

*Research*

# Multistability, cross-modal binding and the additivity of conjoined grouping principles

Michael Kubovy\* and Minhong Yu\*

*Department of Psychology, University of Virginia, Charlottesville, VA 22904, USA*

We present a sceptical view of multimodal multistability—drawing most of our examples from the relation between audition and vision. We begin by summarizing some of the principal ways in which audio-visual binding takes place. We review the evidence that unambiguous stimulation in one modality may affect the perception of a multistable stimulus in another modality. Cross-modal influences of one multistable stimulus on the multistability of another are different: they have occurred only in speech perception. We then argue that the strongest relation between perceptual organization in vision and perceptual organization in audition is likely to be by way of analogous Gestalt laws. We conclude with some general observations about multimodality.

**Keywords:** multistability; cross-modal binding; additivity; conjoined grouping principles; auditory necklaces; indispensable attributes

## 1. INTRODUCTION

This paper is an attempt to refocus the question of the relation between multistable stimuli in vision and audition. The question before us is whether *multistability in one affects multistability in the other*. We believe that such effects are uncommon; hence, we focus on parallels between the behaviour of multistable stimuli in vision and multistable stimuli in audition. We will show that in both domains the interaction of sources of multistability (grouping principles) is subject to similar laws of additivity.

The question of cross-modal influences in the perception of multistable stimuli requires prior answers to two other questions:

- Let us assume that stimuli in the two modalities undergo binding if and only if they seem to come from a single source or event in the environment. If the assumption holds, is the binding of the stimuli in the two modalities a prerequisite for such a cross-modal influence? We believe that the answer to this question is Yes.
- Can one design a visual stimulus that is multistable to the eye and an auditory stimulus that is multistable to the ear *such that they can undergo binding*? We believe that the answer to this question is (with the exception of speech) No.

Because of these answers, we are sceptical regarding cross-modal influences on the perception of multistable stimuli. Nevertheless, as we will show later, there are fascinating and important cross-modal parallels between multistable stimuli.

\* Authors for correspondence (kubovy@virginia.edu; minhongyu@virginia.edu).

One contribution of 10 to a Theme Issue ‘Multistability in perception: binding sensory modalities’.

We begin our paper by listing four conditions for the binding of auditory and visual stimuli. Then we compare and contrast emergence and binding. Then we argue that only emergent properties can be multistable. We continue by asking under which conditions two multistable stimuli can affect each other. We then answer in the negative the question of whether there are stimuli that are multistable in each modality *and* lend themselves to binding according to these determinants. Next, we describe parallel results we have obtained in vision and in audition regarding the additivity of organizational cues when we apply them conjointly to multistable stimuli within each modality. We conclude that the exploration of the scope and limits of cross-modal analogies in the perception of multistable stimuli is a fruitful direction for future research.

## 2. BINDING

For cross-modal influence to occur, one or more of four conditions must hold: synchrony, common embodiment, common source or causality.

### (a) *Binding by synchrony*

Two notable examples of binding by synchrony are the McGurk effect [1] and the ventriloquism effect [2,3]. Both involve speech. In the McGurk effect, a voice repeats the same syllable, such as /b/, whereas visible lip movements correspond to other syllables, such as /g/. The nature of the lip movements has a strong effect on our perception of the speech stimulus. In the ventriloquism effect, a sound comes from one location in space, whereas the lip movement comes from another location. The sound seems to emanate from the location of the mouth. Both of these effects diminish (within limits—see van Wassenhove *et al.* [4] regarding the McGurk effect) with asynchrony between sound and lip movements [5–11].

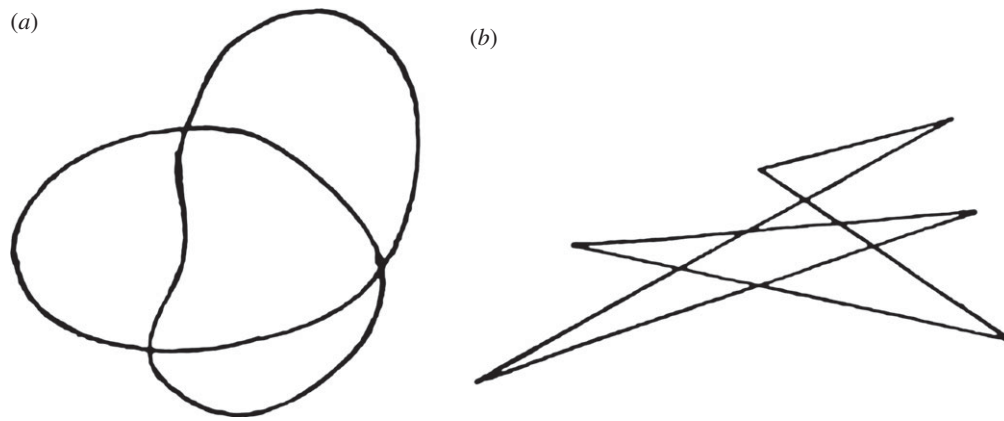


Figure 1. Köhler's demonstration of binding by common embodiment. (a) The *maluma* figure. (b) The *takete* figure.

