ELSEVIER

Cognition 80 (2001) 97–126

COGNITION

www.elsevier.com/locate/cognit

Auditory and visual objects

Michael Kubovy*, David Van Valkenburg

University of Virginia, Charlottesville, VA, USA

Received 8 December 1999; accepted 17 November 2000

Abstract

Notions of objecthood have traditionally been cast in visuocentric terminology. As a result, theories of auditory and cross-modal perception have focused more on the differences between modalities than on the similarities. In this paper we re-examine the concept of an object in a way that overcomes the limitations of the traditional perspective. We propose a new, cross-modal conception of objecthood which focuses on the similarities between modalities instead of the differences. Further, we propose that the auditory system might consist of two parallel streams of processing (the 'what' and 'where' subsystems) in a manner analogous to current conceptions of the visual system. We suggest that the 'what' subsystems in each modality are concerned with objecthood. Finally, we present evidence for – and elaborate on – the hypothesis that the auditory 'where' subsystem is in the service of the visual-motor 'where' subsystem.

Keywords: Auditory; Visual; Objecthood

1. Introduction

In this article we argue for the concept of an *auditory object*. Although some have found such a concept so strange that they avoid the term altogether in favor of 'auditory event' (Blauert, 1997, p. 2), we are convinced that it is both a useful and important concept. To clarify it, we offer a distinction between an auditory 'what' subsystem and an auditory 'where' subsystem (in a manner analogous to

* Corresponding author. Department of Psychology, P.O. Box 400400, University of Virginia, Charlottesville, VA 22904-4400, USA. Fax: +1-804-982-4766.

E-mail addresses: kubovy@virginia.edu (M. Kubovy), dlv6b@virginia.edu (D. Van Valkenburg).

0010-0277/01/$ - see front matter
PII: S0010-0277(00)00155-4

Milner & Goodale, 1995), and argue that the 'what' subsystem forms auditory objects, and that the 'where' subsystem is in the service of vision.

The bias against the idea of auditory objecthood is embedded in folk ontology. Language itself[1] may lead us to believe that objects are visible by definition. For example, according to the *Oxford English Dictionary*, *object* means "Something placed before the eyes, or presented to the sight or other sense; an individual thing seen or perceived, or that may be seen or perceived; a material thing" (Object, 1993). The etymology of the word *object* explains the visuocentric connotation of the word: it derives from the Latin *ob-*, 'before' or 'toward', and *iacere*, 'to throw'. It used to mean, "Something 'thrown' or put in the way, so as to interrupt or obstruct the course of a person or thing; an obstacle, a hindrance" (Object, 1993). Indeed, most visible things are obstacles or a hindrance to sight; they prevent you from seeing something that lies behind them because they are opaque.[2]

In this paper we will deviate from our everyday notion of object in order to extend it to audition. We will do this by finding a different criterion for objecthood, one that does not rely on the notion of opacity. We must do this because the notion of opacity simply does not apply to auditory perception. Material things can of course be opaque to sound (Beranek, 1988, Chapter 3). But we do not listen to *material* things, we listen to *vibrating* things – *audible sources*. One sound source does not in general prevent you from hearing another: many natural sounds, especially biological ones, are composed of a fundamental frequency and discrete harmonics – i.e. they are sparse, like fences. Furthermore, masking is rare in nature because the masking sound must be considerably louder than the masked one (e.g. it takes the sound of a waterfall or thunder to mask our voices).

Although one sound can mask another, Bregman (1990), in his discussion of the auditory continuity illusion, shows that audible sources do not offer a natural analog to opacity. The auditory continuity illusion is created when one deletes part of a signal and replaces it with a louder sound: the signal is perceived to continue uninterrupted 'behind' the sound. Bregman compares this illusion with the visual experience of continuity behind an occluder (Fig. 1): "Let us designate the interrupted sound or visual surface as A, and consider it to be divided into A1 and A2 by B, the interrupting entity… [In vision one] object's surface must end exactly where the other begins and the contours of A must reach dead ends where they visually meet the outline of B. In the auditory modality, the evidence for the continuity occurs in the properties of B itself as well as in A1 and A2; B must give rise to a set of neural properties that contain those of the missing part of A. In vision, on the other hand, if objects are opaque, there is no hint of the properties of A in the visual region occupied by B" (p. 383).

We pointed out earlier that we do not listen to material things, but to audible sources. The auditory system is generally concerned with *sources* of sound (such as speech or music), not with *surfaces* that reflect the sound (Bregman, 1990, pp. 36–

[1] Indo-European languages in particular.

[2] There are two exceptions: things that are transparent and things that are sparse. There are two kinds of sparse things: things with holes in them (e.g. fences) and things that branch (e.g. plants).

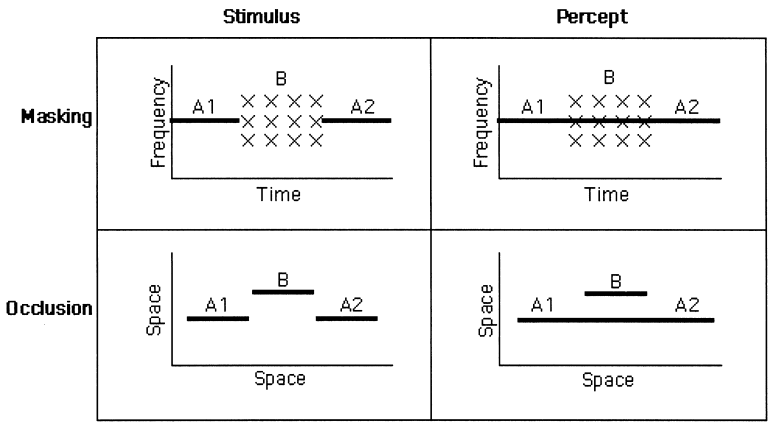

Fig. 1. The auditory continuity illusion (top) compared to the visual effect of continuity behind an occluder (bottom).

38). In a series of experiments, Watkins (Watkins, 1991, 1998, 1999; Watkins & Makin, 1996) has explored how the auditory system compensates for the distortion of spectral envelope (the major determinant of the perceived identity of many sounds) caused by factors such as room reverberation.

For the visual system just the opposite is true: it is generally concerned with *surfaces* of objects, not with the *sources* that illuminate them. As Mollon (1995) points out (giving credit to Monge, 1789):

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

These differences are summarized in Table 1.

For these reasons, we believe that to understand auditory objects we will have to rethink certain commonly accepted analogies between visual and auditory perception. In particular, we will show that both modalities are endowed with 'what' and 'where' subsystems, but that the relation between these four subsystems is complex. Obviously it is the 'what' subsystem of each modality that deals with objects, and so we will devote considerable attention to the auditory 'what' subsystem. But before we do, we must attend to the evidence connecting the auditory 'where' subsystem and the visuomotor orienting subsystem. We will claim that auditory localization is in the service of visual localization. This assertion is one of the cornerstones of our argument that space is not central to the formation of auditory objects.

Table 1
Comparison of vision and audition

Source of information	Vision	Audition
Primary	Surfaces	Sources
Secondary	Location and color of sources	Surfaces

2. Auditory 'where' in the service of visual 'where'

When two auditory sources appear to come from different spatial locations, shouldn't we say that they constitute different auditory objects, as do Wightman and Jenison (1995, pp. 371–372)? We prefer not to, because we believe that auditory localization is in the service of visual orienting, a hypothesis first formulated at the turn of the twentieth century by Angell: auditory "localisation occurs in the space world of vision–touch–movement… Most persons seem to make their localisation of sounds either in the form of visual imagery, or in the form of quasi-reflex localising movements of head and eye" (Angell, 1906, pp. 154–155). In this section we review the evidence supporting Angell's hypothesis.

The earliest sign of directional discrimination of sound in the human newborn is head orientation (Clifton, 1992), which suggests that the newborn is optimizing its head orientation to see the source.

Auditory localization is malleable, and can be influenced by the spatial location of a simultaneous visual stimulus. Bertelson and Aschersleben (1998) asked observers to judge whether the source of a sound was to the left or right of the median sagittal plane. When they presented the sound synchronously with an off-center flash, the sound appeared to be shifted in the direction of the light flash.

Sound localization itself is influenced by the act of visual orienting. Rorden and Driver (1999) presented observers with a noise burst from one of four speakers arranged in a fronto-parallel square array, and asked them to indicate whether the sound came from above or below their head. Before the noise was played, the observers were given a signal instructing them to move their eyes to the left or to the right. Reaction times (RTs) for correct up/down decisions were shorter when the direction of the intended eye movement was ipsilateral with the source of the noise than when it was contralateral to the source of the noise (regardless of whether the noise was heard before or after the eye movement was initiated).

Even clearer evidence of the role of auditory localization in visuomotor orientation is provided by Goldring, Dorris, Corneil, Balantyne, and Munoz (1996). On each trial they presented (for 1 s) either a visual target (an LED), an auditory target (broadband noise), or both, from a variety of azimuths. The participants' task was to turn their gaze towards the target as quickly as possible. When the targets were unimodal the relative eye–head latency depended on the eccentricity of the target: if eccentricity was less than 20° visual angle (dva), visual targets elicited a lower latency than auditory targets; beyond 20 dva, this order was reversed. For our purposes this result has one major implication: auditory localization is in the service of visual orientation where vision is weakest. (See also, Hughes, Nelson, and Aronchick (1998) who develop these findings further.)

To say that auditory localization is in the service of vision does *not* imply that auditory cueing is the most effective way to orient the visual system. Indeed, Jay and Sparks (1990) have shown that auditory-induced saccades are generally slower and less accurate than visually-induced saccades. Moreover, we are not arguing that auditory localization is equivalent to visual localization.

There is evidence of a one-way dependence of the visual modality on the auditory

modality from studies of multimodal cueing. Spence and Driver (1997) presented observers with a light or sound (the target) from one of four positions in a fronto-parallel square array, and asked them to indicate whether the sound came from above or below their head. Some ISI before the target, observers were presented with an uninformative exogenous visual or auditory cue from either the left or right side. RTs for correct localization were compared across conditions. The results showed short-lived (ISI $\leq$ 200 ms) facilitated performance for valid auditory cues when the target was either visual or auditory. Visual cues, however, facilitated performance only for visual targets. Spence and Driver have since replicated this effect numerous times; they interpret their results as evidence for auditory localization in the service of vision: audition influences visual localization but *not* vice-versa (Spence & Driver, 1999).

