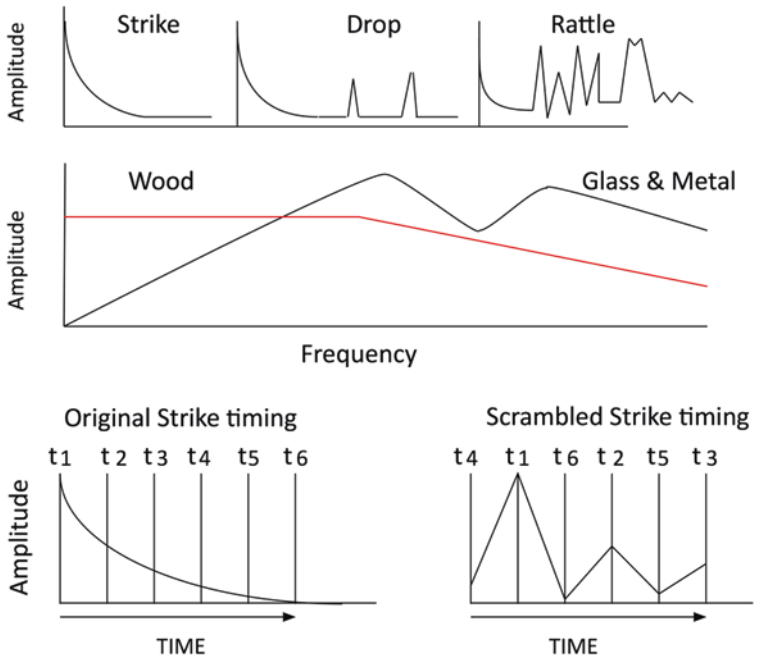

Fig. 5.17 Simplified representations of the temporal patterning of the three actions and the spectra of the three materials used by Hjoetkjaer and McAdams (2016). In the second experiment, the amplitudes of the strikes were scrambled across time to determine the effect of the temporal patterning, and a simplified version of the scrambled amplitudes is shown in the figure

Sound Files 5.17: The nine original sounds from three actions and three materials

higher frequencies. This suggests that while glass and metal are likely to be confused, they will be readily differentiated from wood.

In the initial experiment, participants judged the similarity among the nine types of stimuli. The results indicated that the similarity among the stimuli could be represented in two dimensions, as shown in Fig. 5.18. Wood was distinct from glass and metal probably due to the decline in high-frequency energy; striking tended to be distinct from dropping and rattling probably due to the absence of secondary sound bursts.

In a second experiment, Hjoetkjaer and McAdams (2016) attempted to assess the importance of the material and action separately by zeroing one out. To eliminate the effect of the spectrum, a constant broadband noise was presented using each of the three temporal patterns underlying the different actions. To eliminate the effect of the temporal pattern, the original pattern for each action was broken into time blocks that were then scrambled. The spectrum for each material was maintained. This process is illustrated in Fig. 5.17 for striking. Scrambling produces a pattern that resembles rattling.

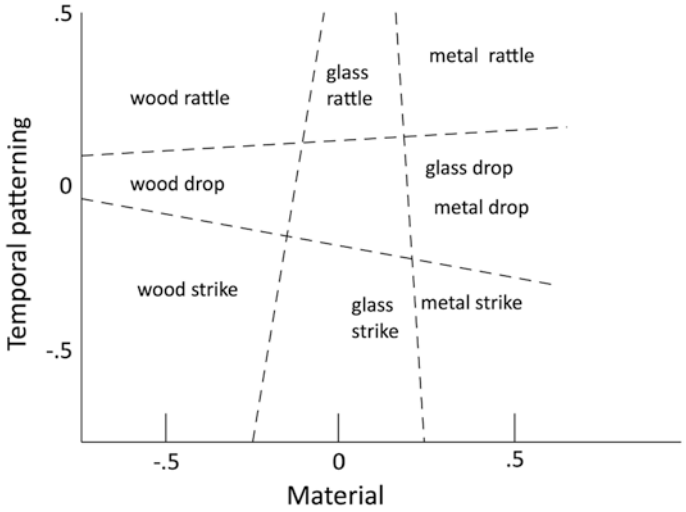

Fig. 5.18 The spatial representation of the similarity judgments among the nine sounds. These sounds roughly present the material and actions as independent factors supporting the results for instrumental sounds shown in Fig. 5.16. (Adapted from Hjoetkjaer & McAdams, 2016)

Sound Files 5.18: Original and modified wood drop and wood strike sounds

For the original sounds, the identification was nearly perfect for either the material (90 percent) or action (96 percent); any errors confused glass and metal. For the conditions that eliminated the spectral information, participants were unable to identify the material though roughly three-quarters of them judged the material to be metal, possibly due to the presence of a high-frequency energy "sizzled" sound**.** For the type of action, drops were nearly always identified correctly (88 percent), but strikes also were usually identified as drops (38 percent strikes to 60 percent drops) because the noise seems to eliminate the sharp onset created by the impact. Rattles were less affected, but somewhat likely to be judged as drops (68 percent rattles to 30 percent drops). In similar fashion, for the conditions that scrambled the temporal pattern, fully 96 percent of the responses identified the action as rattling, probably because the scrambled action had multiple onsets that resembled actual rattling. The percentage correct for material was very good (81 percent), although not quite equal to that for the original stimuli. As found for the original stimuli, wood was identified best, while metal and glass were confused about 25 percent of time.

5.5.4 Timbre of Environmental Sounds

Guyot et al. (2017) investigated Gaver's third category, the perception of liquid sounds. Water by itself does not make hardly any sounds; it is the popping of the air molecules entrained in the water that produces the sound. Smaller bubbles create higher-pitch sounds and aficionados think they can judge the

qualities of sparkling wines by the bubble sounds. The participants' task was to judge the sounds in terms of the physical event that caused the sound, not in terms of the actual sound itself. For example, water spray onto glass, water spray in a full tub, and splatty rain on a roof would all be classified as "jet" sounds even though they sound different. For our purposes, the most important finding was that the timing of the water sounds was a dominant feature. Short repetition sounds (i.e., drips) formed one category, while longer continuous sounds (boiling water) formed another category. The spectral characteristics did not distinguish among the sounds probably because the participants were instructed to judge on the basis of the physical causes of the sounds. As an aside, Katz (1925) has described "film" tactual perception of liquid or airflow as lacking an impression of an object, spatial orientation, or substance. The impression is simply that of immersion.