### (b) *Binding by common embodiment*

A beautiful illustration of cross-modal influences is the Köhler 'takete/maluma' demonstration [12,13] (in the earlier version it was 'takete/baluba'). He would show a person the two drawings in figure 1 and ask, Which is takete and which is maluma? People are near unanimous that the rounded form goes with *maluma*, the angular one with *takete*. The effect holds across different linguistic groups [14] and extends to preliterate toddlers [15].

The phenomenon is probably due to analogies between the gestures required to generate the drawing and produce the sound. The angular drawing requires abrupt reversals of direction; analogously, the sound requires the tongue to sharply strike the palate. In contrast, the rounded drawing requires a single flowing gesture; analogously, the corresponding sound requires a gentle touching of the lips.

Experimental evidence of this phenomenon is now available. Parise & Spence [16] varied the stimulus onset asynchrony (SOA) between the sound and the image, and asked the subject to say whether the sound preceded the image or not. The resulting psychometric functions showed that when the pairs were consistent (a rounded shape paired with a more smooth-sounding pure tone or an angular shape paired with a rough-sounding square wave), participants were more tolerant of SOA, i.e. they were less likely to say that inconsistent pairs were synchronous. Thus, they were able to show greater binding of consistent auditory and visual stimuli than inconsistent ones.

### (c) *Binding by common source*

The relation between vision and touch offers an example of binding by common source. There is a correlation between the likelihood that visual and haptic signals come from different objects and the spatial separation between them. Gepshtein *et al.* [17] asked how this separation affects the binding. Their stimuli were two variable-slant virtual planes presented multimodally, such that the two planes were parallel within and between modalities, but the apparent distance between the planes differed between modalities. In the visual modality, they generated the planes by a random-element stereogram so that they appeared to be transparent; in the haptic modality, they generated the planes by a force-feedback device connected to thimbles into which the participants inserted their

index finger and thumb. The observers' task was to view these surfaces with both eyes and/or 'grasp' them with the index finger (from above) and thumb (from below) and estimate the distance between them. They first determined for each observer the conditions under which the increase in discrimination precision with two modalities relative to performance with one modality is maximal. They manipulated the distance between the signals, and found that discrimination precision was optimal when the signals came from the same location.

### (d) *Binding by causality*

Suppose that a marimba player strikes a key and follows it with a long or a short gesture. The perceived duration of the sound is longer if the gesture is long than if it is short [18,19]. Schutz & Kubovy [18] demonstrated that this effect was due to cross-modal causality, by first showing that the effect depended on the percussive element of the sound. For example, when the sound associated with the long and short gestures (marimba or piano), the effect was present. But when they substituted a non-percussive sound—clarinet, violin or voice—for the percussive sound, the effect disappeared. Thus, they established the first condition of causality: appropriateness. A second condition is temporal order: a cause must precede its effect. They showed that even if they played the sound 400 or 700 ms after the visible impact, the effect was present. But if the sound preceded the impact by 400 or 700 ms, then the effect vanished.

In another experiment, Armontrout *et al.* [20] replaced the video of the percussionist with an animation in which only a white disk on a black background was visible: the disk followed the path of the mallet head, and no impact surface was visible. Despite this extreme simplification, the results held. Thus, the effect does not depend on learned context, such as a person causing the motion.

However, when they reversed the direction of the trajectory of the disk after impact—so that it seemed to go below the impact surface—the effect vanished, even though only the direction changed while velocities and accelerations remained identical. This suggests that this causality effect is hard-wired.

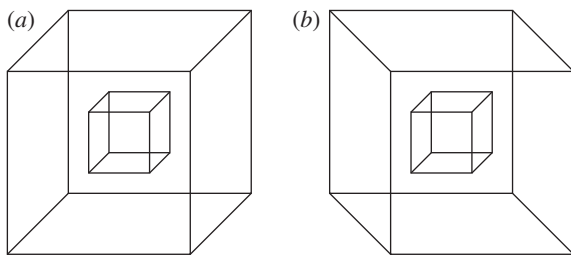


Figure 2. Examples of coherent and unrelated ambiguities: the Adams & Haire [21] nested cubes. (a) Same orientations. (b) Opposite orientations.

### 3. MULTISTABILITY

#### (a) *Related and unrelated ambiguities*

On our way to determining whether multistable stimuli in different modalities can affect each other, we inquire about the conditions under which multistable stimuli affect each other *within* a modality. To that end, we distinguish between *coherent* and *unrelated* ambiguities. Coherent ambiguities share an *emergent property*.

Consider the possibility that the reversals of one Necker cube might affect the reversals of another. Adams & Haire [21] studied the relative reversal rates of a small Necker cube nested in a larger one under two conditions: in one, the orientations of the two cubes were the same (figure 2a); in the other, they were opposite (figure 2b).

In both conditions, none of the participants reported a higher reversal rate for the larger cube. When their orientations were the same, their reversal rates were similar: the nested cube reversed faster for only 5 per cent of the participants. But when the orientations of the small cube and the large cube were opposite, there was no relation between their reversal rates: the nested cube reversed faster for 87.5 per cent of the participants.

The drawing of the cubes in the same orientation has coherent ambiguities. At a given moment, each produces one of two emergent properties—cube seen from right and above or cube seen from left and below. When both fit into the same spatial framework, they share this emergent property. This allows them to be seen as a single object: a solid cube with a cubic chamber inside. There is no question of independent multistability of these two cubes because they have no separate perceptual standing. In contrast, when the two cubes are in opposite orientations, one cannot make them coherent—at a given moment the emergent property produced by one is different from the emergent property produced by the other. We call these ambiguities unrelated.

