Rev.Phil.Psych. (2010) 1:41–61
DOI 10.1007/s13164-009-0004-5

Audio-Visual Objects

Michael Kubovy · Michael Schutz

Published online: 12 January 2010

Abstract In this paper we offer a theory of cross-modal objects. To begin, we discuss two kinds of linkages between vision and audition. The first is a duality. The the visual system detects and identifies *surfaces*; the auditory system detects and identifies *sources*. Surfaces are illuminated by sources of light; sound is reflected off surfaces. However, the visual system discounts sources and the auditory system discounts surfaces. These and similar considerations lead to the Theory of Indispensable Attributes that states the conditions for the formation of gestalts in the two modalities. The second linkage involves the formation of audiovisual objects, integrated cross-modal experiences. We describe research that reveals the role of cross-modal causality in the formation of such objects. These experiments use the canonical example of a causal link between vision and audition: a visible impact that causes a percussive sound.

> [A fire is] a terrestrial event with flames and fuel. It is a source of four kinds of stimulation, since it gives off sound, odor, heat and light .... One can hear it, smell it, feel it, and see it, or get any combination of these detections, and thereby perceive a fire .... For this event, the four kinds of stimulus information and the four perceptual systems are *equivalent*.

The authors work is supported by grants from NEI and NIDCD.

M. Kubovy (✉)
Department of Psychology, University of Virginia, P.O. Box 400400,
Charlottesville, VA 22904–4400, USA
e-mail: kubovy@virginia.edu

M. Schutz
School of the Arts, McMaster University, 1280 Main St. W.,
Hamilton Ontario L85 4MZ, Canada
e-mail: schutz@mcmaster.ca

> If the perception of fire were a compound of separate sensations of sound, smell, warmth and color, they would have had to be associated in past experience in order to explain how any one of them could evoke memories of all the others. . . .
>
> [T]he problem of perception is not how sensations get associated; it is *how the sound, the odor, the warmth, or the light that specifies fire gets discriminated from all the other sounds, odors, warmths, and lights that do not specify fire*. Gibson (1966) (pp. 54–55)

In this paper, we offer a theory of cross-modal objects. We agree with Gibson's assertion that such a theory is unlikely to be an associative theory. Instead, our theory is built on the notion of privileged inter-modal binding. As an example of such privileged binding, we will examine the relation between visible impacts and percussive sounds, which allows for a particularly powerful form of binding that produces audio-visual objects. To motivate these conclusions we devote the first two sections of this article to a review of Kubovy and Van Valkenburg's (Cognition 80(1–2):97–126, 2001) theory of auditory and visual objects. In the final section, we present our new approach and present empirical data to support our view.

1 The Duality of Vision and Audition

The paths to understanding vision and audition differ in many ways. For example, to understand the two mechanisms through which sensory information is delivered (as opposed to the corresponding cortical systems in which the information is processed) we must call upon different physical sciences. The study of the visual system requires an understanding of optics and photochemistry; whereas the study of the auditory system requires acoustics, mechanics, and fluid dynamics. Such differences do not offer a path to understanding cross-modal objects.

For our purposes there is a deeper difference: one of function. The main function of the visual system is to detect and identify surfaces; the main function of the auditory system is to detect and identify sources. They cannot fulfill their main functions, however, without taking other information into account. Surfaces are illuminated by sources of light; therefore sources are an inevitable aspect of our visual world. Sound is reflected off surfaces; therefore surfaces are an inevitable aspect of our auditory world. Neither vision nor audition can take place without a source of energy: a source of light or a source of sound. Both occur in a world of objects bounded by surfaces.

The visual system evolved to allow mammals to navigate through a cluttered environment, detect danger, find nourishment, and engage in social interaction. For these it processes information about surfaces: the location of an obstacle, the threat of a predator, the ripeness of a fruit, and the friendliness of a conspecific. All the while it has information about the nature of the light that illuminates the scene. Although this information may be used to compute

features of the scene, once used it is discounted. By this we mean that it is generally not assessed or measured accurately.

That the visual system discounts source information is evident from *lightness* and *color constancy*. Despite diurnal and seasonal variations in the composition and the intensity of the light that illuminates surfaces, animals and humans do not perceive much variation in the lightness and the colors of objects. The visual system has evolved so as to make this correction independent of explicit information about the spectrum of the light source (Amano et al. 2006). These perceptual constancies are no doubt valuable adaptations, without which we could not reliably distinguish degrees of fruit ripeness under different kinds of illumination. As Mollon (1995) writes:

> our visual system is built to recognize ... permanent properties of objects, their spectral reflectances, ... not ... the spectral flux ... (pp. 148, 149)

As I write about light source information, I notice that my desk (which happens to be in an unfamiliar environment) is lit by two strong direct lights and several more distant and diffuse ones. Diverse objects on my desk cast shadows that are not consistent with a single source of light. Although in principle I could have inferred the location of the sources from these shadows, I had to look up to see where these lights are.

Although we are generally not aware of light source information, our visual system uses this information tacitly. In the two panels of Fig. 1 the green squares and the checkered backgrounds are the same, but the green square on the right seems to hover further above the surface than the green square on the left. Your visual system is "assuming" that a source of light is illuminating the scene from the upper left. It is taking the separation between the green square and the shadow it casts as a cue to the green square's elevation.

As Fig. 2 shows, this "inference" by the visual system is by no means necessary. The same change in the cast shadow could have been achieved by assuming that different lights are illuminating the left and right sides of the scene.

As mentioned earlier, the auditory system is more interested in sources than surfaces. Even though a sound in a room is repeatedly reflected by the walls, we hear a single sound at the location of the source. The reflected sounds are suppressed; we hear them as echoes (in a cave, for example) only when the

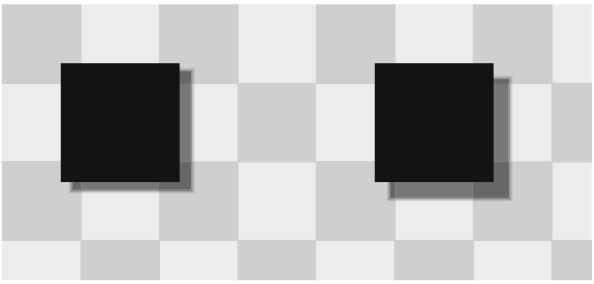

Fig. 1 Cast shadows, an indirect effect of illumination. The *two squares* are of the same size and in the same position relative to the background checker board. Illumination is tacitly used and discounted