We can summarize these results in two ways. First, these outcomes illustrate how timbre is based on the evolution of spectral information over time. For struck instruments and events, the rapid, wide-frequency source impact onset causes the sound body to vibrate at its resonant frequencies, and these resonances die away at different rates. For bowed string and wind instruments, a noisy initiation period precedes a stable vibration, that is, after the bow fully engages the string. In turn, those source frequencies result in the vibrations of the sound body at the resonant frequencies. The onset and amplitude of the resonances are determined by the degree of coupling between the source and filter. For both kinds of instruments, the relative amplitudes of the sound body resonances yield the sense of brightness that has been correlated to the calculation of the spectral center.

Second, we can generalize the outcomes for single sounds to sequences. The spectrum and timing within each sound characterize the material and the timing between sounds yields the type of action. This is true for solids as well as liquids. Jumbling either the temporal or spectral properties kills the ability to identify those sounds.

Furthermore, the evolution of spectral information seems to be critical information for the identification for all sorts of sounds, although the relative importance of each possible cue depends on the sound and task. Ogg, Sleve, and Idsardi (2017) compared the categorical identification of musical instruments, speech, and human-environmental events (e.g., keys jangling, deforming newspaper) and found that listeners used temporal, spectral, noisiness, and spectral variability in various degrees to categorize this wide variety of sounds.

The descriptors for timbre often are cross-modal. Timbre is described in visual terms: sounds with higher spectral centers are bright, clear, active, or colorful while sounds with lower spectral centers are dark, dull, wooden, or colorless. But, timbre also is described in textural terms: rough or sharp versus smooth or delicate, warm versus cold, or heavy versus light. Thus, several experiments have compared the perception of the same object properties through visual, auditory, and tactual presentation individually and through multisensory presentation. Fujisaki, Soda, Motoyoshi, Komatsu, and Nishida (2014) paired the visual presentation of six materials (e.g., glass, ceramic, metal, stone, wood, and bark) with the sounds of eight materials (e.g., glass, ceramic, metal, stone,

wood, vegetable (pepper), plastic, and paper) being tapped with a wooden mallet. In some instances the visual and auditory presentations were congruent: visual presentation of a piece of glass with the sound of the glass being tapped. In other instances the presentations were incongruent: visual presentation of a piece of glass with the sound of a piece of wood being tapped. The results for the congruent presentations were straightforward; participants rated the correct material category highest. If the presentation was incongruent, for example, visual glass with auditory wood, participants compromised and chose the material that may have been a second or third choice for each modality, but was plausible for the individual visual glass and auditory wood presentations. Specifically, visual glass was sometimes categorized as plastic. Similarly, auditory wood was sometimes categorized as plastic. When presented together, the preferred description was therefore plastic. If participants were asked to judge the visual properties (e.g., uniform surface versus rough surface, or opaque versus transparent) or the auditory properties (e.g., dampened versus ringing sound or low pitch versus high pitch) for congruent and incongruent pairings they simply judged according to the relevant modality. If asked to judge other properties such as cold versus warm or hollow versus solid that were not modality specific, participants tended to average the judgments from each modality. The authors argue that these results can be understood in terms of Bayesian outcomes, in which the most reliable modality for the task is most heavily weighted.

The sensations from viewing wood samples, listening to the sound after tapping the samples, or running a finger along those surfaces is obviously different. To test whether the perceptions are equivalent and independent of the modality, Fujisaki, Tokita, and Kariya (2015) asked participants to judge the material and affective properties of the wood samples after viewing, listening, or rubbing them. The first dimension for each modality reflected its intrinsic properties: light/heavy, sparse/dense, and fragile/study for visual presentation; dull sound/sharp sound, mixed sound/pure sound, and damped sound/ringing sound for auditory presentation; rough/smooth, matte surface/gloss surface, and dull sound/sharp sound for touch. The second and third factors reflected the evaluative properties and these were similar for each modality. The second factor grouped terms like expensiveness, rareness, sophistication, and interestingness while the third factor grouped terms like relaxed feeling, pleasantness, and liked. Thus the ratings on the affective or emotional terms were the same across the three kinds of modalities in spite of the large differences in sensations.

5.6 Timbre Constancy

We have argued that timbre adheres to an object much like color. The fundamental problem for color perception is to discount any changes in illumination and recover the reflectance to discover what the color would be under daylight. But the problem for timbre perception is not comparable because the timing and amplitude of the resonances changes at different frequencies. There is nothing that resembles a stable reflectance.

5.6.1 *Independence of Spectral Center and Frequency*

Yet, it is possible to imagine questions and experiments that resemble those for asymmetric color matching. First, the spectral center is based on the relative strength of the harmonics of the fundamental frequency. The spectral center can be raised by increasing the fundamental frequency or by increasing the amplitudes of the higher harmonics. Can participants distinguish the two? If the spectral center is increased, do listeners mistake that for an increase in pitch? Second, the timbre of instruments and singers change at different notes. Can experienced and inexperienced listeners recognize if instruments or singers are identical at different notes?

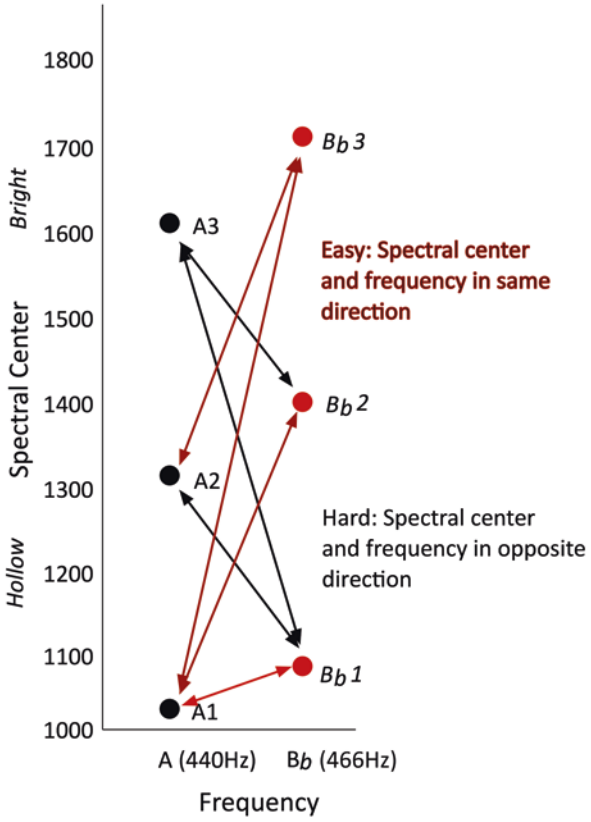

Fig. 5.19 If both the spectral center and the frequency increased for one tone, it was easy to judge which tone had the higher pitch. But, if one tone had the higher spectral center, but the other tone had the higher frequency, it was difficult to judge which tone had the higher pitch. The first five harmonics were used to calculate the spectral center: 440, 880, 1320, 1760, and 2200 Hz for A, 466, 932, 1398, 1864, and 2330 Hz for B_b. The relative amplitudes for spectral center 1 were 0.5, 0.4, 0.3, 0.2, and 0.1; the relative amplitudes for 2 were 0.3, 0.3, 0.3, 0.3, and 0.3; the relative amplitudes for 3 were 0.1, 0.2, 0.3, 0.4, and 0.5