A more striking example of unrelated ambiguities (figure 3a) comes from Bradley & Petry [22]. The figure is multistable in two ways.

First, it shows the usual multistability of the Necker cube—cube seen from right and above or cube seen from left and below.

Second, it is multistable with respect to depth, or layering. In both interpretations, we see the Necker cube as a white paper cutout. But in one, we see it floating above eight black disks painted on white paper (figure 3b), whereas in the other, we see it through eight holes in a sheet of paper against a black backdrop

seen through the holes behind the cutout (figure 3c,d). In the first case, we see the edges of the cube uninterrupted; this is the phenomenon of modal completion discovered by Schumann [23] and made important by Kanizsa [24,25]. In the second case, the completion is amodal, we see that the page occludes parts of the cutout, but we see the cutout as being complete.

Because unrelated ambiguities are logically and perceptually independent, we conjecture that unrelated ambiguities never affect each other.

These notions of coherence and independence of ambiguous patterns are still intuitive, but they offer a principled way to think about the influence of one multistable pattern on another.

#### (b) *Cross-modal effects of one multistable stimulus on the multistability of another*

We know of only one example of *cross-modal coherent ambiguities*. The phenomenon in question—the verbal transformation effect [26]—occurs both in audible speech (AVTE) and in visible speech (VVTE). In the AVTE, as we listen to a repeatedly played syllable, such as a /psə/, after a number of repetitions we hear it as a /səp/ and vice versa. In the VVTE, discovered by Sato *et al.* [27, experiment 1], as we look at images of moving lips that we see as speaking a /psə/, we eventually see it as saying /səp/. Sato *et al.* [27] found that reversals in the VVTE occurred in synchrony with reversals in the AVTE.

Why are these ambiguous stimuli coherent? The reason is that they both can produce the same emergent phonological entities: a /psə/ and a /səp/.

We are doubtful that other cases exist because of fundamental differences between perceptual organization in the two modalities. To illustrate, we compare two patterns: one visual, the other auditory (figure 4).

The visual pattern (figure 4a) forms a tight grouping of three items, and a loose grouping of two. We cyclically play the auditory pattern, represented in figure 4b, called an *auditory necklace* (to which we will return later). If we extended figure 4a indefinitely to the right and left, and played figure 4b repeatedly, either diagram would represent either stimulus.

In perceiving both patterns, groupings of three elements emerge. Grouping by proximity (spatial or temporal, as the case may be) emerges in both modalities; in that respect, they are analogous. But the visual pattern produces a further set of emergent properties based on mirror symmetry (a topic explored by Strother & Kubovy [28]), which are not present when we listen to the auditory stimulus because that requires audition to reverse time. In contrast, the auditory pattern produces its own emergent property, which has no analogue in the perceptual organization of the visual stimulus: an accent on the initial note of the pattern (the first note of the run of three), its ‘clasp’.

What do the stimuli in the two modalities have in common that could allow them to influence each other? What would make these ambiguous patterns coherent, and not independent? Either *space* or *time*.

#### (i) *Space*

Space cannot be the common medium for multistability in the two modalities, thanks to the theory of

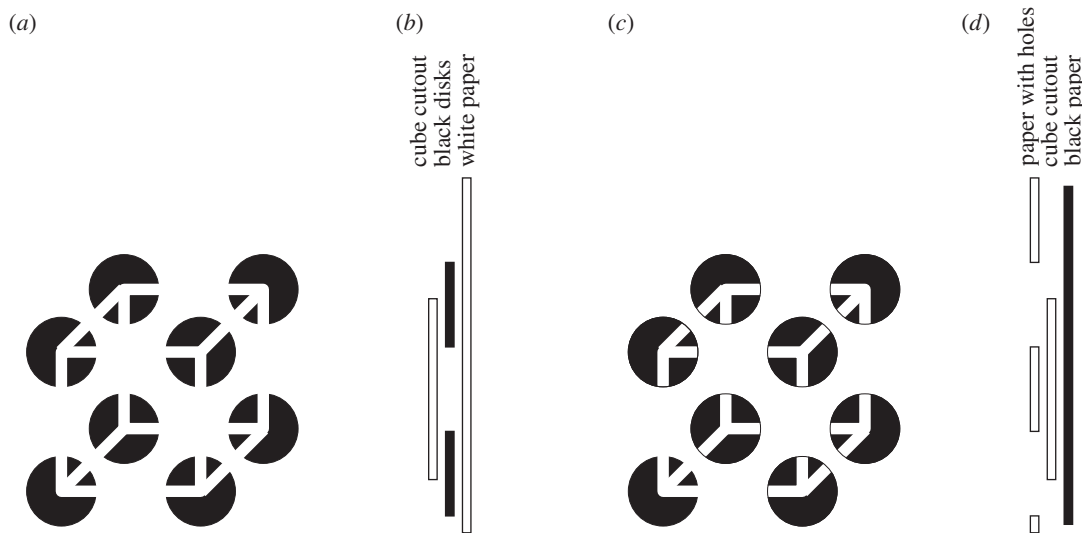


Figure 3. An example of unrelated ambiguities: *Ambiguity 1*: is the Necker cube interpreted as a cube seen from a vantage point above to its right or from a vantage point below to its left? *Ambiguity 2*: is the paper cutout of a Necker cube seen floating in front of eight black disks painted on a white background (illustrated in figure 3b) or is it seen through eight holes in a white surface, against a black backdrop? (illustrated in figure 3d and rendered in figure 3c) (a) The Bradley & Petry [22] cube. (b) Default interpretation: modal completion. (c) The modified Bradley & Petry [22] cube. (d) Alternative interpretation: amodal completion).