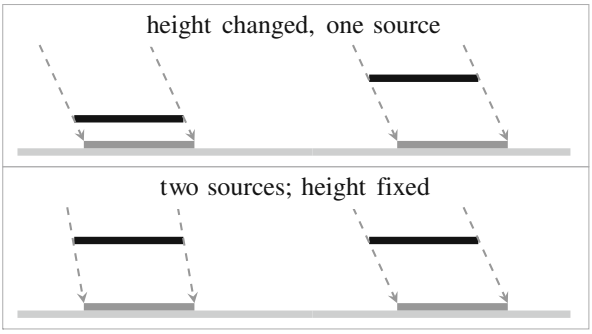

Fig. 2 The ambiguity of shadows. Alternative schematic elevations of the situation in Fig. 1. The *dashed lines* represent hypothetical light rays. The *gray bars* represent the shadows cast by the squares, which are represented by *black bars*

delay of the reflected sound is greater than the echo threshold. The first sound in a sound train comes from the source itself—it precedes the reflected sounds. The suppression of the reflected sounds and the veridical localization of the sound source are known as the precedence effect (Blauert 1997). Moreover, the auditory system achieves timbral constancy (i.e., the perceptual invariance of a source) despite variations in the spectral envelope of the sound caused by room reverberation (Watkins 1991, 1998, 1999; Watkins and Makin 1996).

Just as the inferred direction of a light source allows the visual system to compute the altitude of the green square, the information available to the auditory system is not lost; it's recycled and used to characterize the size of the space in which the sound is produced. When we listen to a recording of a sound, we can tell whether the sound was recorded in a gym, a restroom, a classroom, or a small lab room (Robart and Rosenblum 2005). The use of this information by the auditory system depends on a rapidly constructed internal model of the environment. When a lagging sound is inconsistent with the prevailing internal model, the precedence effect will fail, and it will be heard as an echo (Clifton et al. 1994, 2002).

To paraphrase Mollon (1995):

> our auditory system is not built to recognize the acoustic flux but permanent properties of objects, i.e., the nature of the sound they produce by their own activity if they're animate, or by their audible response to physical energy if they're not.

The foregoing considerations suggest that with respect to the concepts of source and surface, vision and audition are duals (Table 1). The notion of dual is hard to define, because different fields of mathematics have different versions of this concept. It may suffice to give an example. Imagine that in Fig. 3 the black drawing represents a map of four countries on an island. In this map no two countries that share a boundary, or the sea, should have the

Table 1 The first duality of vision and audition

Modality	Sources	Surfaces
Audition	Primary	Secondary
Vision	Secondary	Primary

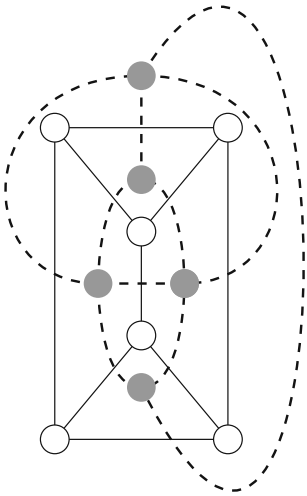

Fig. 3 An example of dual graphs

same color. One step toward solving such coloring problems is to construct the *dual graph* of this map. In each country and in the sea we place a vertex (in gray); we connect only vertices that correspond to countries that share a boundary. All the countries have a coast-line, so the sea-vertex is connected to all the country-vertices. For each country with i sides there will be a gray vertex that is connected to i vertices. For each vertex (white dot) connected to j vertices in the map there will be a face with j sides in the dual graph. Thus when we go from the map to its dual, we exchange the roles of *vertex* and *side*. Similarly when we go from audition to vision, we exchange the roles of *source* and *surface*.

2 The Audio-Visual Linkage

In the remainder of this article we will frequently refer to Fig. 4 which summarizes our view of the relations between vision and audition, and form what we call the *audio-visual linkage*. The figure is divided into two regions: audition on the left and vision on the right. In the lower left and right corners of these regions we reiterate the conclusion reached in the preceding section: that audition and vision are concerned with different aspects of the world. To reinforce this observation, we perform two thought experiments.

2.1 The Theory of Indispensable Attributes

2.1.1 The Visual Thought Experiment

The first thought experiment is summarized in Fig. 5. We begin with the situation depicted in the left-hand panel. There are two spotlights, and each

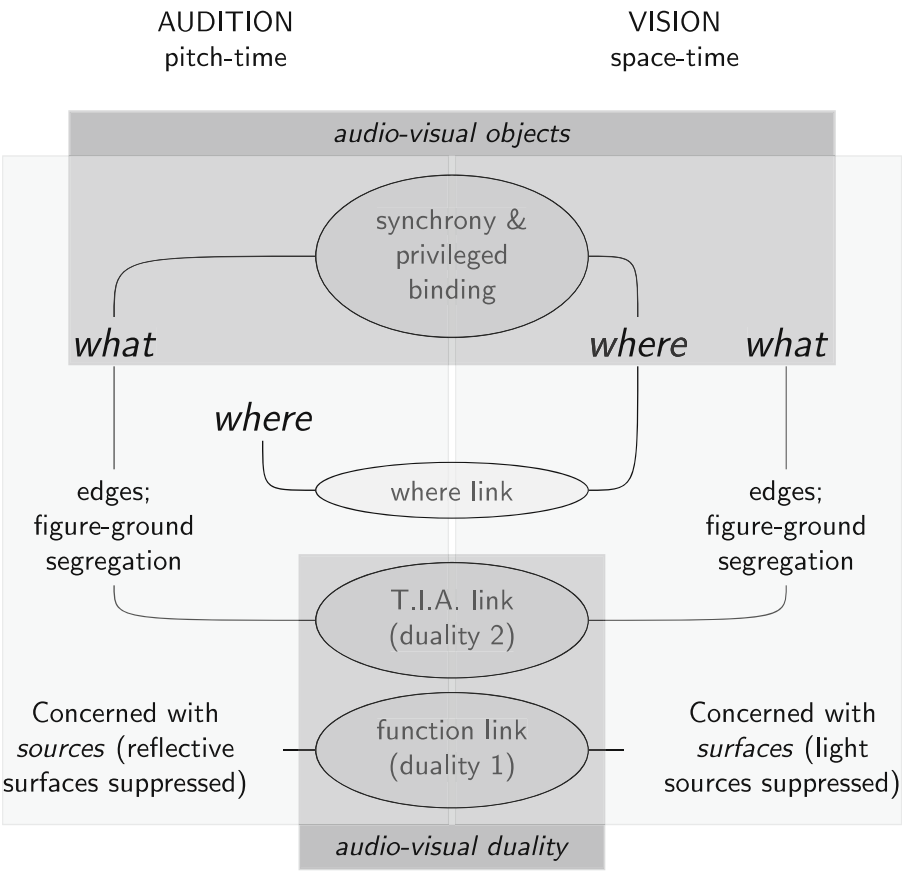

Fig. 4 Audio-visual objects and the audio-visual linkage. T.I.A. stands for theory of indispensable attributes