Sound Files 5.19: Comparisons among sounds that vary in fundamental frequency and spectral center

The basic experiment to determine if individuals can judge timbre independently of frequency can be visualized in Fig. 5.19. There would be six sounds: two adjacent frequencies and three levels of timbre based on the spectral center. Differences in the spectrum bring about different perceptions of brightness that may be confused with differences in pitch. On each trial, the participant would be presented two of the sounds and identify the one with the higher pitch. This would be easy when the spectral center is similar, A1-B_b1, A2-B_b2, and A3-B_b3, or when the spectral center and frequency both increase, shown in red. The difficulties arise when the spectrum and frequency differences are in the opposite direction. For example, consider the comparison between A3 and B_b1. Although the actual frequency of A3 is lower than B_b1, if timbre and frequency are not independent the higher spectral center of B_b1 might lead participants to judge A3 as being the higher pitch. In similar fashion, confusions might occur between A2 and B_b1 and between A3 and B_b2.

Experiments (e.g., Allen & Oxenham, 2014) have used a similar procedure. The results indicated that timbre differences based on changes in the spectral center are often confused with independent changes in the fundamental frequency, particularly for the incongruent conditions in which increases in spectral brightness conflicted with decreases in frequency. The greatest number of confusions occurred between similar sounds; A2 versus B_b1 and A3 versus B_b2 were more likely to be confused than A3 versus B_b1. Probably the latter two sounds were so different that confusions did not occur. Surprisingly, performance was identical for musicians and non-musicians. What this means is that timbre and pitch can be judged separately, but that changes in either the height of the spectral center or frequency can be confused with each other. It seems unlikely that changes in timbre due to variation in the onset time would be confused with changes in frequency.

5.6.2 Timbre of Sources at Different Frequencies

5.6.2.1 Instruments

Figure 5.5 illustrates that the sound body filters of most objects contain multiple resonance modes. Hence, we might expect that the sound quality of objects, that is, timbre, will vary as the source frequencies change. A compelling demonstration of the change in timbre across frequencies is found in the Acoustical Society of America Auditory Demonstrations CD. On this track, first you will hear the actual notes of a bassoon across three octaves. It does sound like a bassoon throughout. Then the spectrum of the highest note was determined, and all the other notes were synthesized based on this spectrum. Suppose the relative amplitudes of the first four harmonics of the highest note were 2, 4, 1, and 3. For a 800 Hz tone, the amplitudes of the first harmonic would be 2(800 Hz), the second harmonic would be 4(1600 Hz), the third harmonic would be 1(2400 Hz), and the fourth harmonic would be 3(3200 Hz). For a 1200 Hz tone, the first four harmonics would have the same relative amplitudes: 2(1200 Hz), 4(2400 Hz), 1(3600 Hz), and 3(4800 Hz). The spectrum, that is, the relative amplitude of the harmonics, remains constant throughout the playing range. The synthesized notes should sound like the actual notes of the

bassoon, but that obviously does not occur. At the different scale notes, different resonances of the bassoon body emerge, changing the timbre.

Sound File 5.20: The original bassoon notes and the simulated notes created by keeping the spectrum constant. By permission, Acoustical Society of America

Given this variability in the sound reaching the listener, it is no surprise that the ability to recognize instruments, objects, and events is relatively poor. While it is possible to imagine that colors possess a reference under daylight, there are no such references for timbre. In addition to the inherent acoustic variability, there are environmental differences (enclosed rooms versus outdoors), memory limitations, individual differences, and expectations that also act to depress recognition.

The simplest experiments present two wind instrumental sounds at different pitches (Handel & Erickson, 2001). In some trials, the instruments are identical, (e.g., Clarinet G_3/Clarinet G_4) and in others the instruments and pitches differ (Clarinet G_3/Trombone G_4). The participant's task was simply to judge whether the two sounds were from the same instrument or not. For non-musicians, this is a relatively difficult judgment, and if the pitch difference was an octave or more, all pairs were judged as being different instruments. At intervals less than one octave, participants tended to judge two instruments within the same type (French horn, clarinet, and trombone or English horn and trumpet) as identical. Sound examples are found in Sound Files 5.21A–D.

Sound Files 5.21: Comparisons of instrumental notes that differ by less than one octave and by more than one octave

These outcomes should be treated with caution, however. All were wind instruments, and we certainly would not expect confusion between a wind and a stringed instrument. Moreover, the participants were inexperienced and musically trained participants would be expected to do better. Steele and Williams (2006) found that experienced musicians were able to correctly discriminate between a bassoon and French horn at separations of two octaves, while non-musicians were unable to make that discrimination beyond one octave. It is interesting to note that while musicians and non-musicians make exactly the same similarity judgments between instruments, musicians are better at timbre and pitch discrimination.

The argument made here is that is that because timbre from one source changes from note to note due to the different resonances, one needs to create a transformation that would connect the sounds. From this perspective, only two sounds are presented in the above experiments and that would preclude the formation of such a transformation. Musicians, having experienced many more notes from instruments are more likely to have created such internal transformations that enable them to extrapolate and interpolate what other notes would sound like.

5.6.2.2 The Oddball Task: Instruments and Singer's Voices

To investigate whether a richer set of notes can improve the ability to recognize instruments and singers at different pitches, Erickson and co-workers made use of an oddball procedure. Three or six notes were presented sequentially, the

same instrument or singer presenting all but one of the notes. The participant's task was to identify the odd note. Compared to the experiments that used but two notes, the three-note sequence should provide a slightly stronger sense of the transformations across pitch, while the six-note sequence should generate a much stronger sense of the change in timbre across the notes. This should lead to the paradoxical outcome that the six-note sequence would produce better performance, although it is a more complex task.