The dominance of vision over audition is confirmed in a case of left visual neglect, i.e. a derangement of visual space representation (Ladavas & Pavani, 1998). The patient was asked to point to left, center or right acoustic stimuli under visual control or blindfolded. Her pointing to left auditory stimuli was influenced by visual spatial information, i.e. she manifested left neglect. But when she was blindfolded, she pointed to the previously ignored left space.

In macaque monkeys, Jay and Sparks (1984) found that the auditory receptive fields shifted with changes in eye position, allowing the auditory and visual maps to remain in register. Even in the barn owl, for which auditory localization is of primary importance in predation, Brainard and Knudsen (1993) found that individuals reared wearing prisms undergo visually-induced changes in the tuning for sound localization cues in the tectum and in the external nucleus of the inferior colliculus (see also Aitkin, 1990; Aronson & Rosenbloom, 1971; Stryker, 1999; Zheng & Knudsen, 1999).

If an auditory 'where' subsystem exists, it would have to combine spatial and somatosensory information. Young, Spirou, Rice, and Voigt (1992) have produced intriguing evidence on this matter. They suggested that the dorsal cochlear nucleus in the cat DGN is responsible for early analysis of sound source location on the basis of two observations: (1) the "DGN principal cells are exquisitely sensitive to spectral features of stimuli that carry sound localization information…", and (2) there is a somatosensory input to the DGN which may be providing information about the orientation of the cat's mobile pinnae, and thus allowing the DGN to integrate pinna-position information with sound localization cues. In humans, there are numerous subcortical areas that are believed to be responsible for cross-modal integration, and which possibly contain a supramodal representation of space (Andersen, Snyder, Bradley, & Xing, 1997; Stein & Meredith, 1993). Audio-visual speech integration studies using fMRI (Calvert, Campbell, & Brammer, 2000) as well as PET studies examining visual-tactile integration (Banati, Goerres, Tjoa, Aggleton, & Grasby, 2000; Macaluso, Frith, & Driver, 2000) provide converging evidence that the superior colliculus, as well as portions of the heteromodal cortex, are likely candidate areas. (See Spence and Driver (1997) for a more complete review of the neurological evidence supporting this idea.)

Let us summarize our argument to this point: (a) sound appears to inform us about

Fig. 2. Grouping by proximity (after Wertheimer, 1923/1938).

Fig. 3. Grouping by similarity (after Wertheimer, 1923/1938).

sources and events rather than surfaces and material objects; (b) our language suggests to us that objects are visual; and (c) the visual objects we see have considerable control over what we hear. Wightman and Jenison (1995, pp. 371–372) distinguish between *concrete* auditory objects "formed by sounds emitted by real objects in the environment" (i.e. an orchestra) and *abstract* auditory objects, which "do not often correspond to real environmental objects" (i.e. a melody). We differentiate the auditory subsystem that processes these 'concrete' objects – the 'where' subsystem – from the auditory subsystem that processes 'abstract' auditory objects – the 'what' subsystem. To understand the auditory 'what' subsystem, we must abandon visuocentric notions of objecthood and offer a more general definition of perceptual object, be it visual or auditory.

3. 'What' subsystems: objects, grouping, figure-ground, and edges

A *perceptual object* is that which is susceptible to figure-ground segregation. This definition will allow us to develop a useful concept of auditory object. A critic who *defines* figure-ground segregation as a process applied to objects might claim that our definition is circular. But we believe that the benefit of the new definition outweighs the cost of abandoning the definition of figure-ground segregation in terms of objects. We believe that the process of grouping and most forms of feature integration are pre-attentive (Kubovy, Cohen, & Hollier, 1999; see also Bregman, 1990, pp. 206–209). We propose the following view of the relation between early processing, grouping, figure-ground segregation and attention. Early processing produces elements that require grouping. Grouping occurs following the principles described by the Gestalt psychologists (Figs. 2 and 3, from Wertheimer, 1923/1938, further developed by Kubovy, Holcombe, & Wagemans, 1998); it produces Gestalts, or perceptual organizations, which are also putative perceptual objects. Attention selects one putative object (or a small set of them) to become figure (Fig. 4) (Peterson & Gibson, 1994) and relegates all other information to ground (Fig. 5). The putative objects that become figure are perceptual objects, whereas the ground remains undifferentiated information (see Brochard, Drake, Botte, & McAdams, 1999 for evidence that the ground remains undifferentiated in audition).

There is little doubt that grouping and figure-ground segregation describe processes that are meaningful for auditory perception. Grouping is a well-estab-

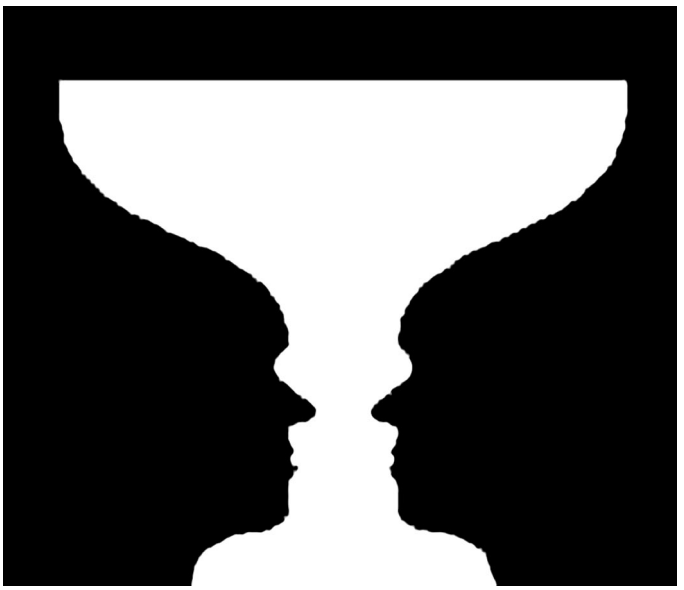

Fig. 4. Rubin vase/face.

lished auditory phenomenon. In particular, auditory stream formation – the auditory analog of visual grouping – has been studied in depth (Bregman, 1990; Handel, 1989, Chapter 7).

Figure-ground segregation is mentioned less frequently.[3] Nevertheless, in studies of streaming it has often been observed that when an auditory sequence breaks into two streams, we cannot pay attention to more than one of them at a time. For example, Bregman and Campbell (1971) presented observers with a repeating sequence of three high pitch notes (*ABC*) and three low pitch notes (123) in the order *A*–1–*B*–2–*C*–3. Observers typically reported the order of notes as *A*–*B*–*C*–1–2–3 or 1–2–3–*A*–*B*–*C*; they were able to attend to one stream or the other, but not both streams simultaneously. As in vision, whichever stream is being attended becomes the figure and the other the ground.

3.1. Plensensory functions and edges

Another phenomenon characterizes visual objects: the formation and assignment of edges. When the faces in Fig. 4 are seen in the foreground, they take ownership of

[3] Many forms of music use a drone, a low-pitched sustained tone that serves as an auditory background for music played at a higher pitch. The word suggests that a musical drone resembles the hum produced by bees. Drones were common in antiquity; today they appear in the music of the Balkans, Sardinia, Appalachia, and India, to name just a few.

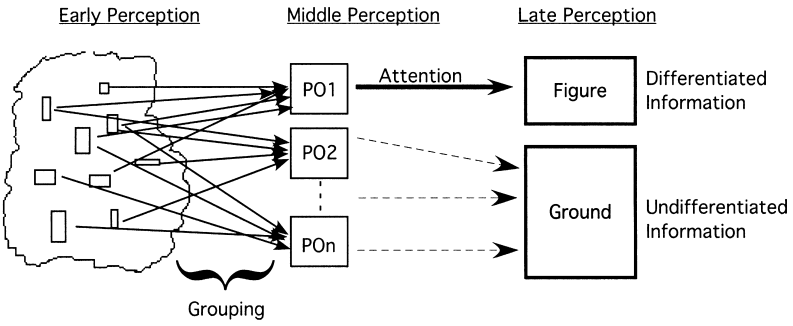

Fig. 5. A conservative view of the formation of perceptual objects. Early processing produces elements that undergo grouping to produce PO_1, PO_2, …, PO_n, a set of perceptual organizations (or putative objects – recall that we have defined an object as the result of figure-ground segregation; therefore, these elements are merely *candidates* for objecthood). Attention is required to produce figure-ground segregation, which allocates processing capacity to the figure and leaves the ground relatively undifferentiated.

the two edges and the background appears to continue behind them. Let us suppose that *any object, be it visual or auditory, must have an edge or a boundary*.

What kinds of edges can be found in optic and acoustic information? To answer this question for vision, Adelson and Bergen (1991) developed the *plenoptic function*, which allows us to characterize edges in the optic information available at every point in space and time. But, as we will see, as soon as we try to construct an analogous *plenacoustic function*, we will face difficult theoretical questions, which we will try to answer in the course of this paper.

The plenoptic function (Fig. 6) is a formalized answer to the question, What can potentially be seen? It is the culmination of a line of thought that began with Leonardo da Vinci that was introduced into the study of perception by Gibson (1979) and further explored by Johansson and Orjesson (1989), who put the idea as follows:

> The central concept of ecological optics is the ambient optic array at a point of observation. To be an *array* means to have an arrangement, and to be *ambient at a point* means to surround a position in the environment that could be occupied by an observer. The position may or may not be occupied; for the present let us treat it as if it were not.[4] (p. 65)

The construction of the plenoptic function starts with a viewpoint, $V(x,y,z)$. We place a box with a pinhole (called a *pinhole camera*) at V, pointing in the direction (ϕ,θ)

[4] In this sense, our figures are slightly misleading because we have inserted pictures of eyes at points of observation in space.

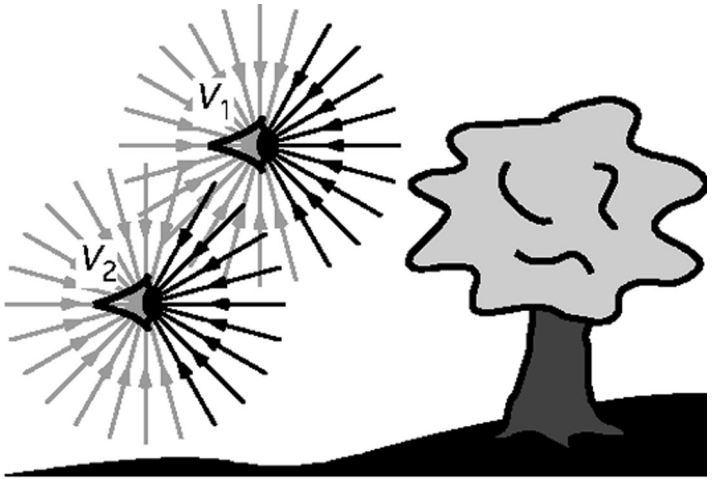

Fig. 6. Two points of the plenoptic function for viewpoints V_1 and V_2. Each arrow, light or dark, corresponds to a pencil of rays in direction (θ,ϕ). The dark arrows correspond to what might be seen by eyes at V_1 and V_2. (Slightly modified and redrawn from Fig. 1.3 in Adelson and Bergen (1991).)