emits a beam of light consisting of a single wavelength. The two beams are cast onto a light surface, creating two spots, which are seen as red and green. The first step in the experiment is to obtain an observer's description of the illuminated surface. *If and only if* the observer describes the situation in a manner that expresses the proposition, "I see two spots of different colors," the experiment can proceed. (This precondition was not made sufficiently clear in previous expositions, which has led to misunderstandings.) For example, suppose that rather than being circular and separate (●●), the spots were square and shared a side (■). It would not be unreasonable for the observer to describe this as one rectangular spot with two parts, one red and one green. If this happens, change the display (or replace your observer if you suspect he's idiosyncratic, uncooperative, or lazy). Furthermore, if the colors were not easy to distinguish (say their wavelengths were 520 and 530 nm), the observer might consider the colors to be the same; this too would vitiate the experiment. We must have no doubt that the observer *sees* that the two spots are separate

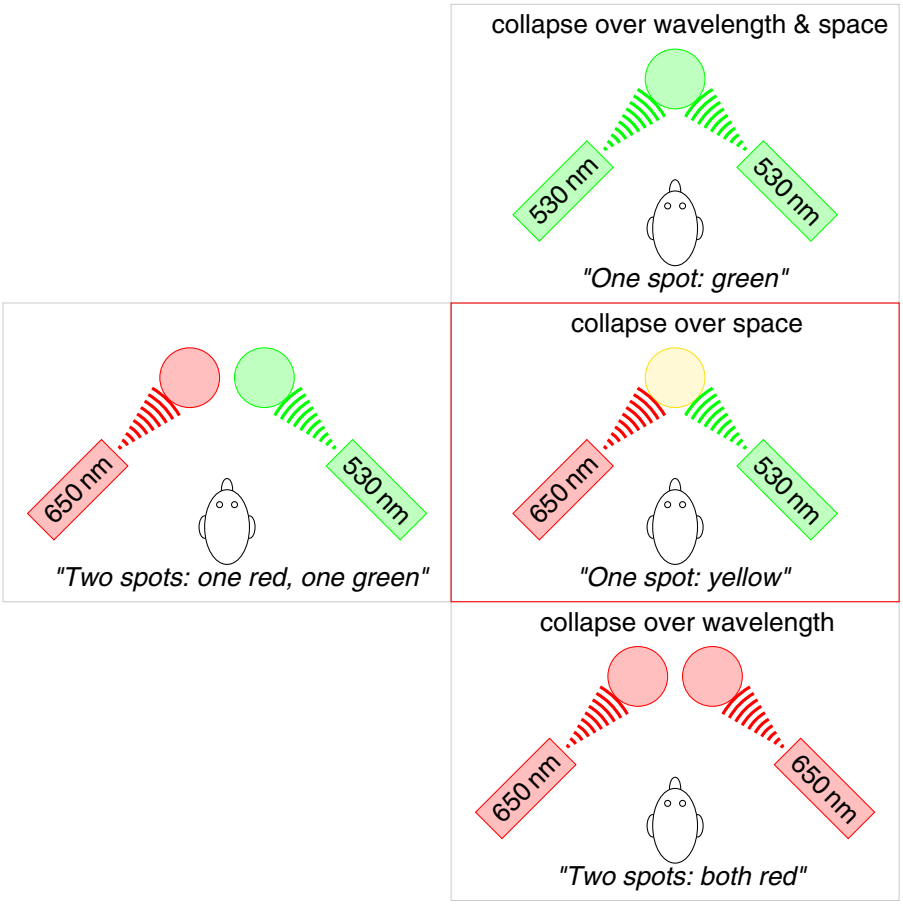

Fig. 5 TIA for vision: spatial separation is an indispensable attribute for visual numerosity; wavelength is not. Each disc represents a projected spot of light. The *top part* of each panel represents the physical conditions of the thought-experiment. The statement at the *bottom of each panel* represents the proposition entertained by the observer. The predicted statements in the *right-hand panels* are conditional on (1) the observer having entertained the proposition on the left when the two projectors projected two spots of light of different wavelengths, and (2) the experimenter having maintained *ceteris paribus* when performing the collapse(s) indicated at the *top of each panel*

and of different colors, which is a more stringent requirement than setting up the display so that the two spots *are* physically separate and of different wavelengths.

Once the precondition has been satisfied we perform an operation which we call *collapsing with respect to wavelength*, depicted in the bottom right-hand panel. It simply involves changing the wavelength emitted by the right-hand projector to be the same as the wavelength emitted by the left-hand projector, and asking the observer what she sees. If any numerosity remains

in her judgment, we will say that wavelength is not indispensable for visual numerosity. There is no doubt that our observer will still report seeing two spots.

The middle right-hand panel depicts *collapsing with respect to space*. The beams are of different wavelengths, but now we have contrived to have their projected spots coincide perfectly. The mixture of two lights in this manner is called additive color mixture. It is known that the additive mixture of these two wavelengths will be called yellow. Because of a phenomenon of color vision called *color metamerism* this yellow is indistinguishable from a yellow that is not a mixture, a so-called spectral yellow. In other words when the mixed yellow is seen, it contains no perceivable trace of either red or green. So our observer's description of this displays will have lost its numerosity. She will say something to the effect that she sees one yellow spot.

For the sake of completeness we show in the upper right-hand panel the consequence of collapsing with respect to wavelength and space. The indispensability of space is inherited by this situation: numerosity is lost.

2.1.2 The Auditory Thought Experiment

Our second thought experiment is depicted in Fig. 6. Here too we begin with the situation depicted in the left-hand panel. Now two loudspeakers are emitting two keyboard notes, a C and F. Here it is not unlikely that the listener will describe what she hears as a dyad, perhaps observing that it was played over two loudspeakers. In such a case the preconditions of the experiment are not satisfied, because from the outset numerosity has been lost. So we assume that the listener has said that she heard two notes coming from two loudspeakers. Furthermore, if the notes were not different enough (say their frequencies were 262 and 267 Hz), the observer might hear a single beating tone; this too would vitiate the experiment. Or if the two loudspeakers were not far enough apart the observer might not notice that the tones are coming from different directions. We must have no doubt that the observer *hears* that the two tones come from different locations and differ in pitch, which is a more stringent requirement than setting up the situation so that the two loudspeakers *are* physically separate and emit different frequencies.