To provide a baseline, the three-note sequences were presented using one instrument or voice. In all but one condition more than 80 percent of the responses chose the most dissimilar note as the oddball. If there actually was an oddball instrument or voice, the results were more complex. If the two instruments were both woodwinds, then the same errors occurred; participants nearly always judged the most dissimilar pitch as the oddball. However, if one instrument was a woodwind and the other a brass, then the ability to identify the oddball when it was not the most dissimilar pitch increased dramatically. In sum, if the two instruments were the same type, then participants were unable to differentiate their extrapolations and pitch dominated the judgments. If the

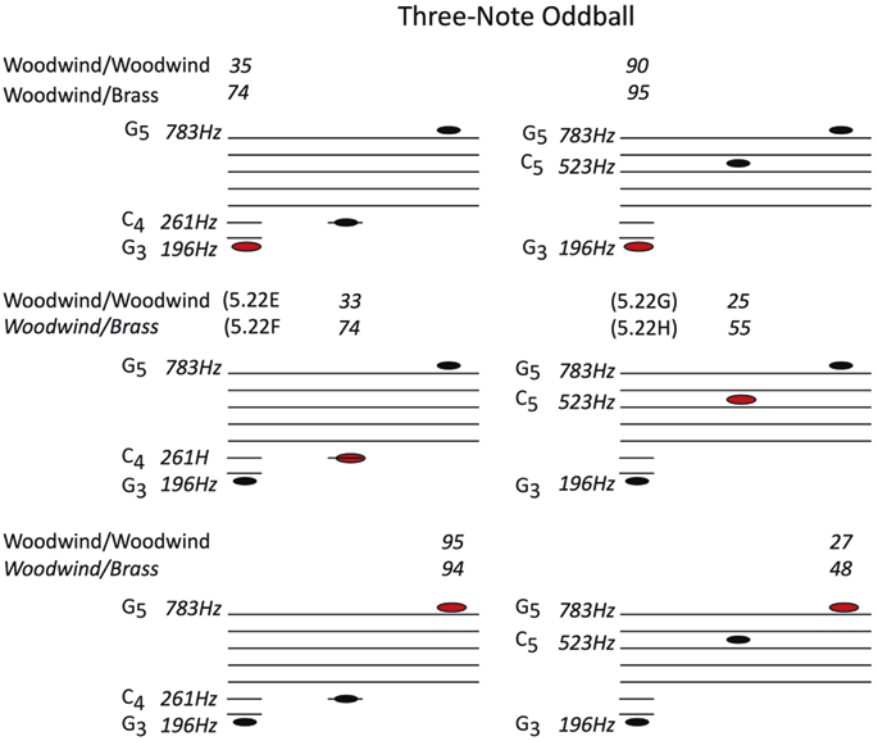

Fig. 5.22 The percentages correct from the three-note oddball task. The oddball note played by the second instrument is portrayed in red. The oddball note was usually picked as being the extreme pitch particularly if both instruments were in the same class. If an instrument in another class played the oddball note identification improved

Sound Files 5.22: The oddball comparisons shown in Fig. 5.22E–5.22H (Middle row)

instruments were from different types, participants were able to perceive the differences in timbre even across dissimilar pitches. These outcomes can be seen in Fig. 5.22.

The results from the three-note oddball task are a mixed bag. There is some evidence that the third note aided identification, but the performance within the same class of instruments just slightly improved. To investigate whether a richer set of six notes would truly increase the identification of the oddball, Erickson and Perry (2003) constructed three-note and six-note sequences sung by sopranos and mezzo-sopranos. For soprano/soprano sequences, one soprano sang two of the three or five or the six notes and the other soprano sang the remaining note. The same procedure was used for the mezzo-soprano/mezzo-soprano and soprano/mezzo-soprano sequences.

The identification results for the three-note oddball experiments mirrored those for instruments. Participants were unable to identify the oddball singer. Surprisingly, the performance for the six-note sequences was better absolutely, even though chance performance was 16 percent versus 33 percent for the three-note sequences. What is more important was that the errors were scattered through the six notes instead of bunching at the lowest and highest pitches. The errors reflected idiosyncratic variation in the singers, further supporting the conclusion that participants were constructing a transformation that enabled them to link the different sung notes. These transformations created a frame of reference that allowed the idiosyncratic variation to be isolated (and possibly misjudged as the oddball).

All of this occurred in a controlled laboratory setting, but in our daily lives nearly every sound could have come from many different sources. For example, a *click* could be due to a ballpoint pen, light switch, computer keyboard, or a simulated camera snap on a smart phone. Ballas (1993) presented everyday sounds and asked participants to identify them. The responses to some sounds were almost unanimous and were easy to identify in later experiments. The responses to other sounds were quite diverse, and more difficult to identify.

What this all means is that there is no single cue for identifying the source of sounds. Ballas (1993) concludes that there are many domains that contribute to identification. First, there are temporal properties ranging from the timing of multiple impacts to the length of the onset transient. Second, there are spectral properties such as the spectral center arising from the resonances of the source. Third, there are small variations in the resonances that yield the sense of change. Fourth, there is the familiarity with sounds (e.g., musicians do better than non-musicians) that arises from the frequency of occurrence in the environment. Finally, we should not forget expertise that arises from extensive study of types of sounds. Bird-watchers, railroad buffs, or car enthusiasts can make fine distinctions that novices cannot. (This may be identical to chess experts being able to detect small differences among the placement of pieces).

Perhaps there is “no smoking gun” for any sort of perception. There is always the inverse problem for color and timbre and doubly so for the objects in the world.

5.7 Echolocation

Echolocation is the process of emitting sounds in order to determine from these reflections and echoes the location and material properties of objects. Both sighted and blind individuals make use of the information gained through echolocation. Builders and carpenters tap walls to determine the location of supporting studs; boat builders and physicians tap enclosed spaces to determine if the spaces are hollow or filled with fluid. Blind individuals snap fingers, make clicking sounds with tongues, stomp feet, and tap canes to locate vertical material surfaces. The source is the produced sound, for example, the rapping against the wall or the tongue click, and the filter is the reflecting surface or hollow that changes the frequency spectrum from the source to the echo. As detailed below, accurate perception depends on the comparison of the source and the echo. The reflection by itself is usually insufficient. Moreover, as found previously, all perceiving is based on the combination of cues evolving over time.

The current interest can be traced to the discovery by Donald Griffin in 1938 of the ultrasonic calls of bats used to navigate and to hunt insects. The comparable use of ultrasonic calls by dolphins and other toothed whales was discovered later. It is not surprising that the ultrasonic calls of bats and dolphins differ in such dissimilar surrounds. In air, the speed of sound is slow and the higher frequencies rapidly dissipate. Narrow frequency sounds can be sustained for the best detection of prey while the frequency of repeating short broadband sounds best for localization can be increased as the bat approaches prey without creating confusion between the outgoing sound and the echo. Bats have time-delay neurons that respond best to signals separated by a fixed delay. (I have often wondered how bats when moving in a large group avoid being confused by the calls of other bats. It seems that the bats vary the frequency of their calls to reduce acoustic interference. This seems analogous to blind individuals moving their head direction when echolocating and individuals changing their speaking voice in crowded noisy rooms). In water, the speed of sound is roughly five times faster than in air so that the calls of dolphins tend to be shorter wide broadband signals at slower rates. In contrast to bats, dolphins reduce the amplitude and rate of clicks as they approach prey. Otherwise, the signals and echoes would overlap and interfere.