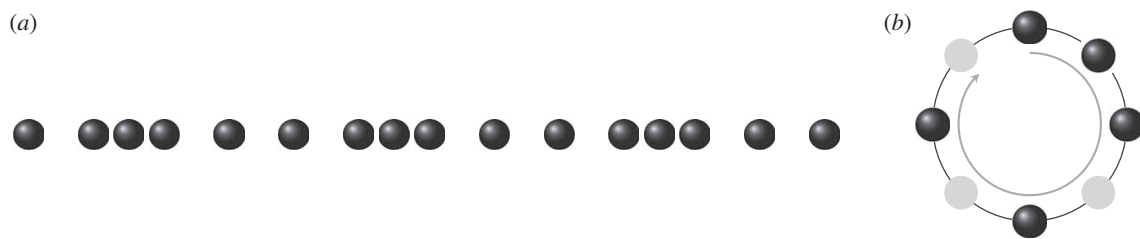


Figure 4. Two examples of grouping. (a) Visual. (b) Auditory.

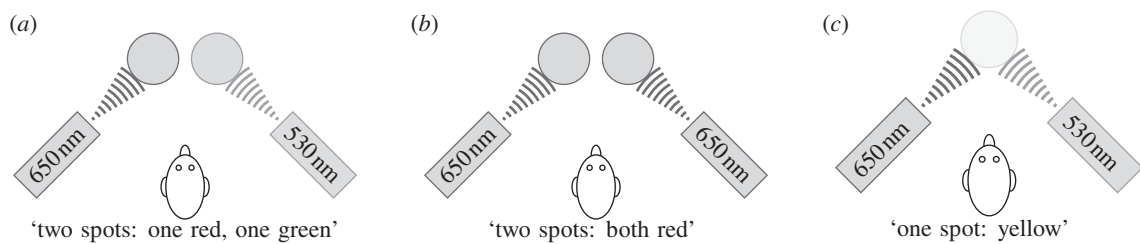


Figure 5. Theory of indispensable attributes: the visual thought-experiment. (a) Start. (b) Collapse over wavelength: not indispensable. (c) Collapse over space: indispensable.

indispensable attributes [29–32]. This theory specifies the aspects of stimulation required for perceptual numerosity in each modality. This is critical for understanding perceptual organization. Without multiple perceptible entities, there is nothing to organize, no Gestalts, no multistability.

Imagine the following thought-experiment: we ask an observer who is looking at two coloured spots of light (figure 5a), How many entities are visible? We assume that the observer answers, Two. (If the observer gives any other answer, then the experiment is invalid.) If we collapse the display with respect to wavelength (figure 5b), then the observer will still say, Two. Hence, *wavelength is not an indispensable attribute for visual numerosity*. But if we collapse over space (figure 5c), then the observer will respond

(because of colour metamerism), One. Hence, *spatial location is an indispensable attribute for visual numerosity*.

The auditory thought-experiment runs as follows: we ask an observer who hears two sounds played over two loudspeakers (figure 6a), How many entities are audible? We assume that the observer answers, Two. If we collapse the display with respect to space (figure 6b), then the observer will still say, Two. Hence, *spatial location is not an indispensable attribute for visual numerosity*. But if we collapse over frequency (figure 6c), the observer will respond (because of auditory localization mechanisms), One. Hence, *frequency is an indispensable attribute for auditory numerosity*.

Because space cannot serve as a common medium for the mutual influence of an auditorially multistable and a visually multistable stimulus, we turn to time.



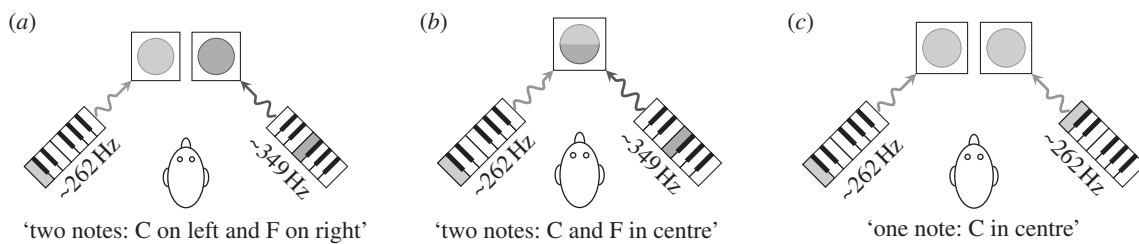


Figure 6. Theory of indispensable attributes: the auditory thought-experiment. (a) Start. (b) Collapse over space: not indispensable. (c) Collapse over frequency: indispensable.