Now we can *collapse with respect to frequency*. If the two loudspeakers are equidistant from the listener, theories of auditory localization tell us that she will hear a single note, coming from a location between the speakers. If not, the precedence effect will cause her to hear a single sound coming from the right or the left speaker. But she will not experience twoness. Thus frequency is an indispensable attribute for auditory numerosity.

If we *collapse with respect to space*, she will hear two notes coming from a single speaker, thus preserving numerosity, and showing that space is not an indispensable attribute for auditory numerosity. For completeness we also show the case in which we collapse over both space and frequency, which inherits the loss of auditory numerosity from the collapse with respect to frequency.

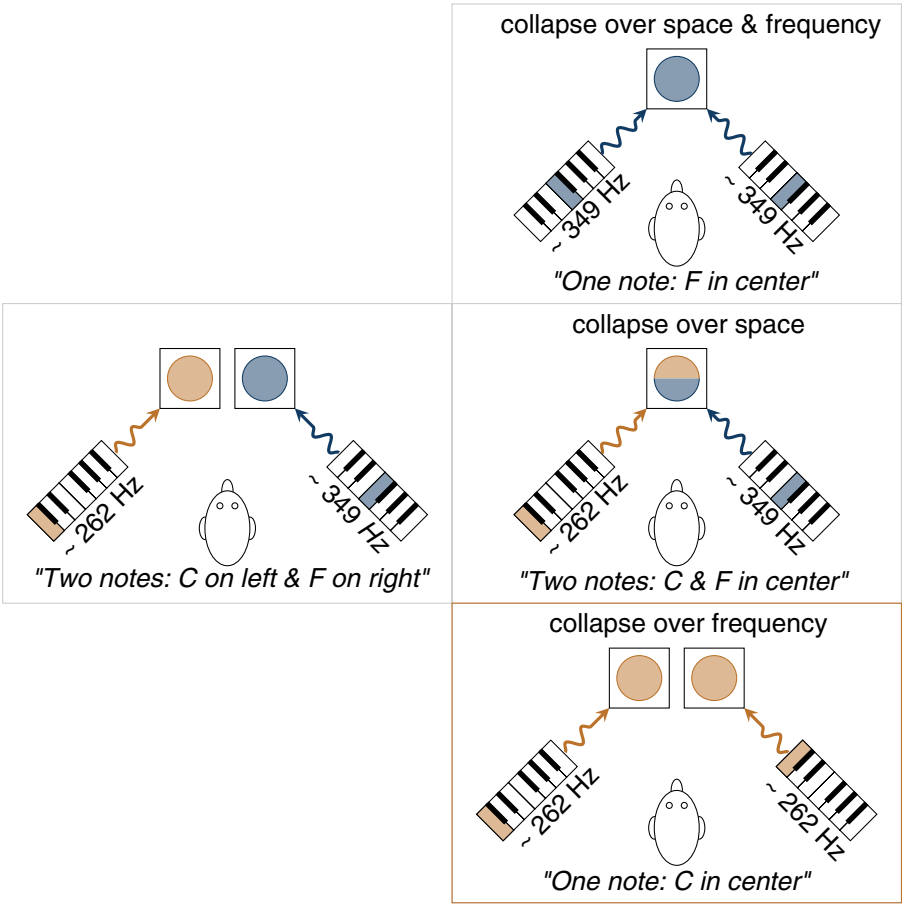

Fig. 6 TIA for audition: frequency is an indispensable attribute for auditory numerosity; spatial separation is not. Each disc in a square represents a loudspeaker. The *top part of each panel* represents the physical conditions of the thought-experiment. The statement at the *bottom of each panel* represents the proposition entertained by the observer. The predicted statements in the *right-hand panels* are conditional on (1) the observer having entertained the proposition on the left when the two frequencies were played over different loudspeakers, and (2) the experimenter having maintained *ceteris paribus* when performing the collapse(s) indicated at the *top of each panel*

2.2 Audio-Visual Duality Amplified

To compare our two thought experiments, we look at the middle right-hand panels of Figs. 5 and 6 in which we represent collapsing with respect to space. Only in vision does a reduction in numerosity ensue (indicated by a red frame); in audition the reduction in numerosity occurs only when we collapse with respect to frequency (also indicated by a red frame).

These conclusions are consistent with what we said earlier about the different functions of vision and audition. The visual world is spread out over space; color is useful, but not essential. Witness the relatively mild consequences of color blindness, and the effectiveness of black-and-white movies. In contrast, the important characteristics of an auditory source are not spatial; they reside by and large in its timbre. The auditory world is spread out over frequency (and timbre, which involves harmonics that are spread over frequency, and envelopes that are modulated in time, and deserves a more thorough analysis than we can give here); stereophonic hearing is useful, but not essential. Witness the relatively mild consequences of unilateral deafness, and the effectiveness of monophonic recordings.

For the sake of efficiency we have postponed our discussion of the role of time. Suffice it to say that one can show, by replacing space with time in the visual thought experiment, and pitch with time in the auditory thought experiment, that time turns out to be an indispensable attribute for both modalities, an Aristotelian common sense as it were.

Armed with the theory of indispensable attributes we can add another layer to the duality of vision and audition (Table 2).

Kubovy and Van Valkenburg (2001) have argued that because indispensable attributes are the spaces that can sustain a manifold of entities (i.e., numerosity), they must also be the spaces that sustain objecthood. The noun *object* comes from Latin by way of Medieval Latin. *Objectus* is derived from *ob-* in the way + *jacere* to throw. According to the Oxford English Dictionary (2004), *object* originally meant "something placed before or presented to the eyes or other senses." Now it means "a material thing that can be seen and touched, and "the presentation of something to the eye or perception."

If we relied on folk-ontologically based lexicography, we could speak of visual or tactile objects, but not of auditory ones. Kubovy and Van Valkenburg (2001) proposed an alternative definition, which is not inconsistent with the dictionary definitions, but allows us to readily speak of auditory objects. It is: "A *perceptual object* is that which is susceptible to figure-ground segregation" (p. 102).