The ability of humans to echolocate has been known for hundreds of years. But the fact that it is based on acoustic information has been discovered only in the last 60–70 years because the blind could not identify the cues they used. Many attributed their ability to heightened facial or tactile sensations such as shadows or pressures across the eyes. Studies begun in the 1940s showed that auditory information was primary. Blind and deaf people were unable to judge when to stop walking in front of a wall, but blind and hearing individuals could. If blind people wore hearing protectors or were placed in noisy rooms, they increased the intensity of their foot shuffling or made other noises to compensate. Facial masks made from cotton did not affect localization, undercutting the facial vision theory (Ammons, Worchel, & Dallenbach, 1953).

The vocal sounds humans use for echolocation are usually short wide band clicks produced by tongues or snapping fingers, with maximum energy in the range between 2000–4000 Hz with higher frequencies from 6000 Hz to 10,000 Hz. The duration ranges from 3 ms to 8 ms and the clicks repeat from 1.5 to 2 times per second. The outgoing click is beam-like; the amplitude is symmetrical and relatively constant out to 60^0 both horizontally and vertically (Thaler et al., 2017). Given the beam characteristics, it would be difficult for a walking person to use echolocation to investigate ground level objects, and that accounts for the use of canes. Although a hissing continuous noise would make an effective source, it has not been employed in the experimental tasks. Based on these sounds, we can identify potential auditory cues that could be used to determine the surface, location, and orientation of objects (Kolarik, Cirstea, Pardhan, & Moore, 2014).

5.7.1 Acoustic Cues

A. The most obvious cue for distance is the loudness of the returning echo. In general, loudness decreases as a function of the distance squared ($1/d^2$) between the source and the listener. The sound moves out to the distant object and back again as the echo, so that as the distance doubles from the source, the distance the sound travels increases fourfold and the level drops to 1/16th (if there is no absorption at the target).

But loudness of the returning echo is inherently ambiguous. If we were simply trying to estimate the distance of an external source, any given loudness would be due to the inherent loudness of the source and the size, distance, or orientation of the source. A low-intensity sound could have come from a nearby soft source or a loud distant one. Echoes too are ambiguous. A soft returning echo could come from a distant object, but also could be due to other properties of the object; a smaller as opposed to a larger reflecting surface will return a smaller percentage of the outgoing source energy; an offset or angled surface also will return just a fraction of the energy because much of the echo is reflected away from the source; or a soft textured surface might absorb much of the incoming energy so that little of the energy is returned. To untangle the ambiguity, the listener could make use of prior knowledge about the size, orientation, and surface of the objects in the environment, or some of the other available cues discussed below.

B. The second obvious cue is the time delay between the outgoing click and the returning echo. The speed of sound is roughly 1100 ft./sec or 340 meters/sec. For an object 1 ft. away, the echo delay would be about 2 msec (2/1100) and for an object 3 ft. away the echo delay would be about 6 msec.

The difficulty in making use of the echo delay is due to the *precedence or Haas* effect. In general, any external sound is likely to reach the listener multiple

times after being reflected by the surfaces in our cluttered environment. The paths will be different lengths bombarding the listener with the more or less identical sounds, one after the other. If we analyzed each sound separately, each one would seem to have come from a different direction. But we do not hear all of these echoes; instead we hear the sound as coming from the direction of the first arriving sound. The later arriving echoes reinforce the direct sound and lead to a sense of spaciousness, but they do not affect the perceived location. A neat example occurs if several hi-fi speakers are arrayed in a line in front of a listener. As the listener walks back and forth the sound seems to travel along because the sound from the speaker in front has the shortest path and arrives first.

The precedence effect has two consequences. First, the emitted click, perceived acoustically or through bone conduction, can merge with and dominate the returning echo suppressing it so that both the emitted click and echo seem to be in front of the face. Second, more distant surfaces can be hidden. If we have two surfaces 10° either side of the midline, the echo from the nearer surface will reach the listener first and the echo from the further surface is perceived to be the same direction as the first echo. Another consequence of the precedence effect is to minimize the effectiveness of time differences between echoes reaching each ear as described below. The act of vocalizing the outgoing source may act to minimize the suppression of the echo due to the precedence effect (Wallmeier, GeBele, & Wiegrebe, 2013).

C. The third cue is the ratio of direct to reflected sound. A sound in an enclosed environment will reach the listener from many directions. The direct sound from the object is the loudest and arrives first so that due to the precedence effect it defines the direction. Following that are sounds that have bounced off the walls and ceiling once or multiple times along many paths of differing lengths so that they reach the listener from many angles with varying delays. The reflected waves are weaker due simply to air friction and at each reflection, the sound loses energy; fuzzy materials absorb more energy, hard non-porous material absorb less. Still, overall, the energy of the reflected waves is roughly 10 times that of the direct wave. But, if the energy of direct wave is quickly swamped by the reflected waves, the direction and clarity of the source is lost. Nonetheless, while the reflected waves reduce the perception of direction, it seems to have little effect on the perception of distance (Wallmeier & Wiegrebe, 2014)

If the sound continues until the level is uniform and then stopped, the reverberation time is defined as the duration of the sound until it is reduced to one millionth of the original level. The energy lost in any given time interval is proportional to the remaining energy. In general, the reverberation time for smaller rooms is shorter because it will have more reflections in a given time interval and therefore lose energy more quickly.

The intensity of the reflected echo decreases with increasing distance according to $1/d^2$ but the intensity of the reflected sound off the surround decreases only slowly as the room size increases. Thus, the ratio of direct to reverberant energy is smaller for larger rooms, which becomes an important cue for judging the source distance. However, for ongoing sounds it is unlikely that listeners can distinguish between the direct and reflected sound. They probably make use of the amount of reverberant energy, the correspondence between the reverberant energy in each ear, and/or the overall level. The best strategy of course is to make use of the optimal cue for each listening situation (Kopco & Shinn-Cunningham, 2011).

Sound Files 5.23: Simulated Reverberations in differently sized spaces

D. Proximity resonance. As a sound wave moves toward and bounces off a surface the energy at that surface increases due to constructive addition, where the amplitudes of the incoming wave and reflected wave combine so that the amplitude at the wall is twice the initial amplitude. Moving away from the wall, the amplitude oscillates; at some points higher-pressure regions overlap, but at other points the higher-pressure regions overlap the lower pressure regions producing "destructive addition." The pressure variation decreases further away from the wall because the returning wave has lost energy.

The important point for echolocation is that the change in amplitude due to reflecting wave is a function of wavelength. Simply put, the buildup in amplitude near the wall is greater for low frequencies. For a tone of 110 Hz, the amplitude starts to increase 15″ (0.38 m) from the wall (wavelength/8), but for an 1100 Hz tone the increase begins just 1.5″ (0.038 m) from the wall. Thus it seems likely that blind individuals would attend to the low-frequency variation.