(Fig. 7).[5] At the back of the camera is a spectrograph. We obtain a record of the intensity of light as a function of two variables: wavelength, λ, and time, t. Since we can change the viewpoint and rotate the camera in any direction, what can potentially be seen of a scene from any position is the plenoptic function, $P(\phi,\theta,\lambda,t,x,y,z)$.

Objects that reflect light readily create discontinuities in the plenoptic function: edges. By slicing the plenoptic function along various planes, as in Fig. 8, we can see how edges in these planes correspond to familiar features of the visual world. For example, Fig. 8a shows an edge that does not change in azimuth (ϕ); it is therefore vertical, whereas in Fig. 8b it does not change in tilt (θ), and it is therefore horizontal. Fig. 8h describes the effect of a horizontal movement of the viewpoint without changing either ϕ or θ.

We now turn to the plenacoustic function, in the hope that it will help us think about acoustic edges. But we encounter difficulties from the very outset. We are not sure how to illustrate it. Can we use Fig. 6, except that we replace the eyes with ears but leave the ecology (here, a tree) unchanged? That cannot be right. Even students of auditory navigation (Arias, Curet, Moyano, Joekes, & Blanch, 1993; Ashmead & Wall, 1999; Ashmead et al., 1998; Stoffregen & Pittenger, 1995) do not claim that a tree is a natural object of auditory perception. So we are reminded that the acoustic ecology differs from the optic ecology: as we pointed out earlier, auditory perception is more concerned with sources than with surfaces.

Should we then start from an illustration like Fig. 9? We think that even this

[5] We adopt the convention of Euler angles (Goldstein, 1980) according to which ϕ (azimuth) and θ (pitch or tilt) are successive counterclockwise rotations of the camera: ϕ is a rotation about the z-axis, and θ is a rotation about the ξ-axis of the camera in the xy plane.

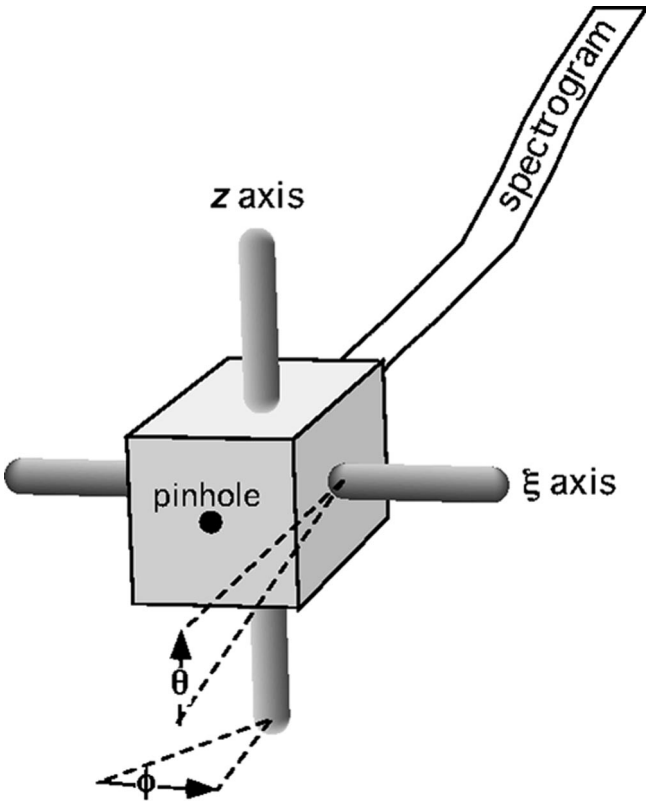

Fig. 7. The pinhole camera.

illustration is misleading, albeit subtly. It draws our attention to the 'where' aspect of audition: *Where* is the screaming boy? *Where* is the crying girl?

Before we can construct a plenacoustic function, we must think through the 'what' aspect of auditory perception. The Kubovy (1981) theory of indispensable attributes (TIA) will prove to be a useful tool in this endeavor.

3.2. Indispensable attributes, emergent properties, and grouping

Plensensory functions would suggest to us where edges might occur in the information available to an observer. Here we move beyond the perceptual ecology to examine the evidence that auditory objects are formed in pitch-time, whereas visual objects are formed in space-time. We begin with a series of thought experiments proposed by Kubovy to illustrate his TIA, after which we will present empirical evidence in support of this theory.

The TIA focuses on an important aspect of object formation: the aggregation of elements to form an emergent object. An emergent property is a property of an aggregate that is not present in the aggregated elements. For example, at room

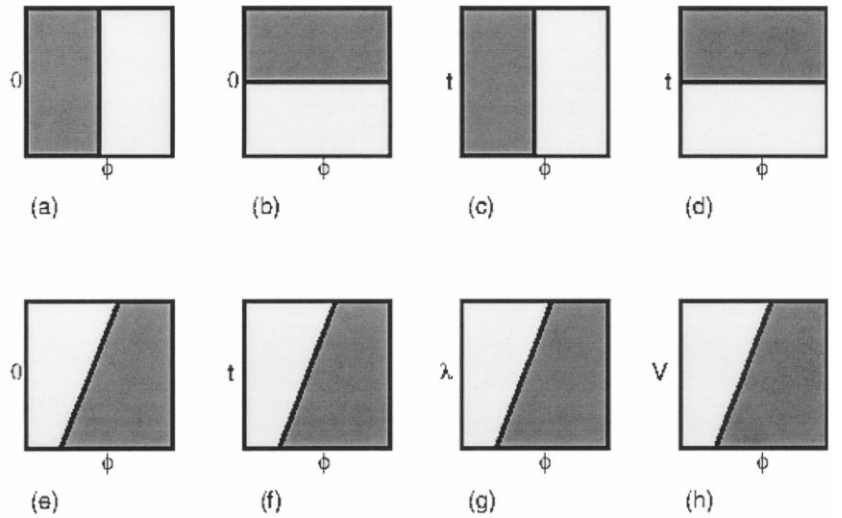

Fig. 8. Some edge-like structures that might be found along particular planes within the plenoptic function (note the varying axes, as labeled on each panel): (a) a vertical edge; (b) a horizontal edge; (c) a stationary edge; (d) a full-field brightening; (e) a tilting edge; (f) a moving edge; (g) a color sweep; (h) an edge with horizontal binocular parallax. (Slightly modified from Fig. 1.4 in Adelson and Bergen (1991).)

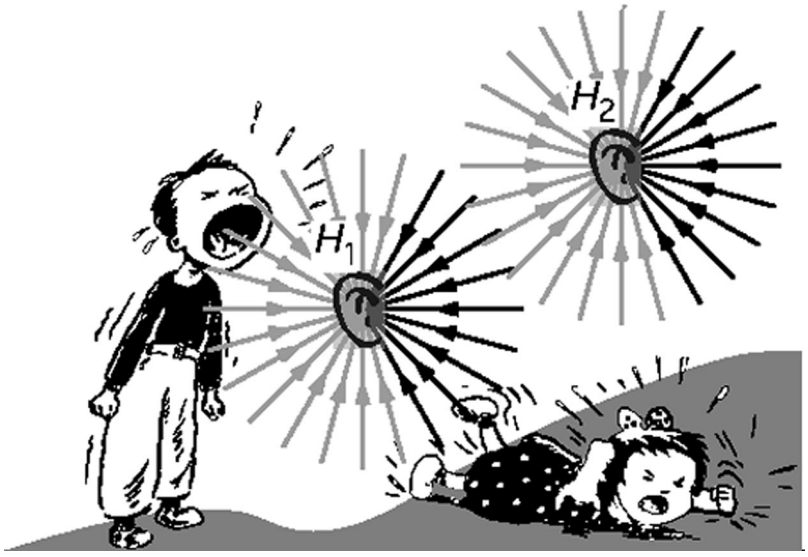

Fig. 9. Two points of the plenacoustic function for hearpoints H_1 and H_2. The dark arrows correspond to the sounds that might be least attenuated by the shadow of the head at H_1 and H_2.

temperature, water is a liquid, but the elements that compose it are both gasses. Thus, at room temperature, the property *liquid* is an emergent property of water. There are two kind of emergent properties: *eliminative* and *preservative*. When hydrogen and oxygen combine to form water, the properties of the elements (being gasses) are not observable; they are eliminated by the process of aggregation. In the human sciences such eliminative emergent properties are also common: we can mix two colored lights (say, red and yellow), and observers will not be able to tell whether the orange they see is a spectral orange or a mixture. Thus, color mixture is an eliminative emergent property. Preservative emergent properties were first noticed by Ehrenfels (1890/1988), who described a *melody* as being an emergent property of the set of notes comprising it. The notes can be heard; indeed they *must* be heard for the melody to be recognized. In a melody the elements are preserved in the process of aggregation; in fact the emergence of the melody is conditional upon the audibility of the elements.

The TIA offers a heuristic argument that suggests the conditions under which a perceptual aggregate will preserve the individuality of its elements. More simply, what are the features of stimuli that enable a perceptual system to determine that there is more than one entity in the environment? An attribute (or dimension) is defined as indispensable if and only if it is a prerequisite of perceptual numerosity. As we will show, these attributes are different for vision and for audition.

Spatial separation is an indispensable attribute for vision. Imagine presenting to an observer two spots of light on a surface (Fig. 10a). Both of them are yellow and they coincide; the observer will report one light. Now suppose we change the color of the lights, so that one spot is blue and the other is yellow, but they still coincide (Fig. 10b); the observer will report one white light. For the observer to see more than one light, they must occupy different spatial locations (Fig. 10c).

Pitch separation is an indispensable attribute for sound. Imagine simultaneously playing two 440 Hz sounds for a listener (Fig. 11a). Both of them are played over the same loudspeaker; the listener will report hearing one sound. Now suppose we play these two sounds over two loudspeakers (Fig. 11b); the listener will still report hearing one sound. For the listener to report more than one sound, they must be separated in frequency (Fig. 11c).