Their view on the passage from indispensable attributes to objects is summarized in Fig. 7, where the processes underlying perception are divided into three classes, corresponding roughly to a movement from peripheral to central processing, from early to late stages of processing. From our point of view the important consequence of this peripheral to central and early to late contrasts

Table 2 The parallel dualities of vision and audition

Modality	Functional duality		T.I.A. duality	
	Primary	Secondary	Indispensable	Dispensable
Audition	Sources (frequency)	Surfaces (space)	Frequency (sources)	Space (surfaces)
Vision	Surfaces (space)	Sources (wavelength)	Space (surfaces)	Wavelength (sources)

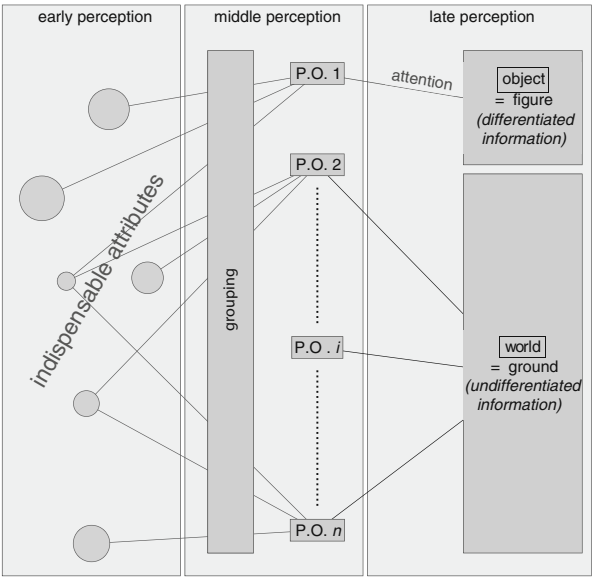

Fig. 7 Objects and attention. P.O. = putative object

is this: early perception is bottom-up, late perception is top-down, and middle perception is both.

The operations of early processing are thought to be sub-personal and cognitively impenetrable, that is, intentions and propositional attitudes cannot affect them. For our purposes the most important aspect of early perception is the detection and isolation of scene fragments. Early vision detects and isolates the blobs in Figs. 8 and 9.

Fig. 8 Dalmatian dog

Fig. 9 Dalmatian dog sniffing the ground facing left and away from us (with blurred background)

Middle-level perception is akin to respiration: it usually runs bottom-up, but can come under top-down control. Take, for example, the complex ambiguous figure in Fig. 10 (Bradley and Petry 1977). With thick white lines (segments of which are illusory) we draw the so-called Necker cube, which can be interpreted either as a wire-frame cube seen from above right, or a wire-frame cube seen from below left. A second level of ambiguity is introduced when we realize that the eight black dots can be seen as black discs on a sheet behind a wire-frame cube, or as holes in a white sheet through which we can see the white wire-frame cube against a black background. Some of these interpretations will occur to you spontaneously, bottom-up, and some you can control, top-down. It is here that Gestalt laws of grouping apply: a sequence of notes can become a melody, collections of dots can coalesce into organized patterns (Fig. 11). When multiple organizations are available to be experienced, one of them may be selected by a bottom-up process with-

Fig. 10 A doubly ambiguous figure (Bradley and Petry 1977)

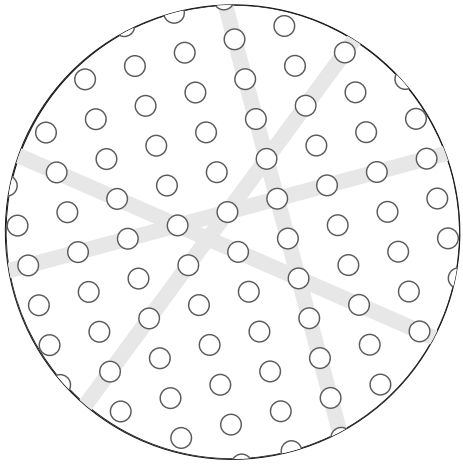

Fig. 11 A pattern (a rectangular dot lattice) that can be grouped in four different ways

out voluntary intervention. These multiple organizations are *putative objects*, because once they win the competition for our awareness, they become objects. Middle-level vision begins to pull together the sharp-edged blobs as in Fig. 9 relying on alignments that are not likely to be accidental.

This is where high-level perception comes into play. Whether the selection of an organization as an object occurred spontaneously, or was affected by attention, it is now differentiated from all other entities. It has undergone *figure-ground segregation*, which is the hallmark of objecthood, whether auditory or visual. High-level vision uses world-knowledge to intervene in the grouping process, and eventually identifies it as an object: a spotted dog on a dappled background (Figs. 8 and 9). If the breed is known to the perceiver, she will think "Dalmatian."

2.3 What and Where

Kubovy and Van Valkenburg (2001) review the evidence that both audition and vision have "what" and "where" subsystems. The primary function of the visual "what" subsystem is to recognize objects, and serve our social interactions, such as recognizing faces and facial expressions. The primary function of the visual "where" subsystem is to locate objects in space for the purpose of locomotion and other forms of action. The case for this distinction was famously made by Milner and Goodale (1995), and although it been weakened somewhat, by and large it has survived the test of many critical experiments. The primary function of the auditory "what" subsystem is also to recognize the nature of physical events (a glass shattering), and serve our social interactions (recognizing the voice of a friend).

The most important function of the auditory "where" subsystem is to produce an orienting response, the effect of which is to direct the gaze toward

the source of sound (Bon and Lucchetti 2006; Goldring et al. 1996; Hötting et al. 2003; Witten et al. 2006). In other words, the auditory "where" subsystem is primarily in the service of the visual system.

Returning now to Fig. 4, we see that the duality links are links between the auditory and the visual "what" subsystems, which constitute the outer turn of our audio-visual linkage. In contrast, the space link connects the two "where" subsystems in the manner we have just described. We turn to the space link.

3 Audio-Visual Objects

3.1 The Question of Binding

The "where" and the "what" links of the audio-visual linkage provide us with the tools we need to address Gibson's question which opened this article. We can now put the question in the following terms: how do the "what" and the "where" subsystems of vision bind their information with the information provided by the "what" subsystem of audition to form an audio-visual object?