Ashmead and Wall (1999) investigated the ability of sighted people to detect the distance to a frontal wall based on the spectrum of the reflected sound wave. To do so, the authors simulated the spectrum of sound at varying distances from a wall. Sounds that represented those close to a wall had more low-frequency energy than those that represented sounds further out. Overall, participants were able to detect a wall at a distance of about 1.5′ (0.5 m).

Surprisingly, simulating movement toward the wall did not improve detection beyond simply presenting the spectrum of the closest point. The 1.5′ (0.5 m) distance matches measurements when blind children walk along a hallway. If the children walked along a narrow hallway less than 6′ (about 2 m) wide, they were able to maintain a smooth trajectory. If they veered slightly, they would be able to make use of the spectral change perceivable near the opposite wall and correct their course.

E. Spectrum of returning echo. Due to friction, high frequencies decay more rapidly than low frequencies. Listeners judge the distance of sounds with high frequency as being closer than sounds without high frequency. Furthermore, the material of the reflecting surface will affect the spectrum of the returning echo. Softer materials such as carpet absorb a higher per-

centage of higher-frequency energy than harder materials such as metal or rigid plastic so that comparing the spectrum of the source to that of the echo can be used to make a somewhat rough guess as to the surface.

Experts can discriminate among shapes, such as a square versus a triangle. For two- and three-dimensional objects, the reflected echoes may arrive at different times and therefore interact with each other creating spectral changes. There may also be specific spatial reflections at the edges of objects that change the spectrum of the returning echoes. Although the cues used to make these object discriminations are not known, it is likely that spectral changes and loudness differences contribute.

F. Prior knowledge. Experience with events and objects can yield cues to distance. We have heard thunder, ambulances, and fire engines at close hand, have learned that whispering does not include voicing, and that shouts are characterized by high-frequency energy. That knowledge allows us to disregard the overall level in judging distance and allows us to make estimates based on the level and spectrum of those sounds.
G. Repetition pitch. Experts often comment that changes in pitch are important cues for estimating distance. As suggested above, this may be due to enhancement of different frequencies due to proximity resonance. But a sense of pitch can arise as a consequence of the overlap of the source and echo or the delay between the direct echo and the later echoes reflected off a wall.

Imagine standing in the middle of a room with a sound source directed at the front wall. The direct echo will arrive first followed by reflections off the floor, the sidewalls, or the back wall. Suppose the sound bouncing off the floor is delayed by 10 msec, that is, the reflected sound has travelled an additional 11′. Assume that the direct echo and the floor bounce echo are identical. What one hears is the combination of the two sounds, the direct echo plus the floor echo delayed by 10 msec away (Fig. 5.24). The same outcome would occur near the sidewall of a concert hall. You would hear the sum of the direct sound plus the sound delayed by 10 msec after bouncing off the sidewall.

The correlation between adjacent elements in the direct echo and floor echo is zero because they are a sequence of random numbers. But, the correlation between adjacent elements in the sum has increased to some degree because adjacent values share the same amplitude. For example, 8+3 and 4+3 share 3, 8+3 and 8+6 share 8, and so on. Part of the sum is identical between pairs. This correlation yields the faint pitch based on the 10 msec delay or 100 Hz. If the delay increases to 20 msec (the reflected echo traveled an addition 22′), the perceived pitch would become 50 Hz, or if the decay decreased to 5 msec (the reflected echo traveled an addition 5.5′) the perceived pitch would increase to 200 Hz. On this basis, the pitch from the repetition delay could bring about the perception of distance. As the pitch increases, surfaces would appear closer and conversely as the pitch decreases surfaces would appear further.

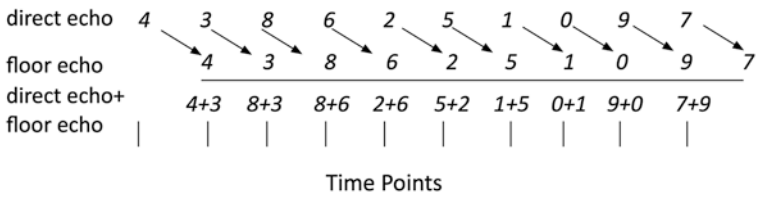

Fig. 5.24 The amplitudes, represented by the random numbers, of the direct and floor bounce echo are identical. The sense of pitch caused by the delay+add of the two echoes is created by the partial correspondence between the amplitudes in adjacent points, for example, (4+3) & (8+3) and (8+3) & (8+6). Notice that there is no correspondence in amplitude between time points two steps apart

Sound Files 5.24: Examples of repetition pitch at different delays and number of echoes

While repetition pitch may seem an unlikely phenomenon, it occurs in nearly all environments. Even in an open field, for example, the delayed sound bouncing off the ground leads to repetition pitch. In concert halls, seats near sidewalls can be susceptible to repetition pitch from those delayed sound waves. It is better to sit in the middle. (In general though, the reflections occur at different delays that would eliminate the repetition pitch). If a sound source is in front of a staircase, each step can yield one reflection, and equally spaced steps will result in a series of equally spaced echoes that result in a strong sense of pitch. The only sure way of avoiding the effects of repetition noise is to float in outer space. But, then again sound pressure will not travel in a vacuum.

H. Binaural cues. Due to the circular head and outer ear (i.e., pinna), lateral sources give rise to interaural time and interaural level differences. An informal observation shows the advantage of binaural hearing. If one approaches a wall sideways so the sound in each ear differs, it is far easier to detect the wall than if one approaches the wall straight on.

The maximum interaural time and level difference would occur when the source is directly opposite one ear. Even at +/−90°, the time difference is only 1 msec but does not change as the distance from the source varies. Even though we can easily make discriminations at this short time interval, we might expect that in normal environments this cue could not be used to determine distance due to the precedence effect. The interaural level difference changes as a function of the frequency of the source and its distance. Low frequencies at the near side are enhanced at close distances due to proximity resonance (i.e., sound waves bouncing off the head). But at greater distances the low frequencies bend around the head so that there is little difference in intensity between the near and far ears. (Sounds will bend around the head when the wavelength of the head matches the diameter of the head, roughly 1500 Hz). For higher frequencies, the head creates a shadow such that the sound at the closer ear is more intense than the farther one that does not vary with the distance of the source. Thus, the relative intensities of the low and high frequencies can provide a cue to distance.

Multiple studies provide strong evidence that expert echo locators make very accurate judgments about distance, location, and surface qualities and that sighted people can learn to make the same discriminations after a relatively short training period (e.g., Donelli, Brayda, & Gori, 2016). (It is true that the performance of such sighted individuals does not match those of experienced echolocators, but this is probably due to the short training). It is frustrating given the strong performance and the wide variety of cues available for distance, orientation, shape, and surface texture that there is very little agreement about how judgments are made.