By analogous argument, time is an indispensable attribute for both vision and audition. Time thus takes on the role of a common indispensable attribute.

We would like to head off several possible misinterpretations of the TIA. (a) We do *not* claim that auditory spatial cueing is ineffective. On the contrary, we have no doubt that auditory spatial cueing gives rise to costs and benefits (Scharf, 1998). Our claim is that although spatial cueing may be sufficient to draw attention to a pitch, attention is allocated to the *pitch*, not to its *location*. (b) We do not claim that indispensable attributes are prerequisites of perceptual numerosity at every point. For example, consider Fig. 12. It would be foolish of us to claim that we do not see two planes at X. Rather, we see two overlapping objects, and we see them occupying different extensions in space. In audition, an analogous case is homophonic induction (Warren, 1982). Homophonic induction occurs when observers are played a long pure tone with periodic amplitude increases: we hear one continuous tone and a

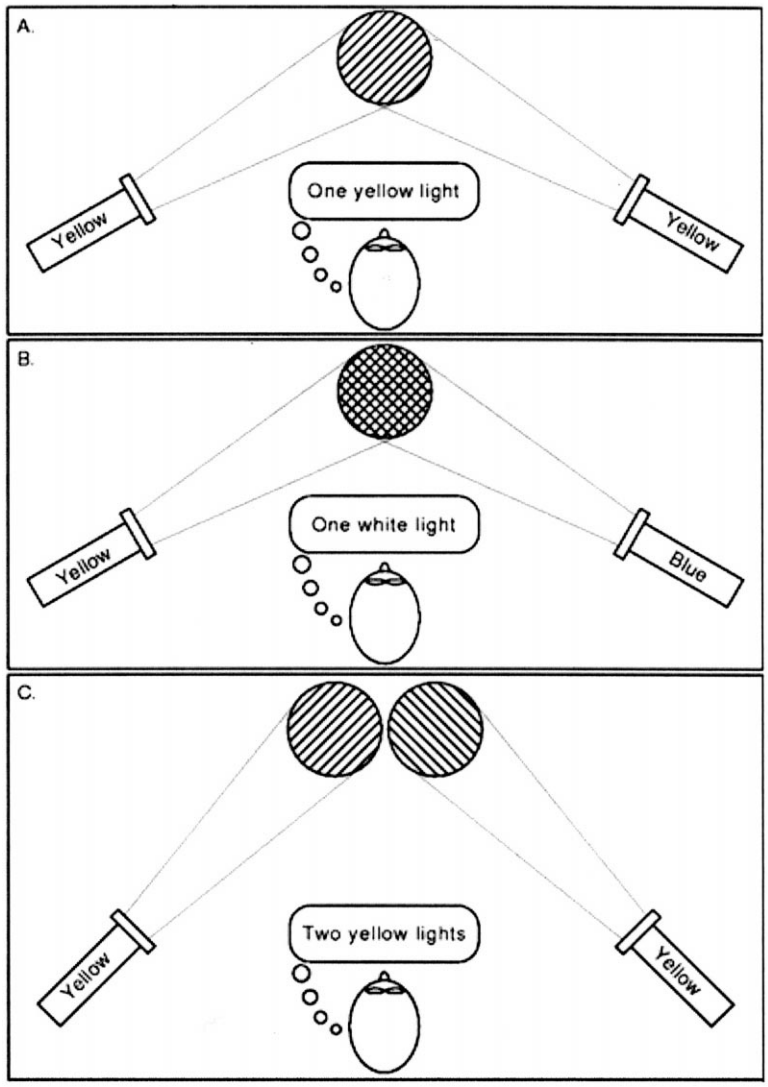

Fig. 10. (a) Two yellow spotlights create coincident spots. The observer sees one yellow spot. (b) One yellow spotlight and one blue spotlight create coincident spots. The observer sees one white spot. (c) Two spotlights create separate spots. Regardless of their color, the observer sees two spots.

second intermittent tone of the same pitch. Thus, we hear two overlapping objects, and we hear them occupying different extensions in time.

We draw the following conclusions from our thought experiments. (a) In vision space is an indispensable attribute for perceptual numerosity; color is not. (b) In audition frequency is an indispensable attribute for perceptual numerosity; space is not. (c) In both, time is an indispensable attribute.

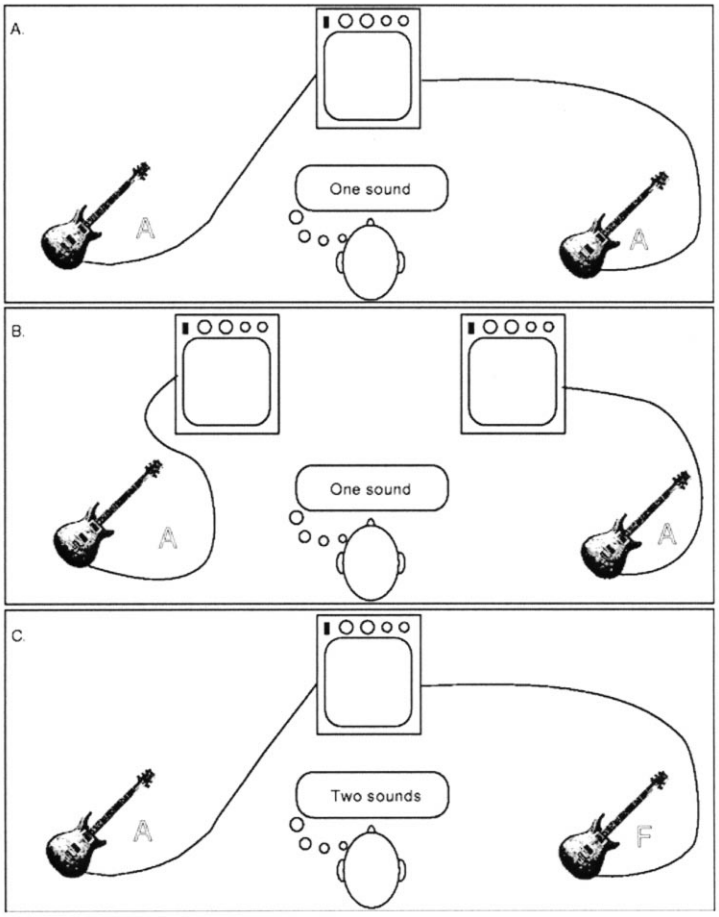

Fig. 11. (a) One loudspeaker plays two *A*s. The listener hears one *A* sound. (b) One loudspeaker plays an *A* while another speaker plays an *A*. The listener hears one *A* sound. (c) An *A* and an *F* are played. Regardless of whether they are played over one loudspeaker or two, the listener hears two sounds, an *A* and an *F*.

3.3. Indispensable attributes and edges

Earlier we could not construct a plenacoustic function because we did not know enough to map optics onto acoustics. With the TIA in hand, we can conjecture that we will find contours in the indispensable attributes of each modality. Looking back at Fig. 8 we observe that in each of the eight panels, one of the axes of the plane is spatial or temporal.

Likewise, in audition, we find edges in pitch, and edges in time, but not in space. The claim that there are edges in pitch may seem strange, but a moment's thought will show that the idea is quite natural. A biologically important source is periodic, at least over short periods of time. Therefore, it is characterized by a *fundamental frequency* that can be thought of as its lower edge in pitch. As we mentioned earlier,

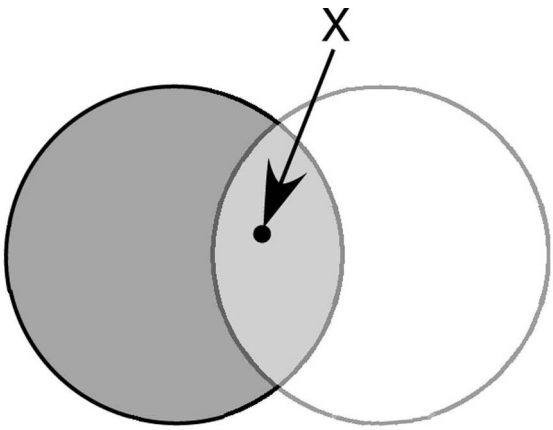

Fig. 12. Transparency: observers perceive two planes at X.

the object delimited can be schematically represented by its spectrum, as shown in Fig. 13. It is further characterized by the shape of its leading edge in pitch-time, its *attack*, and its trailing edge in pitch-time, its *decay*.

3.4. Interim conclusions

Up to here we have done two things. (a) We have argued in favor of an auditory 'where' subsystem that is linked to visuomotor orientation. (b) We have offered a new definition of perceptual objecthood, and we have shown that it implies a new way of thinking about auditory objects. We also made an assumption: in perceptual systems that have separate 'what' and 'where' subsystems, the 'what' subsystems are responsible for the generation of perceptual objects. We now turn to a fundamental question: what is the evidence for two auditory subsystems?

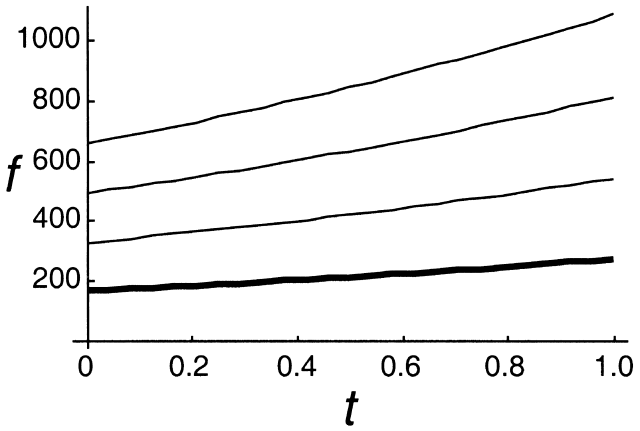

Fig. 13. An example of an auditory object. The first four harmonics of a tone gliding from 165 to 272 Hz over 1 s. The fundamental (the thick line) is the edge of the object.

4. Evidence for two auditory subsystems

The idea of a parallel between the two visual subsystems and two auditory subsystems is gaining favor (Cisek & Turgeon, 1999). Unfortunately, the evidence for a separation of streams in the auditory system is scattered in the literature and may not be sufficiently strong to be conclusive. We turn first to behavioral evidence, and then present neurophysiological evidence.