The evidence for binding can be of two kinds: phenomenological or experimental. To use phenomenological evidence we would need a clear criterion for saying when an acoustic event and a visual event were bound to form a single audio-visual object. Some cases are clear. The sound and sight of a glass shattering leave no doubt in our minds that what we heard and saw was caused by the same physical event. We cannot prevent the binding of the ventriloquist's voice with the movement of the dummy's lips, even though we know it's an illusion. We do not experience the flash of lightning and the sound of thunder as one event, even though we know that they were both caused by same electrical discharge.

Unfortunately phenomenology does not give us quantitative information. It cannot tell us how strong binding is, which prevents us from undertaking a systematic phenomenological study of the effects of spatial separation (as in the case of ventriloquism) or the effects of temporal separation (as in the case of an atmospheric electrical discharge) on the experience of cross-modal binding.

When the direct approach of phenomenology fails, we turn to indirect indicators of binding. When we see and hear a person speaking, we do not experience an auditory and a visual event, but one. We can put this binding to the test by exploring the limits of ventriloquism. If ventriloquism succeeds we experience one unremarkable event: the sound comes from the speaker's mouth. Unless we pay particular attention to the speaker's facial movements, we do not see a jaw wagging, lips opening and closing, and tongue gestures. If it fails, we experience the auditory and the visual

unbound, we experience perceptual numerosity: we see lip movements and hear speech, and we localize them in difference places. The synchronization of speech sounds with the speaker's mouth movements is necessary for ventriloquism; the effect is notably reduced with a delay as short as 0.2 s between the visual and the auditory. On the other hand, the effect is not much decreased when the sound source and the speaker's mouth are 30° apart (Jack and Thurlow 1973).

We discover that binding occurs by finding the conditions under which it fails. To do that we must either present consistent audio-visual information in an inconsistent manner, or inconsistent information in a consistent manner. Ventriloquism is an example of the former: the lip movements and speech sounds are consistent. The McGurk Effect (McGurk and MacDonald 1976) is an example of the latter: if an auditory /b/ is dubbed onto a visual /g/, listeners perceive a fused phoneme, a /d/. With the reverse presentation, they experience a combination such as /bg/.

Studies of audio-visual binding have often focused on asymmetries such as ventriloquism, where the location of the sound is captured by the location of its apparent visual source. Here visual location trumps auditory location in binding, as indispensable attributes might lead us to expect. We hope some day to generalize this observation: if (1) two modalities, X and Y can process a stimulus attribute A, (2) the information about A from X and Y is in conflict, and (3) A is indispensable for X and not for Y, then X will trump Y with respect to A. Location also affects judgments of audio-visual simultaneity. Bertelson and Aschersleben (2003) had observers judge whether a burst of sound or a flash of light came first. They manipulated the temporal and spatial separation between the acoustic and optical events. The observers' judgments of the temporal order of the events were better when the sound and the flash coincided in space than when they were apart.

In matters of time, however, the auditory often has greater weight. Aschersleben and Bertelson (2003) created a temporal analog of ventriloquism: they asked observers to tap in time with a regular pulse-train of flashes of light and ignore a sound that preceded (or followed) each flash. Despite these instructions, the observers' taps gravitated toward the sounds. When the role of visual and auditory pulse was reversed, and the observers were asked to keep time with a pulse-train of sounds while ignoring a flash that preceded (or followed) each sound, the taps gravitated toward the flashes much less than in the previous experiment. This and other experiments have led researchers to conclude that audition plays a greater role than vision in the processing of temporal information.

Although both temporal simultaneity and spatial coincidence facilitate audio-visual binding, an experiment by Schutz and Lipscomb (2007) has led us to a new hypothesis about its nature: binding depends on the ecological fit between of visible events and sounds. For example, the sound of a marimba binds with the visible impact because such an impact could have produced such a percussive sound.

Fig. 12 The video showed the upper body of the marimbist, and included the performer's stroke preparation and release (reprinted from Schutz and Lipscomb 2007, Fig. 1, with permission of Pion Limited, London)

3.2 The Discovery of Privileged Binding

The seminal experiment, conducted at the Northwestern University School of Music, where Schutz was studying, originated as an attempt to resolve an ongoing debate among marimba players: Can a marimbist's gesture affect the duration of the sound produced by the impact of the mallet on the wood key? Schutz and Lipscomb (2007) recorded a sound video of a world-class marimbist who believed that it could (Fig. 12). Their study had two phases. In one they did not show the video, they just played sounds produced by gestures intending to produce long sounds (L-gestures), and sounds produced by gestures intending to produce short sounds (S-gestures). The observers–undergraduate students of music, none specializing in percussion—were asked to rate the duration of each sound in the absence of the video, using a rating scale shown on the screen of a computer (Fig. 13). In the second phase they were asked to rate the duration of the same sounds in the presence of the video, *disregarding the video*, having been informed about possible mismatches between the audio and the video.

The marimbists who thought that they could affect the duration of the sound were wrong. When the participants only heard the sounds, the fact that they had been produced by different gestures had no effect on the perceived duration of the sounds. This was not merely a failing of the listening skills of the

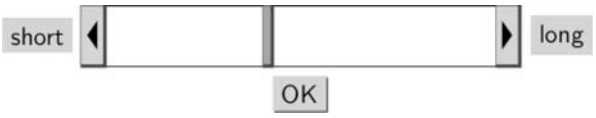

Fig. 13 The scale with which participants rated the relative duration of sounds. They clicked at the point they wanted to place the slider until they were satisfied with their rating. They then clicked on the "OK" button, which triggered the next trial

participants: they found no consistent measurable physical difference between the sounds.

But even though this school of marimba performance was wrong, the results produced an ironic twist in their favor. When the observers heard the sound while seeing the marimbist perform it, the gesture affected the perceived duration of the sounds: when the sounds were accompanied by the video, the participants gave much higher ratings of duration for the L-gestures than for the S-gestures.

Such an effect of visual information on a judgment of duration is not in keeping with the widespread view that auditory temporal information should trump visual temporal information in audio-visual binding. This inconsistency with the prevailing consensus led us to ask two questions: (1) Could these results be anything other than an effect of visual information on a temporal judgment? (2) If they are not, are they merely the effect of the simultaneity of the visible impact and the audible impact?

The key experiment for our story involved a replication of the Schutz and Lipscomb experiment, but with several new sounds in addition to those of the marimba (Schutz and Kubovy 2009a). We added one percussion instrument, a piano sound, and several non-percussion instruments, such as a sung tone, and notes played by a clarinet, a trumpet, and a burst of white noise. We found an effect of the video on the judged duration of the sounds only with the two percussion instruments. If the effect of the video on the perceived duration is due to binding, then it cannot just be due to the simultaneity of the seen moment of impact and its acoustic effect, since the seen moment of impact was also simultaneous with the onset of the non-percussive sounds. Indeed when people hear the sound of a voice at the moment of the impact of the mallet, the combination is so incongruous that it can elicit laughter, suggesting an experience of failed binding.