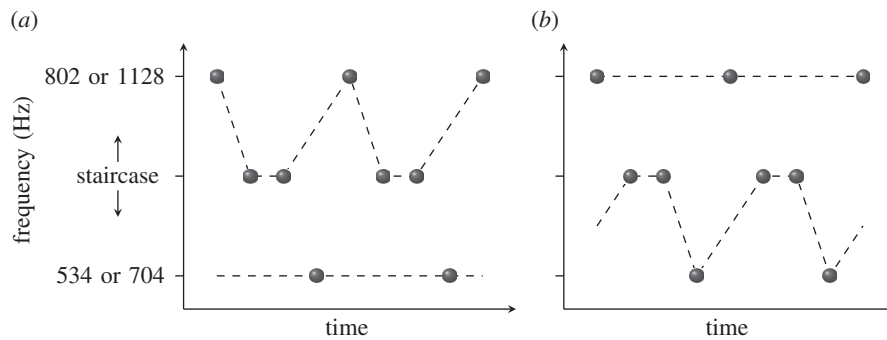


Figure 7. Four-tone bistable tone sequence used by Cook & Van Valkenburg [33]. (a) The middle tones are heard grouped with the high tone, leaving the low tone perceptually isolated. (b) The middle tones are heard grouped with the low tone, leaving the high tone perceptually isolated.

#### (ii) Time

One can construct a sequence of visible events that form a multistable stimulus accompanied by a sequence of audible events that also form a multistable stimulus, and ask whether the perceptual organization of one affects the perceptual organization of the other. Cook & Van Valkenburg [33] have provided the best evidence to date that they do not. They used a four-tone bistable sequence (figure 7). In such sequences, one hears the middle tone group either with the high or with the low tone, leaving the other tone perceptually isolated. Because the frequency of the middle tone determines the perceptual organization of this stimulus, they used an adaptive method to determine the frequency at which the sequence is most ambiguous, where the probabilities that the middle tones group with the high tone and the low tone are equal.

Concurrently with the sounds, they flashed two horizontally separated light-emitting diodes three times on the left and once on the right (or vice versa) to form two temporal groupings at two spatial locations. The main variable in the experiment had two levels: the single flash coincided with the high tone or it coincided with the low tone.

The participants reported whether they heard the pattern of figure 7a or 7b. The results showed that the visual perceptual organization had no effect on the auditory perceptual organization. In a further experiment, they obtained a temporal ventriloquism effect (perceived simultaneity of the two isolated events) even when the tone and the light were 120 ms out of sync. Together, these findings suggest that unimodal grouping precedes interaction between auditory and visual stimuli.

Similarly, Pressnitzer & Hupé [34] have shown that although ambiguous auditory streaming [35,36] and

moving visual plaids [37] follow similar temporal courses, they alternate independently.

#### (iii) Binocular rivalry

Of course, unambiguous stimulation in one modality can have an effect on multistability in another. These examples are not central to our concerns for two reasons. First, they are not about one multistable stimulus affecting another. Second, they are generally about binocular rivalry, which is probably the product of different mechanisms than the classic multistable phenomena that we have been discussing [38]. For the sake of completeness, we briefly review them.

Regarding the effect of haptics, Binda *et al.* [39] and Lunghi *et al.* [40] found that orientation-matched haptic stimuli can extend the dominance and reduce the suppression of visual stimuli. Maruya *et al.* [41] found that the congruence between voluntary action and the motion of one of the rival visual stimuli prolongs the dominance times of that stimuli.

For audition, Conrad *et al.* [42] found that when the perceived unambiguous motion of a sound is directionally congruent with one of two binocularly rivalrous visual motion stimuli, it prolongs the dominance periods of the directionally congruent visual motion percept. Kang & Blake [43] found that an unambiguous amplitude-modulated sound can influence the rate of alternation in a synchronously flickering binocularly rivalrous grating.

## 4. CROSS-MODAL ANALOGIES IN THE PERCEPTION OF MULTISTABLE STIMULI

We now can state our view of the relationship between the perceptual organization of multistable stimuli in

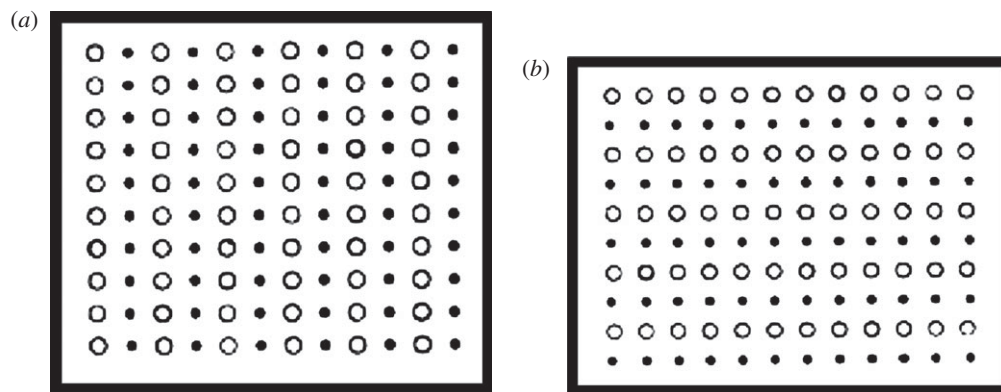


Figure 8. Wertheimer's [44] rectangular dimotif dot lattices apply grouping by proximity and grouping by similarity concurrently to the same stimuli. (a) Wertheimer's figure (xii): both proximity and similarity favour columns (in terms defined in figure 9,  $|b|/|a| = 1.083$ ). (b) Wertheimer's figure (xiii): proximity favours columns (in terms defined in figure 9,  $|b|/|a| = 1.104$ ) but similarity favours rows.