It is possible to create an auditory illusion in which the 'what' of a stimulus is perceived correctly, but the 'where' is perceived incorrectly. Deustch and Roll (1976) created a stimulus that consists of a sequence of five pairs of concurrent dichotic tones. Let us denote by *A* a 400 Hz tone lasting 250 ms, and by *B* an 800 Hz tone of the same duration. The right ear receives *AAABB* (with 250 ms pauses between tones) and the other receives *BBBAA*. Right-handed listeners report hearing *AAABB*, but they hear the first three tones as if they came to the right ear and the remaining two as if they came to the left. Thus, they dissociated the two subsystems, because the dominant ear determines 'what' is heard, but the localization of the sounds (the 'where') is illusory.

Under some conditions, adaptation affects localization but not pitch perception. Hafter (1997) studied adaptation to trains of high-frequency clicks. He presented click trains in which the clicks had an inter-aural time disparity (ITD), so that listeners heard the clicks localized to the right or to the left. The listeners showed *binaural adaptation*: the effectiveness of the inter-aural information in the click train gradually diminished. He then presented listeners with trains of clicks that give rise to *periodicity pitch*, which is controlled by the inter-click interval (ICI). He asked listeners to discriminate changes in the pitch they heard when he changed the ICI. He found that the listeners suffered no adaptation in the performance of this task.

A somewhat indirect piece of behavioral evidence in favor of two auditory subsystems comes from experiments on the distribution of auditory response times. Wynn (1977) collected more than 300 000 response times from observers who were asked to tap in synchrony with a visual or auditory pulse. The distribution of RTs to visual pulses was unimodal. But the distribution of RTs to auditory pulses was bimodal, with many more fast responses than slow ones. As he increased the intensity of the sounds, three phenomena emerged: (a) the modes did not shift; (b) the humps became narrower; and (c) the mode of slow RTs diminished. From this Wynn concludes that there are two pathways in audition: one 'slow' and one 'fast'.

There also is neurophysiological evidence in favor of our hypothesis. On the basis of single-unit and tracer studies in cats and macaques as well as results from PET and fMRI experiments on humans, Rauschecker (1997, 1998a,b) argues that the auditory system has a dorsal 'where' stream and a ventral 'what' stream. Vocal communication sounds – presumably processed by a 'what' subsystem – project into areas of the lateral belt (and parabelt) of the superior temporal cortex. Auditory spatial information – presumably processed by a 'where' subsystem – projects into parietal areas. Studies of human patients with left hemispheric lesions provide a

double dissociation between the two subsystems (Clarke, Bellmann, Meuli, Assal, & Steck, 2000). Patients with lesions in the medial, posterior auditory cortex (which Clarke et al., 2000 call the 'putative spatial pathway') exhibit localization deficits, whereas more lateral lesions (in the 'putative recognition pathway') cause recognition deficits.

5. Relations between 'what' and 'where' in vision and audition

We have suggested that the auditory 'where' subsystem is probably in the service of the visuomotor subsystem. The visual 'where' subsystem provides us with spatial information about the world in egocentric terms (Milner & Goodale, 1995). We believe the same to be true about the auditory 'where' subsystem. In other words, both the visual and the auditory 'where' subsystems may be thought of as being in the service of action.

It is harder to describe the relation between the visual and the auditory 'what' subsystems. We suggested earlier that they both are concerned with the segregation of figure from ground and the grouping of elements into objects. In vision this conception does not create a disjunction between the products of the 'what' and the 'where' subsystems, although the mechanisms are dissociable. Visual objects are extended in space, and they are located in space. So in vision the two subsystems seem to work to the same end: collecting information about a unified spatiotemporal reality.

In audition, however, distinguishing between the two subsystems creates what might be called an ontological chasm. The kinds of objects the auditory 'what' subsystem constructs need not be located in space, nor do they need to be defined in spatial terms. Nevertheless, if we surmise that the origin of the auditory 'what' subsystem is in vocal communication, then its non-spatiality becomes somewhat less puzzling.

The problem we encountered earlier, when we were considering the possibility of constructing a plenacoustic function, may be called the *dimension mapping question*: which dimensions of the optical input are analogous to which dimensions of the acoustic input? *The answer depends on our view of these inputs*. If we think of them in spatiotemporal terms, then the spatial and temporal dimensions in the two inputs will correspond to each other. Thus, six dimensions of the plenoptic function, ϕ, θ, t, V_x, V_y, V_z, map onto six dimensions of a putative plenacoustic function, ϕ, θ, t, H_x, H_y, H_z. We call this the *spatiotemporal mapping*. If, on the other hand, we think of the plensensory functions in terms of object formation, then the analogous dimensions in the two functions are those that allow for the formation of edges in the plenoptic and the plenacoustic functions. We call this mapping the *indispensable attributes mapping*.

5.1. The spatiotemporal mapping

Under the spatiotemporal mapping visual space and time are mapped onto auditory space and time. This is a natural view to take in the light of the assumptions

made by Newton (1726/1999): "Absolute, true, and mathematical time, in and of itself and of its own nature, without reference to anything external, flows uniformly… Absolute space, of its own nature without reference to anything external, always remains homogeneous and immovable … times and spaces are, as it were, the places of themselves and of all things" (pp. 408–410). In his *Critique of Pure Reason*, Kant (1787/1996) took Newton's idea further: "Space is not an empirical concept that has been abstracted from outer experiences. For the presentation of space must already lie at the basis in order for certain sensations to be referred to something outside of me…" (p. 77, B 38).[6] "Time is not an empirical concept that has been abstracted from any experience. For simultaneity or succession would not even enter our perception if the presentation of time did not underlie them a priori" (p. 85, B 46).

Newton and Kant were powerful influences. It was natural for early psychologists to adopt the spatiotemporal mapping. From this mapping they concluded that color gets mapped onto pitch, since they are non-spatial and non-temporal, and since they are both caused by waves. For example, we read a footnote by A.J. Ellis, translator of *Sensations of Tone* by Helmholtz (1885/1954):

> Assuming the undulatory theory, which attributes the sensation of light to the vibration of a supposed luminous 'ether', resembling air but more delicate and mobile, then the phenomena of 'interference' enables us to calculate the lengths of waves of light in empty space, &c., hence the numbers of vibrations in a second, and consequently the ratios of these numbers, which will then clearly resemble the ratios of pitch numbers that measure musical intervals. Assuming, then, that the yellow of the spectrum answers to the tenor *c* in music, and Fraunhaufer's 'line *A*' corresponds to the *G* below it, Prof. Helmholtz, in his *Physiological Optics* (*Handbuch der Physiologischen Optik*, 1867, p. 237), gives the following analogies between the notes of the piano and the colors of the spectrum:
>
> *G*, Red,
> *G*#, Red,
> *A*, Red
> *A*#, Orange-red,
> *B*, Orange,
> *c*, Yellow,
> *c*#, Green,
> *d*, Greenish-blue,
> *d*#, Cyanogen-blue
> *e*, Indigo-blue,
> *f*, Violet,
> *g*, Ultra-violet,
> *g*#, Ultra-violet,
> *a*, Ultra-violet,
> *a*#, Ultra-violet,
> *b*, end of the solar spectrum.
> The scale therefore
> extends to about a Fourth
> beyond the octave.
>
> (p. 18)

[6] This is the traditional way to denote p. 38 in the second edition of the *Critique*.

As profound differences between light and sound became clear in the twentieth century, psychologists abandoned the exploration of parallels between pitch and color. But the rest of the spatiotemporal mapping has been retained. Research on the 'cocktail party problem' is an excellent example. How do we segregate a speaker's voice from the voices of other concurrent speakers at a cocktail party? To Cherry (1959), who coined the term, the problem was a spatial one. All of his experiments involved dichotic listening: the listener hears a different message in each ear and has to report something from one or both. The assumption of the primacy of space in audition is even clearer in Broadbent (1958): "Sounds reaching the two ears are of course often perceived as coming from different directions … and such sounds we will regard as arriving by different 'channels'" (p. 15). In this context the auditory system was considered to be a 'where' system, and auditory segregation was thought to be spatial in nature.

The implicit assumption of the auditory system as a 'where' system persists. For example, Handel (1988) criticized the original formulation of the TIA (Kubovy, 1981) for this very reason (for a reply, see Kubovy, 1988): "The auditory and visual worlds are inherently both temporal and spatial" (p. 315). For this reason, Handel opposed Kubovy's commitment to the TIA mapping and claimed that all mappings are possible and relevant, depending on the context. Although we have come somewhat closer to Handel's position by proposing two mappings, we are making a more specific claim. For the 'where' subsystems the spatiotemporal mapping is appropriate; for the 'what' subsystems the TIA mapping is appropriate.

5.2. The indispensable attributes mapping

We believe that the TIA mapping will allow researchers to formulate testable hypotheses about the nature of auditory objects and auditory object perception. The TIA is a heuristic tool for extending theories of visual perception into the domain of auditory perception (and perhaps vice-versa). Note that such extensions have been done in the past with considerable success. For example, Bregman (1990) has shown the similarities between the Gestalt principles in vision and in audition. Just as grouping by proximity functions in visual space (Kubovy et al., 1998), it also operates in auditory pitch (Bregman, 1990). McPherson, Ciocca, and Bregman (1994) have shown that good continuation operates in audition in an analogous way to vision. The concept of amodal completion as it is used in vision (Kanizsa, 1979) has been given a number of different names in audition: the acoustic tunnel effect (Vicario, 1960), perceptual restoration (Warren, 1970, 1984), the continuity effect (Bregman, 1990), and the phonemic restoration effect (Samuel, 1991). Since all of these phenomena abide by the same laws of grouping and organization, a desirable goal would be to have a theoretical framework which can account for this. The TIA is such a framework.

5.3. Implications for theories of attention

An attentional counterpart of TIA – which we will abbreviate to ATIA – could

come in two versions: *strong* and *weak*. According to a strong ATIA, selective attention can *only* be directed to indispensable attributes, and not to other stimulus attributes. For example, the strong ATIA predicts that in vision you can pay attention to a region of space (or to an object defined by its spatial boundaries), but not to a part of color space. The work of Shih and Sperling (1996) favors such a position. They have shown that visual selective attention can be directed to space but not to color or size. On each trial of the experiment they presented a series of briefly presented frames (typically 27 of them). In each frame six letters were presented, equidistant from a fixation point. The characters in a frame were uniformly colored red or green. In one of these frames one letter was replaced with a digit. The observer's task was to report the digit's name, location, and color. The dependent variable was the observer's report accuracy. Before each trial the observer was given information (not always valid) about the color of the digit. The main result was this: the observers' accuracy was not affected by the validity of the cue. In a second experiment the color of the digit differed from the color of the letters, and observers did show a cost or benefit that depended on the validity of the cue. Thus, color can *draw* our attention to spatial locations, we cannot selectively attend to color – only to spatial locations.