In order to determine the power of this privileged binding of the two manifestations of an impact, we also studied the effect of synchrony on the Schutz and Lipscomb phenomenon (Schutz and Kubovy 2009a). We found that if the marimba sound preceded the impact, there was no effect of gesture on perceived duration. But if it was simultaneous or even delayed by 0.4 s, the effect was large. We even obtained a measurable effect with a delay of 0.7 s. (Recall that ventriloquism vanishes with an 0.2 s delay.)

We examined the nature of the visual information required for privileged binding (Schutz and Kubovy 2009b). We simplified the video by creating a skeleton performer with three moving joints (shoulder, elbow, and wrist), and a small disk that tracked the head of the mallet (Fig. 14). In a series of experiments we asked how much information is required for the binding to occur. We reduced the number of joints one by one, and each time found the same visual effect, even when only the head of the mallet remained. In an unpublished study we even found that *if we did not move the position of disk that represents the head of the mallet, its duration still affected the perceived duration of the sound*. So the visual information required for the binding is quite abstract. We have made progress toward understanding this abstraction

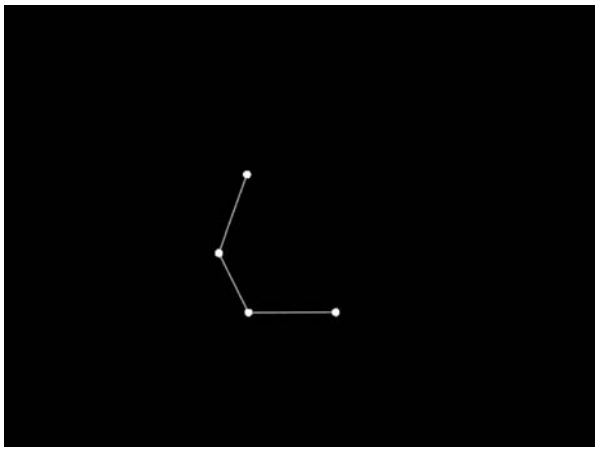

Fig. 14 The skeleton player showed the upper body of the marimbist, and included the performer's stroke preparation and release

by showing that the illusion is driven primarily by the post-impact portion of the gesture. More specifically, it is driven by post-impact motion duration rather than post-impact velocity, acceleration, or distance (Armontrout et al. 2009).

3.3 Philosophical Comments

A reviewer of this article reminded us that important contemporary philosophers (Campbell 2002; Matthen 2004, 2005) hold that perceptual experiences refer to physical objects or changes in these objects understood as external and mind-independent things. They are not mental products. We agree, and do not intend our concept of audio-visual object to undermine the notion of a mind-independent physical objects. Indeed the notion of privileged binding is meaningless unless we know kinematics and dynamics: when object of type O strikes surface of type S, it is likely to have travelled along a visible trajectory of type T, and produce (cause) a audible sound of type A. When we act as stimulus designers we assume that we can step out of our phenomenal world and manipulate apparent physical properties of the quadruple $< O, S, T, A >$ as they are represented on a computer screen. Furthermore, as experimenters we are confident in the mind-independence of our measurement of the intensity of a sound: we know enough about the psychophysical correspondence between sound intensity and loudness to predict to a good first approximation the loudness of a particular sound to any listener.

Even though we perceive the world and not percepts, we cannot dispense with mind-dependent concepts, and indeed entities. We just referred to the indispensable distinction between sound intensity and loudness. It is always useful to keep in mind that cognitive scientists rely on manipulations of the physical world to produce phenomenal effects. Our experimental observers

experience these effects as objects and events in the world; phenomenally they are *in* the environment.

Thus audio-visual objects are constructs of the mind—they are the end-product of a process that operates on sensory information, and attempts to produce the most plausible reading of this information as caused by objects and events taking place in the environment. Adaptation has insured that the errors committed by this process are generally inconsequential.

3.4 Conclusion

We summarize the argument about privileged binding in Fig. 15. It shows in what way we have a more complete understanding of the formation of audio-visual objects, the central part of the audio-visual linkage. We know