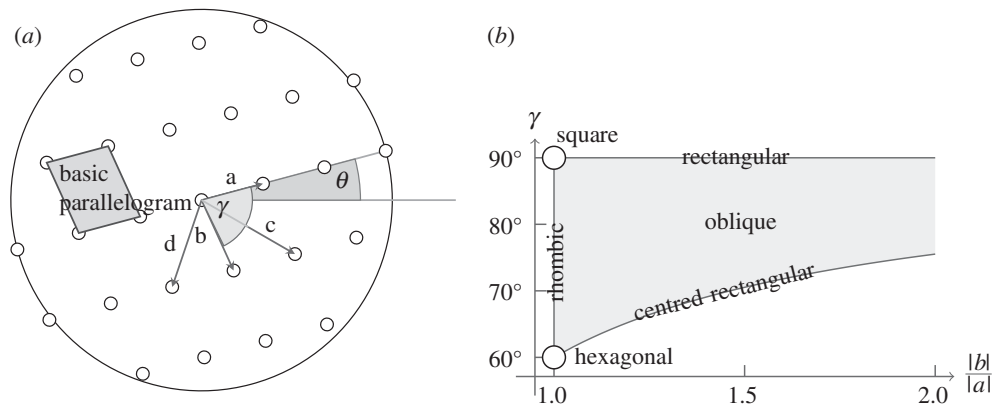


Figure 9. Dot lattices. (a) Defining features. (b) Two-dimensional space and nomenclature.

vision and audition: they proceed independently but show notable parallels. In particular, we find that in both vision and audition, when one applies more than one grouping principle to the same stimulus (an operation we call *conjoining*), they act additively.

#### (a) *Conjoined grouping principles*

As soon as Wertheimer [44] formulated the laws of perceptual organization in vision (anticipated, as Ash [45], points out, by Schumann [46] and Müller [47]), he pitted grouping by proximity against grouping by similarity, as shown in figure 8 [48, pp. 74–75]. In so doing, he showed how to *conjoin* two grouping principles, i.e. apply both principles to the same collection of elements.

Naturally enough, this raised the question of the joint effect of conjoined grouping principles: are they additive, i.e. act independently, or are they non-additive, i.e. create a new emergent property? This is a task for mathematical models.

#### (b) *Quantifying grouping principles in vision*

Our goal in this section is to answer the following question: When grouping by proximity and grouping by similarity are concurrently applied to the same visual pattern, what rule governs their joint application?

Later, we will ask the same question about grouping in auditory perception.

To study grouping, researchers (starting with Wertheimer [44], our figure 8) often used *dot lattices*. To describe the earlier research, we use the taxonomy of dot lattices (Kubovy [49], who extended the work of Bravais [50]). Figure 9 summarizes the taxonomy.

A dot lattice is a collection of dots in the plane that is invariant under two translations,  $\mathbf{a}$  (with length  $|\mathbf{a}|$ ) and  $\mathbf{b}$  (whose length is  $|\mathbf{b}| \geq |\mathbf{a}|$ ). These two lengths, and the angle between the vectors,  $\gamma$  (constrained for purely geometric reasons by  $60^\circ \leq \gamma \leq 90^\circ$ ), define the *basic parallelogram* of the lattice, and thus the lattice itself. The diagonals of the basic parallelogram (shown in figure 9a, but not represented in figure 9b) are  $\mathbf{c}$  and  $\mathbf{d}$  (where  $|\mathbf{c}| \geq |\mathbf{d}|$ ). In its *canonical orientation*,  $\mathbf{a}$  is horizontal; the angle  $\theta$  (measured counterclockwise) is the measure of the *orientation* of a dot lattice ( $\theta = 15^\circ$  in figure 9a); we call  $|\mathbf{a}|$  the *scale* of the lattice.

If we are not interested in the scale of a lattice, then we can locate dot lattices in a two-dimensional space with dimensions  $|b|/|a|$  and  $\gamma$  (figure 9b). In this space, we can locate six different types of lattices, which differ in their symmetry properties.

Before we can explore how the effects of two grouping principles combine when we apply them conjointly, we

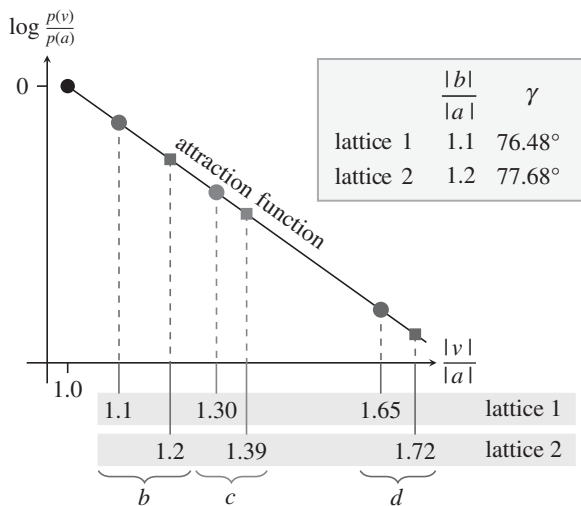


Figure 10. The attraction function of grouping by proximity.

must measure the strength of grouping by proximity on its own.