The lack of consensus is due to many factors. The initial studies were mainly demonstrations that echolocation by humans was in fact possible so that potential acoustic cues were not measured nor correlated to performance. Recent studies have imposed more control to determine which acoustic cues were available and which correlate to accuracy (Kopco & Shinn-Cunningham, 2011; Papadopoulos, Edwards, Rowan, & Allen, 2011). To do so, researchers have first recorded the source and echoes from a surface at the head of a manikin in an anechoic or in a reverberant enclosure. The sounds measured at the manikin's ears (sometimes termed *phantom* sounds) are altered in various ways to accentuate or eliminate acoustic variables (e.g., removing low frequencies), and then presenting the higher frequencies to participants by means of headphones. The importance of the cues can be evaluated from the accuracy of the judgments. The downside is that the participants become passive, and unable either to vary their source output or move their heads. Even so, it is unclear how movement affects the accuracy of echolocation (Wallmeier & Wiegrebe, 2014). In general, we do not know how this research generalizes to real-life judgments.

Probably the biggest stumbling block is the assumption that there is a fixed set of cues. Depending on context, individuals will make use of any cue that works. For example, Kopco and Shinn-Cunningham (2011) suggest that in highly reverberant rooms, participants use the direct to reverberant energy for low frequencies particularly for nearby surfaces, but in low reverberant rooms use the interaural level differences. The highly reverberant room will tend to mask the level difference.

Perceiving should be opportunistic. Blind people can use self-generated sound to navigate, but can also make use of their knowledge of environmental sounds. We have mentioned individual sounds such as thunder and fire engines, but there are also "walls" of sound such as rivers or a lines of traffic, and the trajectories of individual cars or black flies. It seems impossible to construct a "neutral" experiment, all perceiving occurs in a context and the best we can do is discover the optimal cues in each such context.

5.7.2 Physiological Mechanisms

Recent experiments have investigated the neural underpinnings of echolocation for early blind, late blind, and sighted. Several studies have found increased activation in secondary visual areas during echolocation for the early blind but not for the late blind or the sighted participants. Bauer et al. (2017) found

anatomical changes in the early blind; in some brain areas the density of their connections increased, while in other areas the density decreased. As described above, the early-blind experts outperform the others in various tasks and that leads to the notion that the recruitment of traditional visual areas gives rise to the superior discrimination.

I am skeptical about this interpretation.

1. In reality, we do not know if the activation actually causes the improved performance or simply follows the auditory "calculation." All we know is that there is more blood flow which is presumed to indicate increased neural activity. The translation of that activity into distance or surface judgments is completely opaque.
2. Early-blind experts learn to correlate auditory cues to aspects of the environment and that would necessarily require some way to calibrate those cues. The obvious way to do this is reaching out to judge distance and orientation, or running a hand around a surface to judge shape. The motor and tactual systems connect the auditory cues to the environment. It is interesting to note that it is harder to make distance judgments for overhead surfaces that cannot be reached than frontal or lateral surfaces that can be touched. It would seem that echolocation in the early blind would be accompanied by increased activation in the tactual and motor cortical areas, rather than in the visual areas. In contrast, the late blind and sighted would probably calibrate the acoustic cues to objects and surfaces through their visual memory so that it seems reasonable that visual areas would be activated. But, that is not the case.
3. Spatial information from echolocation and from direct visual input is fundamentally different. The spatial information underlying echolocation is temporal, involving the interval between the tongue click and the returning echo, or the intervals between successive echoes in reverberant enclosures. The visual information is spatial, the extension and adjacency of extended surfaces, or the connected movement of surfaces. We do not have any understanding how visual brain circuits presumably organized for spatial relationships can encode temporal relationships to derive the absent spatial relationships. It could be that echolocation occurs by means of rhythmic timing circuits, which occur across cortical regions.

5.7.3 Echolocation Summary

Clearly we have strayed from the technical definition of echolocation to include material on passive acoustic information useful for orientation, navigation, and exploring objects. The source-filter model is most relevant to understand active echolocation, but a more generous definition of the filter, to include the acoustics of the general environment, can provide insights about how blind and

sighted can maneuver and make sense of the external world. What is common is the convergence of multiple cues that may be substitutable in some instances and whose usefulness may vary in other instances.

5.8 Overall Summary

Visual and auditory stimulation results from excitation impinging upon and being modified by the properties of the filter, and then being propagated to the perceiver. Source-filter models explain many of the physical properties of the energy reaching the eyes and ears and give insights into why perceiving is so difficult. The same object can look and sound so different in different contexts because the proximal stimulation can arise from many different combinations of excitation and filter. The perceiver is forced to interpret the proximal stimulation in its particular context. Depending on that interpretation, people can see or hear different objects, as found for "the dress." Given the potential for alternative percepts, it is amazing to me that so few interfere with our actions or cause bodily harm.

References

Allen, E. J., & Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *Journal of the Acoustical Society of America, 135*, 1371–1379. https://doi.org/10.1121/1.4863269

American National Standards Institute. (1973). *Psychoacoustics terminology*. S3.20. New York, NY: American National Standards Institute.

Ammons, C. H., Worchel, P., & Dallenbach, K. M. (1953). "Facial vision": The perception of obstacles out of doors by blindfolded and blindfolded-deafened subjects. *American Journal of Psychology, 56*, 519–553.

Ashmead, D. H., & Wall, R. S. (1999). Auditory perception of walls via spectral variations in the ambient sound field. *Journal of Rehabilitation Research and Development, 36*, 313–322.

Ballas, J. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 250–267.

Bauer, C. M., Hirsh, G. V., Zajac, L., Koo, B.-B., Collignon, O., & Merabet, L. B. (2017). Multimodal MRI-imaging reveals large-scale structural and functional connectivity changes in profound early blindness. *PLoS One, 12*, 1–26. https://doi.org/10.1371/journal.pone.0173064

Brainard, D. H., & Hurlbert, A. C. (2015). Colour vision: Understanding #thedress. *Current Biology, 25*, R549–R568. https://doi.org/10.1016/j.cub.2015.05.020

Brainard, D. H., & Maloney, L. T. (2011). Surface color perception and equivalent illumination models. *Journal of Vision, 11*(5), 1–18. https://doi.org/10.1167/11.5.1

Bramao, I., Reis, A., Peterson, K. M., & Faisca, L. (2011). The role of color information on object recognition: A review and meta-analysis. *Acta Psychologica, 136*, 244–253. https://doi.org/10.1016/j.actpsy.2011.06.010

Cornsweet, T. (1970). *Visual perception*. New York, NY: Academic Press.