According to a weak ATIA selective attention is generally directed towards indispensable attributes, but *can* be directed towards other attributes. For example, color is not an indispensable attribute, yet color-based inhibition of return (IOR) has been reported. Law, Pratt, and Abrams (1995) showed observers two successive color patches in the center of a monitor, with an ISI of 900 ms. They asked observers to respond as soon as they detected the second patch. When the two colors were the same, RTs were ≈ 5.5 ms longer than when the colors were different. We note, however, that the magnitude of the inhibition observed in this experiment is much lower than the effects observed with space-based IOR. The authors themselves acknowledge that "the color-based inhibition of return that we observed might be reduced or eliminated in situations with spatial uncertainty" (p. 407). We therefore await further progress on this topic before we retreat from the strong ATIA.

5.4. Costs of not adhering to indispensable attributes

The spatiotemporal mapping implies that space holds the same status in audition as it does in vision. Culling and Summerfield (1995) have argued the contrary: "the introspective impression that one can concentrate on a particular direction in order to pick out an attended voice from an interfering babble may be misleading. …there is evidence that localization occurs after concurrent sounds have been separated by other processes" (p. 796). In other words, auditory objects are *not* formed in space. This assertion is supported by three experiments in which Culling and Summerfield explored the role of within-channel and across-channel processes in the perceptual separation of competing sounds that had different inter-aural phase spectra. A similar position is adopted by Darwin and Hukin (1999) in regards to speech sounds. They claim (on p. 622, illustrated in their Fig. 3, right-hand panel) that auditory objects are the result of non-spatial grouping processes (e.g. harmonicity and onset

time). Once an object is formed, listeners can direct their attention to it, and even attend to its direction. This is precisely what we have been arguing.

Reliance on the spatiotemporal mapping may have led researchers to the erroneous conclusion that attention operates differently in vision and in audition. For example, Posner (1978) attempted to obtain spatial cueing effects in a variety of tasks, using endogenous, or top-down, cues. Even though his attempt failed, Spence and Driver (1994) determined that "it is clear that endogenous mechanisms of spatial auditory attention must exist at some level of processing: otherwise we could not achieve such textbook feats, such as selectively shadowing a message from one location while ignoring a message from a different location" (p. 557). So Spence and Driver do not believe Posner's data because they are not consistent with the spatio-temporal mapping. Butchel and Butter (1988) used exogenous spatial cues and found spatial cueing effects in vision but not in audition. They argue that due to the lack of a 'fovea-like' area in audition, there is no spatial attention in audition.

According to Spence and Driver (1994) there may be several reasons why Butchel and Butter (1988), Posner (1978), and Scharf, Quigley, Aoki, Peachey, and Reeves (1987) failed to find an auditory IOR. (a) Audition is poorer than vision with respect to spatial localization, and so maybe past experimenters failed to utilize sufficient angular separation. (b) The intensity of the peripheral cues may have been insufficient to draw attention. (c) Because RTs in auditory detection/discrimination tasks are generally faster than in vision, the auditory response speeds may have been at ceiling because the tasks may not have been sufficiently demanding. (d) Perhaps the previous tasks were performed without engaging the spatial aspects of audition. In an extensive set of experiments, Spence and Driver showed that when the auditory spatial localization subsystem is engaged by the task, attentional effects appear. Observers were given exogenous cues at the lateral midline from either the left or the right with a sound that they were to ignore and which did not predict the lateralization of the target. Later observers were presented with a target sound on either the same side or the opposite side as the cue. Their task was to press one of two buttons as quickly as possible to indicate where the target was located. Targets were presented either in front of/behind (Experiment 1) or above/below the lateral midline (Experiment 2). They found an advantage for valid cues at ISI = 100 ms, but not at ISI = 400 ms or ISI = 1000 ms. There was no IOR. This was interpreted by Spence and Driver as being evidence for a short-lived advantage due to exogenous orientation. When the observers were required to make a frequency discrimination instead of a localization judgement (Experiment 3) the advantage on the cued side disappeared. Spence and Driver concluded that when the task demands are not based on localization, there is no attentional effect of spatial cueing in audition.

In a subsequent report, Spence and Driver (1998) argue that IOR operates differently in vision and audition. They showed that auditory RT *was not* affected by the location of the auditory target that appeared on the preceding trial, whereas visual detection RT *was* affected by the location of the visual target on the preceding trial. Only when targets were unpredictably auditory or visual did this effect occur between auditory targets presented on successive trials. Quinlan and Hill (1999)

explained the auditory IOR in the modality-uncertain condition of Spence and Driver (1998) as a 'modality' switching effect. In a series of three experiments, observers made left/right localization judgments to either visual or auditory signals. In the first two experiments, observers were given a cue indicating the signal's modality, whereas in the third experiment the observers were not cued. The difference between the first and second experiments was the ISI between the modality cue and the subsequent presentation. The results of Experiment 1 (ISI = 50 ms) show that there is a cost to switching between modalities between trials. Observers were significantly slower when the modality changed between experimental trials. In Experiment 2 (ISI = 500 ms), where observers had time to prepare for the switch between modalities, the effect disappeared. Quinlan and Hill interpreted this result as an indication that the costs seen in Experiment 1 were due to the modality switching, and further, that modality switching requires attention. Finally, in Experiment 3, Quinlan and Hill showed that when there is no modality preparatory signal and ISI = 500 ms, the costs associated with modality switching re-emerge. They concluded that the effects seen in this and in Spence and Driver (1998) were the result of modality switching, not IOR. We believe that the above studies examined the 'where' component of the auditory system. Many researchers have hypothesized that this link occurs in the superior colliculus (Abrams & Dobkin, 1994; Goldring et al., 1996; Hughes et al., 1998; Spence & Driver, 1994; Tipper, Weaver, Jerreat, & Burak, 1994).

A TIA viewpoint would lead to a different experiment: instead of having tones vary in location, they would vary in pitch. Under *these* circumstances, we would expect to observe auditory IOR. (We are puzzled by the results of Mondor and Breau (1999) and Mondor, Breau, and Milliken (1998), who found frequency-based IOR, but also localization-based IOR. We are currently replicating their work.)

5.5. Benefits of indispensable attributes

When the TIA mapping is used, further analogies as well as an interesting separation between the auditory and the visual 'what' subsystems emerge. Duncan, Martens, and Ward (1997) report on experiments in which observers were asked to identify a stimulus shortly after they had deployed their attention to another stimulus (to study the so-called 'attentional blink', AB). In one experiment the stimuli were visual: they consisted of two streams of written trigrams ('xxx', 150 ms long, separated by an ISI of 100 ms). One stream consisted of a pair of trigrams above and below a fixation cross; the other consisted of a pair of trigrams to the right and the left of the fixation cross. One stream lagged behind the other by 125 ms. Each stream contained one target trigram ('nap', 'nab', 'cod', or 'cot'). The participants were asked to report which target trigrams they had seen. When the two target trigrams were 125 or 375 ms apart, the identification of the later trigram was depressed, i.e. an AB was observed. In a second experiment the stimuli were auditory: they consisted of two streams (one high-pitched, the other low-pitched) of spoken syllables ('guh', 150 ms long, separated by an ISI of 100 ms). One stream lagged behind the other by 125 ms. Each stream contained one target

syllable ('nap', 'nab', 'cod', or 'cot'). The listeners were asked to report which target syllables they heard. If the two target syllables were 125 or 375 ms apart, the identification of the later syllable was depressed, i.e. an AB was observed. The auditory and the visual results were remarkably similar. In a third experiment one stream was visual and the other was auditory. No AB was observed. This result does not only fit what we would expect, given the TIA mapping, it also shows that the two 'what' subsystems are not linked. Despite these results, some research has shown that cross-modal AB can be elicited under certain circumstances (Arnell & Jolicoeur, 1999; Potter, Chun, Banks, & Muckenhoupt, 1998) – therefore tending to favor a weaker version of the ATIA. Arnell and Jolicoeur (1999), for example, have shown that if the target presentation rate is under 120 ms/item, then cross-modal AB is observable. Arnell and Jolicoeur argue that this cross-modal AB occurs because of a central processing limitation.

According to Treisman and Gelade (1980), if a target differs in one feature from a set of distracters (e.g. an O among Xs or a red O among green Os) the RT to find the target is independent of the number of distracters. If a difference between the target and the distracters is a conjunction of features (e.g. a green O among green Xs and red Os) then the RT varies with the number of distracters. It is as if targets in the single feature condition spontaneously segregated themselves from the other elements, but failed to do so when they were defined by a conjunction. An analogous phenomenon has been demonstrated by Lenoble (1986). Her work built upon a demonstration of concurrent-pitch segregation by Kubovy, Cutting, and McGuire (1974), in which listeners heard seven equal-intensity binaural tones. When the ITD of individual tones was manipulated, listeners were able to hear a melody segregated from the complex even though it was not audible in either ear alone. When Lenoble (1986) presented observers with tone complexes in which target tones were defined by a conjunction of ITD and amplitude or frequency modulation, concurrent-pitch segregation did not occur.

An interesting (and testable) hypothesis is that observers are only able to allocate voluntary attention to indispensable attributes. The Shih and Sperling (1996) results indicate that observers are able to voluntarily devote attention to space, but that only space could be attended to in this way. This would mean that in audition, observers would only be able to allocate attention to particular frequencies or pitch space and not to other features such as rise time, intensity, or space. While to our knowledge a critical test of this hypothesis has not been made, there are results which suggest that this may indeed be the case. Mondor and Terrio (1998) conducted a series of five experiments designed to examine the role of selective attention and pattern structure in audition. They presented observers with a sequence of ascending or descending tones and the observer was to make a speeded response to a target tone that differed from the non-targets in duration, rise time, or intensity, or the target tone contained a 1 ms gap. Target tones could be consistent with the overall pattern structure or inconsistent (in frequency). When targets fell on tones that were consistent with the overall pattern structure observers were not more sensitive or faster to make a response. When targets fell on tones inconsistent with the overall pattern structure, however, observers

remained equally sensitive (as measured by d') but were significantly faster to detect the targets. Deviations in frequency from an established pattern were sufficient to draw attention and thus enhance the detection of these features when they fell on an inconsistent target. The implication of these studies is that frequency is more important than duration, rise time, intensity, or the presence of a gap for auditory selective attention mechanisms (although we acknowledge that the evidence would be stronger had Mondor and Terrio (1998) tested the analogous case – where pattern inconsistency is defined by, for example, duration or rise time).