Oyama [51] had shown how to measure the strength of grouping by proximity without recourse to grouping by similarity. Using rectangular dot lattices at fixed orientation, he recorded the amount of time subjects reported seeing the competing vertical and horizontal groupings. He found that the ratio of the time they saw the vertical and the horizontal organizations is a power function of the ratio of the vertical and horizontal distances.

Using dot lattices at near-equilibrium, Kubovy & Wagemans [52] and Kubovy *et al.* [53] demonstrated that we can understand grouping by proximity as the outcome of a probabilistic competition among potential perceptual organizations. Figure 10 shows their results schematically. Consider two dot lattices (in which we assume that  $|a| = 1$ ): in the first,  $|b| = 1.1$  and  $\gamma = 76.48^\circ$ ; in the second  $|b| = 1.2$  and  $\gamma = 77.68^\circ$ . The corresponding lengths of  $c$  are  $|c| = 1.3$  and 1.39, and the lengths of  $d$  are  $|d| = 1.65$  and 1.72.

It is remarkable that all the values of  $\log[p(v)/p(a)]$  (where response  $v \in \{b, c, d\}$ ) fall on the same line, which we call the *attraction function*. Its slope,  $\xi$ , is a person-dependent measure of sensitivity to proximity:

$$\log \frac{p(v)}{p(a)} = \xi \left( \frac{|v|}{|a|} - 1 \right),$$

where vector  $v \in \{b, c, d\}$ . We define  $\phi(v) = p(v)/p(a)$ , and call it *strength of grouping along v*. Thus

$$\phi(v) = e^{\xi(|v|/|a| - 1)}.$$

(i) *The additivity of grouping principles in vision*

Once we have determined how grouping varies as a function of relative distance, we can determine the effect of conjoined grouping principles. This requires us to plot a family of attraction functions and determine whether we can make these functions parallel, suggesting additivity (and hence independence). Kubovy & van den Berg [54] acquired their data using dimotif dot lattices as shown in figure 11. Using these stimuli, they obtained additive results shown schematically in figure 12.

When we conjoin two principles of grouping, what determines the perceived outcome? At least with respect to proximity and similarity, they affect the outcome independently. The conjoint effect of these two grouping principles is equal to the sum of their independent effects. Just as with the attraction function, the additivity of laws is as far as one can imagine from the spirit of Gestalt, even though grouping is an emergent property, and thus a Gestalt.

(c) *Quantifying grouping principles in audition*

The segmentation of auditory rhythmic patterns is an important function of the auditory perceptual system, particularly in the processing of speech and music [55–57]. For example, if ♪ represents a note and 7 represents a rest, then

... ♪ ♪ ♪ 7 7 ♪ ♪ ♪ ♪ ♪ 7 7 ♪ ♪ ♪ 7 7 ♪ ♪ ♪ 7 ...

is a rhythmic pattern with eight repeating elements. Such patterns can be multistable. You can hear the preceding pattern as either a cyclic version of

... ♪ ♪ ♪ 7 7 ♪ ♪ ♪ 7 7 ♪ ♪ ♪ 7 7 ♪ ♪ ♪ 7 ... ,

or of

... ♪ ♪ ♪ 7 7 ♪ ♪ ♪ 7 7 ♪ ♪ ♪ 7 7 ♪ ♪ ♪ 7 ...

We call such stimuli auditory necklaces (ANs) because they are cyclical sound patterns of equal-duration beats (notes or rests) that are best visualized as arranged in uniform spacing around a circle (figure 14). Assume that when we play an AN, we play it fast at first, eventually decelerating into a moderate tempo, in order not to bias the listener—which note will the listener hear as the initial note (its ‘clasp’)?

To describe an AN, we need the notions of *run*, which is a sequence of consecutive notes, and *gap*, which is a sequence of consecutive rests. We describe an AN by

- its length,  $n$ , and
- the number of *runs* it contains (which is the same as the number of *gaps* it contains).

Figure 14 shows an example: ♪ ♪ ♪ 7 7 ♪ ♪ 7, which we code 11100110 (where 1s are notes and 0s are rests). It has two runs (111 and 11) and two gaps (00 and 0).

Garner and co-workers [58–61] proposed two principles to account for the segmentation of ANs—the *run* and the *gap* principles. According to the *run* principle, the first note of the longest run will be the clasp, whereas according to the *gap* principle, the first note following the longest gap will be the clasp. Most listeners hear the AN in figure 14 either as 11100110 (*run* principle) or as 11011100 (*gap* principle).

(i) *The additivity of organizational principles in audition*

Previous researchers have used two tasks to study the segmentation of ANs:

- *Reproduction*. One can have participants report each pattern by tapping it out or by writing it down, from which one can infer the clasp note [58–61]. This made data collection slow and as a result the datasets were too small to be adequately modelled.