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

Dedrick, D. (2015). Some philosophical questions about color. In A. J. Elliot, M. D. Fairchild, & A. Franklin (Eds.), *Handbook of color psychology* (pp. 131–145). Cambridge, UK: Cambridge University Press.

Donelli, A., Brayda, L., & Gori, M. (2016). Depth echolocation learnt by novice sighted people. *PLoS One, 11*, e0156654. https://doi.org/10.1167/journal.pone.0156654

Erickson, M. L., & Perry, S. R. (2003). Can listeners hear who is singing? A comparison of three-note and six-note discrimination tasks. *Journal of Voice, 17*, 353–369. https://doi.org/10.1067/S0892-19970903000021-3

Fujisaki, W., Soda, N., Motoyoshi, I., Komatsu, H., & Nishida, S. (2014). Audiovisual integration in the human perception of materials. *Journal of Vision, 14*, 1–20. https://doi.org/10.1167/14.4.12

Fujisaki, W., Tokita, M., & Kariya, K. (2015). Perception of the material properties of wood based on vision, audition and touch. *Vision Research, 109*, 185–200. https://doi.org/10.1016/j.visres.2014.11.020

Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology, 5*, 1–29.

Gegenfurtner, K. R., Bloj, M., & Toscari, M. (2015). The many colours of "the dress". *Current Biology, 25*, R523–R548. https://doi.org/10.1016/j.cub.2015.04.043

Gilchrist, A. (2015). Perceptual organization in lightness. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 391–412). Oxford, UK: Oxford University Press.

Giordano, B. L., & McAdams, S. (2010). Sound source mechanics and musical timbre: Evidence from previous studies. *Music Perception, 28*, 155–168. https://doi.org/10.1525/mp.2010.28.2.155

Gough, C. E. (2016). Violin acoustics. *Acoustics Today, 12*(2), 22–30.

Guyot, P., Houix, O., Misdaris, N., Susini, P., Pinquier, J., & Andre-Obrecht, R. (2017). Identification of categories of liquid sounds. *Journal of the Acoustical Society of America, 142*, 878–889. https://doi.org/10.1121/1.4996124

Handel, S., & Erickson, M. L. (2001). A rule of thumb: The bandwidth for timbre invariance is an octave. *Music Perception, 19*, 121–0126. https://doi.org/10.1525/mp.2001.19.1.121

Haywood, N. R., & Roberts, B. (2013). Build-up of auditory stream segregation induced by tone sequences of constant or alternating frequency and the resetting effects of single deviants. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 1652–1666. https://doi.org/10.1037/a0032562

Hjoetkjaer, J., & McAdams, S. (2016). Spectral and temporal cues for perception of material and action categories in impacted sound sources. *Journal of the Acoustical Society of America, 140*, 409–420. https://doi.org/10.1121/1.4955181

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kaiser, P. K., & Boynton, R. M. (1996). *Human color vision* (2nd ed.). Washington, DC: Optical Society of America.

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

Kolarik, A. J., Cirstea, S., Pardhan, S., & Moore, B. C. J. (2014). A summary of research investigating echolocation abilities of blind and sighted humans. *Hearing Research, 310*, 60–68. https://doi.org/10.1016/j.heares.2014.01.010

Kopco, N., & Shinn-Cunningham, B. G. (2011). Effect of stimulus spectrum on distance perception for nearby sources. *Journal of the Acoustical Society of America, 130*, 1530–1541. https://doi.org/10.1121/1.3613705

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbre: Common dimensions, specificities, and latent subject classes. *Psychological Research, 58*, 177–192. https://doi.org/10.1007/BF00419633

Nathmann, A., & Malcolm, G. L. (2016). Eye guidance during real-world scene search: The role color plays in central and peripheral vision. *Journal of Vision, 16*, 1–16. https://doi.org/10.1167/16.2.3

O'Modhrain, S., & Gillespie, R. B. (2018). One more, with feeling: Revisiting the role of touch in performer-instrument interaction. In S. Papetti & C. Saitis (Eds.), *Musical haptics* (Vol. 18, pp. 11–27). Springer.

Ogg, l., Sleve, R., & Idsardi, J. (2017). The time course of sound category identification: Insights into acoustic features. *Journal of the Acoustical Society of America, 142*, 3459–3473. https://doi.org/10.1121/1.5014057

Papadopoulos, T., Edwards, D. S., Rowan, D., & Allen, R. (2011). Identification of auditory cues utilized in human echolocation-objective measurement results. *Biomedical Signal Processing and Control, 6*, 280–290. https://doi.org/10.1016/jbspc.2011.03.005

Radonjić, A., Gottaris, N. P., & Brainard, D. H. (2015). Color constancy in a naturalistic, goal-directed task. *Journal of Vision, 15*, 1–21. https://doi.org/10.1167/15.13.3

Rowland, H. M. (2009). From Abbot Thayer to the present day: What have we learned about the function of countershading? *Philosophical Transactions of the Royal Society, Section B, 364*, 519–527. https://doi.org/10.1098/rstb.2008.0261

Saitis, C., Järveläinen, H., & Fritz, C. (2018). The role of haptic cues in musical instrument quality perception. In S. Papetti & C. Saitis (Eds.), *Musical haptics* (pp. 73–93). Springer.

Steele, K. M., & Williams, A. K. (2006). Is the bandwidth for timbre invariance only one octave? *Music Perception, 23*, 215–220. https://doi.org/10.1525/mp.2006.23.215

Thaler, L., Reich, G. W., Zhang, X., Wang, D., Smith, G. E., Tao, Z., … Antonio, M. (2017). Mouth-clicks used by blind expert human echolocators – Signal description and model based signal synthesis. *PLoS Computational Biology, 13*, e1005670. https://doi.org/10.1371/journal.pchi.1005670

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

Wallisch, P. (2017). Illumination assumptions account for individual differences in the perceptual interpretation of a profoundly ambiguous stimulus in the color domain: "The dress". *Journal of Vision, 17*, 5. https://doi.org/10.1167/17.4.5

Wallmeier, L., GeBele, N., & Wiegrebe, L. (2013). Echolocation versus echo suppression in humans. *Proceedings of the Royal Society, B, 280*, 2013.1428. https://doi.org/10.1098/rspb.2013.1428

Wallmeier, L., & Wiegrebe, L. (2014). Ranging in human sonar: Effects of additional early reflections and exploratory head movements. *PLoS One, 9*, e15363. https://doi.org/10.1371/journal.pone.0115363

Warren, W. H. J., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 704–712.

Witzel, C., Racey, C., & O'Reagan, J. K. (2017). The most reasonable explanation of the "dress": Implicit assumptions about illumination. *Journal of Vision, 17*, 1–19. https://doi.org/10.1016/17.2.1