The following experiment could serve as a critical test of the hypothesis that observers can only voluntarily allocate attention to indispensable attributes. We would place listeners in front of a linear array of four loudspeakers (i.e. two on each side of the observer's midline). We would use five different instrumental sounds: a target (i.e. a bassoon) and four distracters (i.e. a flute, a guitar, a piano, and a trumpet), chosen so that they are easily distinguishable. On each trial, we would play a series of brief 'frames' of sound, each of which would comprise the four distracter instruments. Each of the instruments would be played from a different speaker and at a different frequency. In one of the frames we would replace one of the distracters with the target instrument (the bassoon). We would ask the listeners to determine which frame contained the target. Before each trial we would give the listeners one of two types of cues: (a) a spatial cue (e.g. informing them that the target is 80% likely to come from the left), or (b) a pitch cue (e.g. informing them that the target is 80% likely to be high-pitched). According to the TIA there should be no cost or benefit from spatial cues, because listeners cannot allocate attention to a spatial location, but there should be costs and benefits associated with pitch cues.

6. Overview

In summary, consider Fig. 14. On the left side of the diagram we have set out the characteristics of audition, and on the right we have done so for vision. Each of the modalities is represented by two pathways, one labeled 'what' and the other labeled 'where'. We should stress that we are using the term 'where' as shorthand for the sense of Milner and Goodale (1995), i.e. a subsystem that maintains spatial information in egocentric coordinates for the purpose of controlling action. That is why we sometimes refer to it in this paper as a visuo*motor* system.

In the center of the diagram we show that the auditory and the visual 'where' subsystems are tightly linked. This is because of the evidence (see Section 2) that the auditory 'where' subsystem is in the service of the visual 'where' subsystem. We connected these two subsystems with a thick line to indicate their linkage, and added an arrow to this line to indicate the asymmetric relation between them (one is in the service of the other). The two subsystems are mapped onto each other with the traditional spatiotemporal mapping.

To either side of the 'where' subsystem(s) we represent the 'what' subsystems.

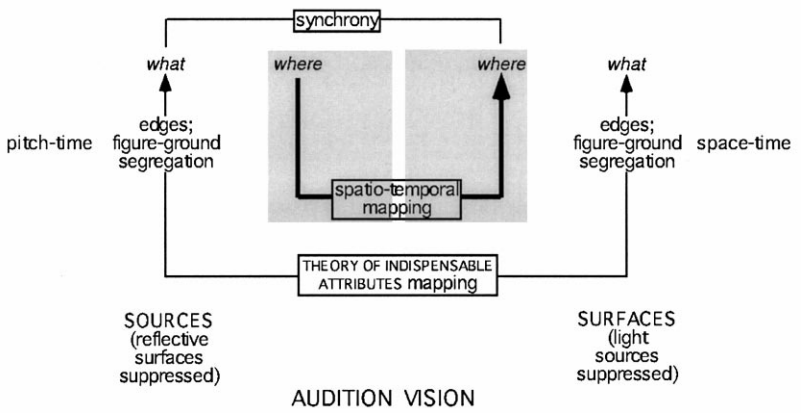

Fig. 14. Summary of the theory.

Here we show that the key operation of both the visual and the auditory 'what' subsystems is figure-ground segregation and edge formation (see Section 3.1). We also indicate that the auditory operation occurs in pitch-time, whereas the visual operation occurs in space-time (see Section 3.2). We connected these two subsystems with a thin line to indicate that they are analogous, and that the heuristic we offer to draw analogies between the two is the TIA. When we use the term 'analogy' we do not wish to take a stand on whether they are *merely* analogous, i.e. that they evolved separately, or whether they might be *homologous*, i.e. they have some common evolutionary origin.

We also note a link between the visual 'where' and auditory 'what' which represents the ventriloquism effect, in which synchronous visual and auditory events can determine the auditory localization.

Finally, we remind the reader (in the lower left and right corners of the auditory and visual boxes) of a fundamental difference between the two subsystems (summarized in Table 1).

7. Conclusion

The human cortex contains 10^{10} neurons. Up to half of these may be involved in visual function (Palmer, 1999, p. 24); the auditory system is much smaller. This seems to confirm that reality unfolds in space and time and that understanding is visual. But we believe that the main source of resistance to a non-visuocentric view of perception is the 'Knowing is Seeing' metaphor. According to Lakoff and Johnson (1999, Table 4.1, pp. 53–54) this metaphor (summarized in Table 2) is a tool all of us use to understand the idea of knowing.

We hope that our analysis will enable us to hear more clearly the polyphony between the two voices in the complex counterpoint between vision and audition.

Table 2
The Knowing is Seeing primary metaphor (Lakoff & Johnson, 1999, pp. 393–394)

Visual domain	Knowledge domain
Object seen	→ Idea
Seeing an object clearly	→ Knowing an idea
Person who sees	→ Person who knows
Light	→ 'Light' of reason
Visual focusing	→ Mental attention
Visual acuity	→ Intellectual acuity
Physical viewpoint	→ Mental viewpoint
Visual obstruction	→ Impediment to knowing

Acknowledgements

We wish to thank B.J. Scholl and J. Mehler for their superb editorial work on this paper. We are also grateful to those who contributed in various ways to this paper: A. Bregman, R.S. Bolia, C. Spence, S. Handel, C.L. Krumhansl, J.G. Neuhoff, B. Repp, M. Turgéon, and A.J. Watkins. Our work is supported by NEI grant No. R01 EY 12926-06.

References

Abrams, R. A., & Dobkin, R. S. (1994). Inhibition of return: effects of attentional cuing on eye movement latencies. *Journal of Experimental Psychology: Human Perception and Performance*, *20* (3), 467–477.

Adelson, E. H., & Bergen, J. R. (1991). The plenoptic function and the elements of early vision. In M. S. Landy, & J. A. Movshon (Eds.), *Computational models of visual processing* (pp. 3–20). Cambridge, MA: MIT Press.

Aitkin, L. (1990). Coding for auditory space. In M. J. Rowe, & L. Aitkin (Eds.), *Information processing in mammalian auditory and tactile systems* (pp. 169–178). New York: Wiley-Liss.

Andersen, R. A., Snyder, L. H., Bradley, D. C., & Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience*, *20*, 303–330.

Angell, J. R. (1906). *Psychology: an introductory study of the structure and function of human conscious* (3rd ed., pp. 141–160). New York: Henry Holt.

Arias, C., Curet, C. A., Moyano, H. F., Joekes, S., & Blanch, N. (1993). Echolacation: a study of auditory functioning in blind and sighted subjects. *Journal of Visual Impairment & Blindness*, *87*, 73–77.

Arnell, K. A., & Jolicoeur, P. (1999). The attentional blink across stimulus modalities: evidence for central processing limitations. *Journal of Experimental Psychology: Human Perception and Performance*, *25* (3), 630–648.

Aronson, E., & Rosenbloom, S. (1971). Space perception in early infancy: perception within a common auditory-visual space. *Science*, *172*, 1161–1163.

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.

Ashmead, D. H., Wall, R. S., Eaton, S. B., Ebinger, K. A., Snook-Hill, M.-M., Guth, D. A., & Yang, X. (1998). Echolocation reconsidered: using spatial variations in the ambient sound field to guide locomotion. *Journal of Visual Impairment & Blindness*, *92*, 615–632.

Banati, R. B., Goerres, G. W., Tjoa, C., Aggleton, J. P., & Grasby, P. (2000). The functional anatomy of visuo-tactile integration in man: a study using pet. *Neuropsychologia*, *38*, 115–124.

Beranek, L. L. (1988). *Acoustical measurements* (Rev. ed.). Woodbury, NY: American Institute of Physics.

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review*, *5* (3), 482–489.

Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization* (Rev. ed.). Cambridge, MA: MIT Press.

Brainard, M. S., & Knudsen, E. I. (1993). Experience-dependent plasticity in the inferior colliculus: a site for visual calibration of the neural representation of auditory space in the barn owl. *Journal of Neuroscience*, *13*, 4589–4608.

Bregman, A. (1990). *Auditory scene analysis: the perceptual organization of sound*. Cambridge, MA: MIT Press.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, *89*, 244–249.

Broadbent, D. E. (1958). *Perception and communication*. New York: Pergamon Press.

Brochard, R., Drake, C., Botte, M., & McAdams, S. (1999). Perceptual organization of complex auditory sequences: effects of number of simultaneous subsequences and frequency separation. *Journal of Experimental Psychology: Human Perception and Performance*, *25* (6), 1742–1759.

Butchel, H. A., & Butter, C. M. (1988). Spatial attentional shifts: implications for the role of polysensory mechanisms. *Neuropsychologia*, *26*, 499–509.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in human heteromodal cortex. *Current Biology*, *10*, 649–657.

Cherry, C. (1959). *On human communication*. Cambridge, MA: MIT Press.

Cisek, P., & Turgeon, M. (1999). 'Binding through the fovea', a tale of perception in the service of action. *Psyche*, *5* http://psyche.cs.monash.edu.au/v5/psyche-5-34-cisek.html, accessed January 2000.

Clarke, S., Bellmann, A., Meuli, R. A., Assal, G., & Steck, A. J. (2000). Auditory agnosia and spatial deficits following left hemispheric lesions: evidence for distinct processing pathways. *Neuropsychologia*, *38*, 797–807.

Clifton, R. K. (1992). The development of spatial hearing in human infants. In L. A. Werner, & E. W. Rubel (Eds.), *Developmental psychoacoustics* (pp. 135–157). Washington, DC: APA Press.

Culling, J. F., & Summerfield, Q. (1995). Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay. *Journal of the Acoustical Society of America*, *98* (2), 785–797.

Darwin, C. J., & Hukin, R. W. (1999). Auditory objects of attention: the role of interaural time differences. *Journal of Experimental Psychology: Human Perception and Performance*, *25* (3), 617–629.

Deustch, D., & Roll, P. (1976). Separate "what" and "where" decision mechanisms in processing a dichotic tonal sequence. *Journal of Experimental Psychology: Human Perception and Performance*, *2* (1), 23–29.

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature*, *387*, 808–809.