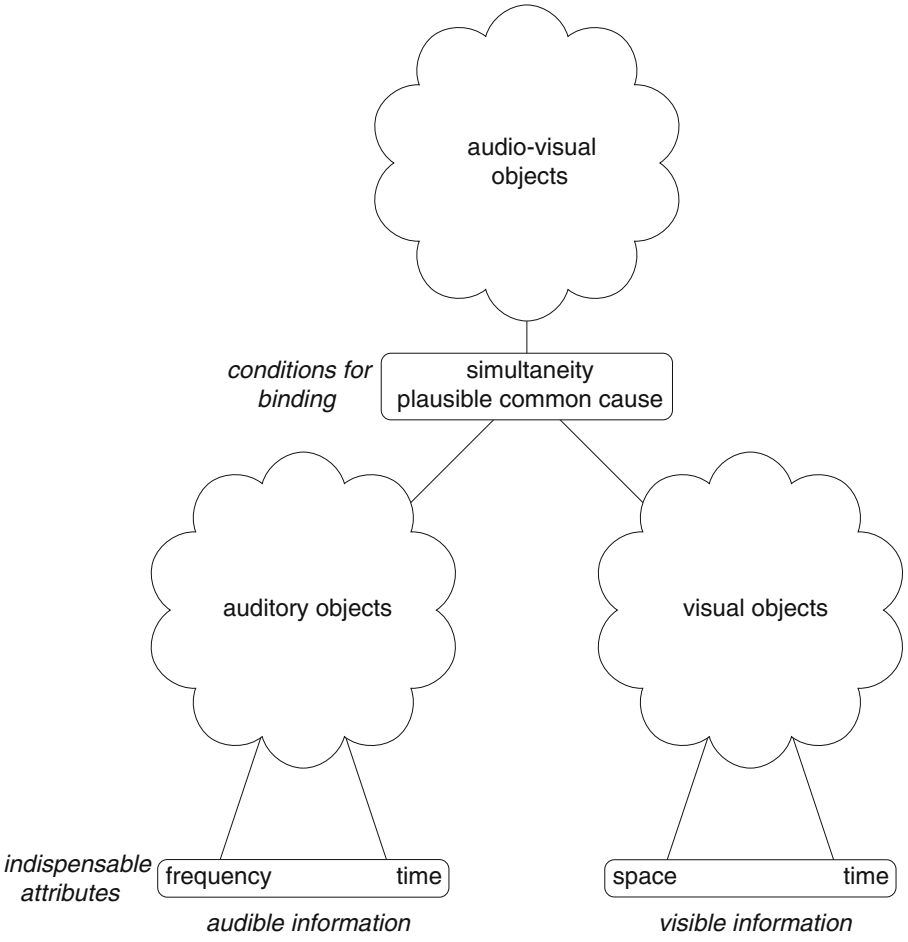

Fig. 15 The formation of audio-visual objects

that synchrony is important, but only when arbitrary sounds and visual events are bound into audio-visual objects. We believe that the most important kind of binding is between ecologically-related auditory and visual information. Gibson was critical of theories that allow arbitrary associations between different kinds of stimulation. But his ideas about cross-modal binding constitute a research program, not a theory. We have now taken a first step toward that theory.[1]

References

Amano, K., D.H. Foster, and S.M.C. Nascimento. 2006. Color constancy in natural scenes with and without an explicit illuminant cue. *Visual Neuroscience* 23: 351–356.

Armontrout, J.A., M. Schutz, and M. Kubovy. 2009. Visual determinants of a cross-modal illusion. *Attention, Perception, & Psychophysics* 71: 1618–1627.

Aschersleben, G., and P. Bertelson. 2003. Temporal ventriloquism: Crossmodal interaction on the time dimension—2. Evidence from sensorimotor synchronization. *International Journal of Psychophysiology* 50(1–2): 157–163.

Bertelson, P., and G. Aschersleben. 2003. Temporal ventriloquism: Crossmodal interaction on the time dimension—1. Evidence from auditory-visual temporal order judgment. *International Journal of Psychophysiology* 50(1–2): 147–155.

Blauert, J. 1997. *Spatial hearing: The psychophysics of human sound localization*. Cambridge: MIT (revised edition).

Bon, L., and C. Lucchetti. 2006. Auditory environmental cells and visual fixation effect in area 8B of macaque monkey. *Experimental Brain Research* 168(3): 441–449.

Bradley, D.R., and H.M. Petry. 1977. Organizational determinants of subjective contour: The subjective Necker cube. *American Journal of Psychology* 90: 253–262.

Campbell, J. 2002. *Reference and consciousness*. Oxford: Oxford University Press. Published to Oxford Scholarship Online: November 2003. doi:10.1093/0199243816.001.0001. Accessed: 25 December 2007.

Clifton, R.K., R.L. Freyman, R.Y. Litovsky, and D. McCall. 1994. Listeners' expectations about echoes can raise or lower echo threshold. *The Journal of the Acoustical Society of America* 95(3): 1525–1533.

Clifton, R.K., R.L. Freyman, and J. Meo. 2002. What the precedence effect tells us about room acoustics. *Perception & Psychophysics* 64(2): 180–188.

Gibson, J.J. 1966. *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Goldring, J., M. Dorris, B. Corneil, P. Balantyne, and D. Munoz. 1996. Combined eye-head gaze shifts to visual and auditory targets in humans. *Experimental Brain Research* 111: 68–78.

Hötting, K., F. Rösler, and B. Röder. 2003. Crossmodal and intermodal attention modulate event-related brain potentials to tactile and auditory stimuli. *Experimental Brain Research* 148: 26–37.

Jack, C.E., and W.R. Thurlow. 1973. Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual & Motor Skills* 37: 967–979.

Kubovy, M., and D. Van Valkenburg. 2001. Auditory and visual objects. *Cognition* 80(1–2): 97–126.

Matthen, M. 2004. Features, places, and things: Reflections on Austen Clark's theory of sentience. *Philosophical Psychology* 17(4): 497–518.

Matthen, M. 2005. *Seeing, doing, and knowing—a philosophical theory of sense perception*. Oxford, UK: Oxford University Press. Published to Oxford Scholarship Online: April 2005. doi:10.1093/0199268509.001.0001. Accessed: 25 December 2007.

[1]The concept of privileged binding is similar to Stoffregen and Bardy's (2001) notion of *global arrays*. We do not agree that their approach (and *a fortiriori* ours) undermines the idea of separate senses; a justification of this disagreement would go beyond the scope of this article.

McGurk, H., and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264(5588): 746–748.
Milner, A.D., and M.A. Goodale. 1995. *The visual brain in action*. Oxford: Oxford University Press.
Mollon, J. 1995. Seeing colour. In *Colour: Art & science*, eds. T. Lamb, and J. Bouriau, 127–150. Cambridge: Cambridge University Press.
Oxford English Dictionary. 2004. Object. Retrieved on 25 December 2006 from the Oxford english dictionary. Online: http://dictionary.oed.com/cgi/entry/00329075.
Robart, R.L., and L.D. Rosenblum. 2005. Hearing space: Identifying rooms by reflected sound. In *Studies in perception and action XIII*, eds. H. Heft, and K.L. Marsh, 153–161. Hillsdale: Lawrence Erlbaum.
Schutz, M., and M. Kubovy. 2009a. Causality and cross-modal integration. *Journal of Experimental Psychology: Human Perception and Performance* 35(6):1791–1810.
Schutz, M., and M. Kubovy. 2009b. Deconstructing a musical illusion: Point-light representations capture salient properties of impact motions. *Canadian Acoustics* 37: 23–28.
Schutz, M., and S. Lipscomb. 2007. Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception* 36: 888–897.
Stoffregen, T.A., and B.G. Bardy. 2001. On specification and the senses. *Behavioral and Brain Sciences* 24: 195–213.
Watkins, A.J. 1991. Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America* 90: 2942–2955.
Watkins, A.J. 1998. The precedence effect and perceptual compensation for spectral envelope distortion. In *Psychophysical and physiological advances in hearing*, eds. A. Palmer, A. Rees, A.Q. Summerfield, and R. Meddis, 336–343. London: Whurr.
Watkins, A.J. 1999. The influence of early reflections on the identification and lateralization of vowels. *Journal of the Acoustical Society of America* 106: 2933–2944.
Watkins, A.J., and S.J. Makin. 1996. Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America* 99: 3749–3757.
Witten, I.B., J.F. Bergan, and E.I. Knudsen. 2006. Dynamic shifts in the owl's auditory space map predict moving sound location. *Nature Neuroscience* 9(11): 1439–1445.