Ehrenfels, C. von (1988). On 'gestalt qualities'. In B. Smith (Ed.), *Foundations of gestalt theory* (pp. 82–117). Munich, Germany: Philosophia Verlag. (Original work published 1890)

Gibson, J. J. (1979). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum.

Goldring, J., Dorris, M., Corneil, B., Balantyne, P., & Munoz, D. (1996). Combined eye-head gaze shifts to visual and auditory targets in humans. *Experimental Brain Research*, *111*, 68–73.

Goldstein, H. (1980). *Classical mechanics* (2nd ed.). Reading, MA: Addison-Wesley.

Hafter, E. R. (1997). Binaural adaptation and the effectiveness of a stimulus beyond its onset. In R. H. Gilkey, & T. R. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 211–232). Mahwah, NJ: Lawrence Erlbaum.

Handel, S. (1988). Space is to time as vision is to audition: seductive but misleading. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 315–317.

Handel, S. (1989). *Listening: an introduction to the perception of auditory events*. Cambridge, MA: MIT Press.

Helmholtz, H. L. F. (1954). *On the sensations of tone as a physiological basis for the theory of music* (2nd ed., A. J. Ellis, Trans.). New York: Dover. (Original work published 1885)

Hughes, H., Nelson, M., & Aronchick, D. (1998). Spatial characteristics of visual-auditory summation in human saccades. *Vision Research*, *38*, 3955–3963.

Jay, M. F., & Sparks, D. L. (1984). Auditory receptive fields in primate superior colliculus shift with changes in eye position. *Nature*, *309*, 345–347.

Jay, M. F., & Sparks, D. L. (1990). Localization of auditory and visual targets for the initialization of saccadic eye movements. In M. A. Berkley, & W. C. Stebbins (Eds.), *Comparative perception. Basic mechanisms*: Vol. 1. (pp. 351–374). New York: Wiley.

Johansson, G., & Orjesson, E. B. (1989). Toward a new theory of vision. Studies in wide-angle space perception. *Ecological Psychology*, *1*, 301–331.

Kanizsa, G. (1979). *Organization in vision: essays on Gestalt perception*. New York: Praeger.

Kant, I. (1996). *Critique of pure reason* (W. S. Pluhar, Trans.). Indianapolis, IN: Hackett. (Original work published 1787)

Kubovy, M. (1981). Concurrent-pitch segregation and the theory of indispensable attributes. In M. Kubovy, & J. Pomerantz (Eds.), *Perceptual organization* (pp. 55–99). Hillsdale, NJ: Lawrence Erlbaum.

Kubovy, M. (1988). Should we resist the seductiveness of the space:time:vision:audition analogy? *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 318–320.

Kubovy, M., Cohen, D., & Hollier, J. (1999). Feature integration that routinely occurs without focal attention. *Psychonomic Bulletin & Review*, *6*, 183–203.

Kubovy, M., Cutting, J. E., & McGuire, R. M. (1974). Hearing with the third ear: dichotic perception of a melody without monaural familiarity cues. *Science*, *186*, 272–274.

Kubovy, M., Holcombe, A., & Wagemans, J. (1998). On the lawfulness of grouping by proximity. *Cognitive Psychology*, *35*, 71–98.

Ladavas, E., & Pavani, F. (1998). Neuropsychological evidence of the functional integration of visual, auditory and proprioceptive spatial maps. *NeuroReport: an International Journal for the Rapid Communication of Research in Neuroscience*, *9*, 1195–1200.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: the embodied mind and its challenge to western thought*. New York: Basic Books.

Law, M. B., Pratt, J., & Abrams, R. A. (1995). Color-based inhibition of return. *Perception & Psychophysics*, *57* (3), 402–408.

Lenoble, J. S. (1986). Feature conjunctions and the perceptual grouping of concurrent tones (Unpublished doctoral dissertation, Rutgers – The State University of New Jersey, New Brunswick, NJ, 1986). *Dissertation Abstracts International, 47-06B*, 2654.

Macaluso, E., Frith, C., & Driver, J. (2000). Selective spatial attention in vision and touch: unimodal and multimodal mechanisms revealed by PET. *Journal of Neurophysiology*, *83*, 3062–3075.

McPherson, L., Ciocca, V., & Bregman, A. (1994). Organization in audition by similarity in rate of change: evidence from tracking individual frequency glides in mixtures. *Perception & Psychophysics*, *55* (3), 269–278.

Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action*. Oxford: Oxford University Press.

Mollon, J. (1995). Seeing colour. In T. Lamb, & J. Bourriau (Eds.), *Colour: art & science* (pp. 127–150). Cambridge: Cambridge University Press.

Mondor, T. A., & Breau, L. M. (1999). Facilitative and inhibitory effects of location and frequency cues: evidence of a modulation in perceptual sensitivity. *Perception & Psychophysics*, *61* (3), 438–444.

Mondor, T. A., Breau, L. M., & Milliken, B. (1998). Inhibitory processes in auditory selective attention: evidence of location-based and frequency-based inhibition of return. *Perception & Psychophysics*, *60* (2), 296–302.

Mondor, T. A., & Terrio, N. A. (1998). Mechanisms of perceptual organization and auditory selective attention: the role of pattern structure. *Journal of Experimental Psychology: Human Perception and Performance*, *24* (6), 1628–1641.

Monge, G. (1789). Mémoire sure quelques phénomènes de la vision. *Annales de Chimie*, *3*, 131–147.

Newton, I. (1999). *Mathematical principle of natural philosophy* (I. B. Cohen & A. Whitman, Trans.). Berkeley, CA: University of California Press. (Original work published 1726)

Object (1993). In *Oxford English Dictionary* (2nd ed.). (http://etext.lib.virginia.edu/etcbin/oedbin/

oed2www?specfile = /web/data/oed/oed.o2w&act =
text&offset = 287948343&textreg = 0&query = object, retrieved 1 October 1999)
Palmer, S. E. (1999). *Vision science: photons to phenomenology*. Cambridge, MA: MIT Press.
Peterson, M. A., & Gibson, B. S. (1994). Must figure-ground organization precede object recognition? An assumption in peril. *Psychological Science*, *5*, 253–259.
Posner, M. I. (1978). *Chronometric explorations of the mind*. Hillsdale, NJ: Erlbaum.
Potter, M. C., Chun, M. M., Banks, B. S., & Muckenhoupt, M. (1998). Two attentional deficits in serial target search: the visual attentional blink and an amodal task-switch deficit. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24* (4), 979–992.
Quinlan, P. T., & Hill, N. I. (1999). Sequential effects in rudimentary auditory and visual tasks. *Perception & Psychophysics*, *61* (2), 375–384.
Rauschecker, J. P. (1997). Processing of complex sounds in the auditory cortex of cat, monkey, and man. *Acta Oto-Laryngologica – Supplement*, *532*, 34–38.
Rauschecker, J. P. (1998a). Cortical processing of complex sounds. *Current Opinions in Neurobiology*, *288*, 516–521.
Rauschecker, J. P. (1998b). Parallel processing in the auditory cortex of primates. *Audiology and Neurootology*, *3*, 86–103.
Rorden, C., & Driver, J. (1999). Does auditory attention shift in the direction of an upcoming saccade? *Neuropsychologia*, *37*, 357–377.
Samuel, A. G. (1991). A further examination of attentional effects in the phonemic restoration illusion. *Quarterly Journal of Experimental Psychology*, *43A* (3), 679–699.
Scharf, B. (1998). Auditory attention: the psychoacoustical approach. In H. Pashler (Ed.), *Attention* (pp. 75–113). Hove: Psychology Press.
Scharf, B., Quigley, S., Aoki, C., Peachey, N., & Reeves, A. (1987). Focused auditory attention and frequency selectivity. *Perception & Psychophysics*, *42*, 215–223.
Shih, S., & Sperling, G. (1996). Is there feature-based attentional selection in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *22* (3), 758–779.
Spence, C., & Driver, J. (1994). Covert spatial orienting in audition: exogenous and endogenous mechanisms. *Journal of Experimental Psychology: Human Perception and Performance*, *20* (3), 555–574.
Spence, C., & Driver, J. (1997). Audiovisual links in exogenous overt spatial orienting. *Perception & Psychophysics*, *59* (1), 1–22.
Spence, C., & Driver, J. (1998). Auditory and audiovisual inhibition of return. *Perception & Psychophysics*, *60* (1), 125–139.
Spence, C., & Driver, J. (1999). Cross-modal attention. In G. W. Humphreys, & A. Treisman (Eds.), *Attention, space, and action* (pp. 130–149). New York: Oxford University Press.
Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.
Stoffregen, T. A., & Pittenger, J. B. (1995). Human echolation as a basic form of perception and action. *Ecological Psychology*, *7*, 181–216.
Stryker, M. P. (1999). Sensory maps on the move. *Science*, *284*, 925–926.
Tipper, S. P., Weaver, B., Jerreat, L. M., & Burak, A. L. (1994). Object-based and environment-based inhibition of return of visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *20* (3), 478–499.
Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.
Vicario, G. (1960). The acoustic tunnel effect. *Rivista da Psicologia*, *54*, 41–52.
Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*, 392–393.
Warren, R. M. (1982). *Auditory perception: a new synthesis*. New York: Pergamon Press.
Warren, R. M. (1984). Perceptual restoration of obliterated sounds. *Psychological Bulletin*, *96*, 371–383.
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 A. Palmer, A. Rees, A. Q. Summerfield, & R. Meddis (Eds.), *Psychophysical and physiological advances in hearing* (pp. 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., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *99*, 3749–3757.

Wertheimer, M. (1938). Laws of organization in perceptual forms. In W. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 71–88). London: Routledge & Kegan Paul. (Original work published 1923)

Wightman, F. L., & Jenison, R. (1995). Auditory spatial layout. In W. Epstein, & S. J. Rogers (Eds.), *Perception of space and motion* (2nd ed. pp. 365–400). San Diego, CA: Academic Press.

Wynn, V. T. (1977). Simple reaction time – evidence for two auditory pathways to the brain. *Journal of Auditory Research*, *17*, 175–181.

Young, E. D., Spirou, G. A., Rice, J. J., & Voigt, H. F. (1992). Neural organization and responses to complex stimuli in the dorsal cochlear nucleus. *Philosophical Transactions of the Royal Society of London, Series B*, *336*, 407–413.

Zheng, W., & Knudsen, E. I. (1999). Functional selection of adaptive auditory space map by $GABA_A$-mediated inhibition. *Science*, *284*, 962–965.