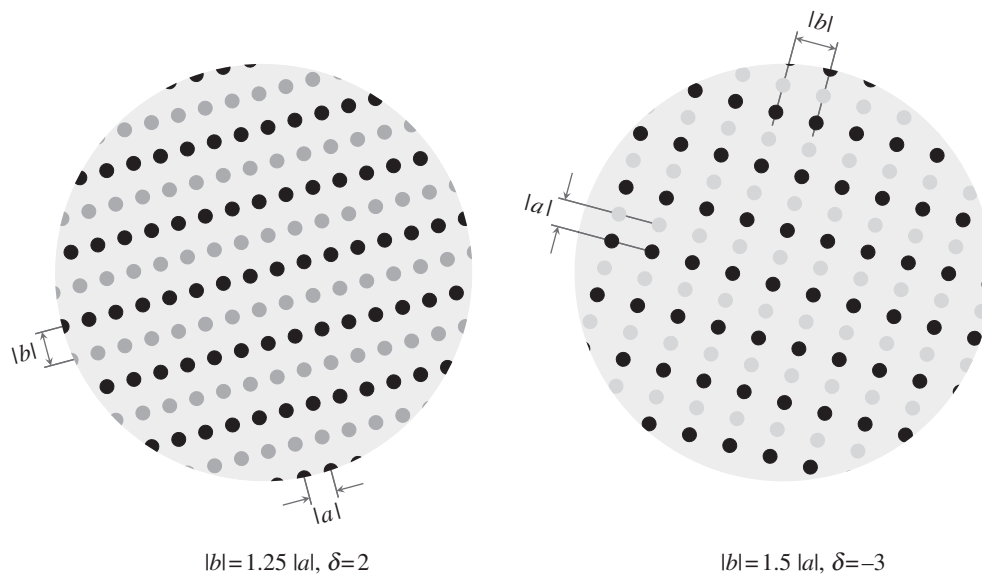


Figure 11. Two dimotif dot lattices. In both, grouping by proximity favours  $a$ , but more weakly in the dot lattice on left, where  $|a| = 1.25|b|$ , than in the dot lattice on the right, where  $|a| = 1.5|b|$ . In the dot lattice on the left, grouping by similarity favours  $b$  ( $\delta > 0$ , where  $\delta$  is a measure of dissimilarity between two kinds of dots), whereas in the dot lattice on the right it favours  $a$  ( $\delta < 0$ ), but because the differences between dot-colours are smaller on the left than on the right, the strength of the grouping by similarity ( $\delta = 2$ ) on the left is smaller than the strength of grouping by similarity on the right ( $\delta = -3$ ).

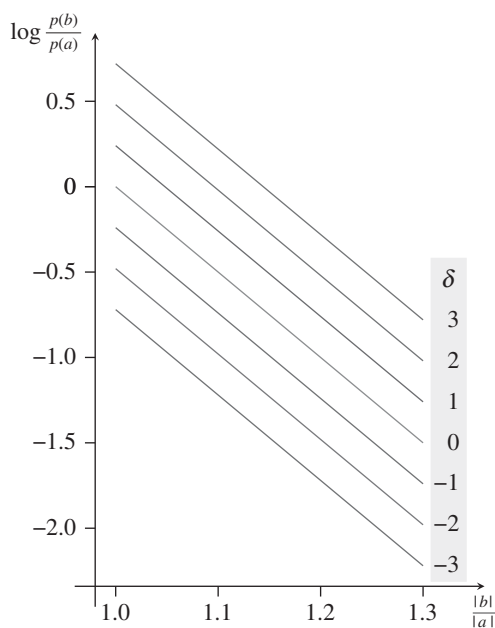


Figure 12. A schematic of the results obtained by Kubovy & van den Berg [54] using the dimotif lattices described in figure 11, showing that the conjoined effects of proximity and similarity are additive when choice probabilities are represented as log-odds. The line for  $\delta = 0$  (light grey) corresponds to the attraction function in figure 10, i.e. all the dots have the same colour, and therefore, grouping by similarity cannot affect the results. The results are equivalent to the multiplicative model of figure 13.

— *Tap the clasp.* One can ask participants to strike a key to coincide with the moment of the clasp [62]. Although the amount of data collected by this method can be copious, it is noisy because participants have trouble synchronizing their taps with the onset of the clasp beat. Another drawback of this task is that the motor control required might affect the perception of the patterns.

We improved upon these procedures by devising a procedure that did not require participants to synchronize their responses with events in real-time. A circular array of  $n$  icons (where  $n$  is the length of the AN) appeared on the computer screen at the beginning of each trial (figure 15). Each icon was associated with one beat (a note or a rest) of the AN. The position of the icon corresponding to a potential clasp can be in any of the  $n$  locations around the circle with equal probability. As each beat of an AN occurred, a square highlighted the icon associated with that note and then moved clockwise to coincide with the next beat.

The participants' task was to click on the icon associated with the beat they heard as the clasp. They could do this at any moment, without regard for the currently highlighted icon; i.e. they did not have to synchronize their response with a sound or an event shown on the screen.

In the research we are briefly reporting here (as yet unpublished), we used a sample of ANs with two runs (a sequence of one or more consecutive notes) and two gaps (a sequence of one or more consecutive rests). We called the runs  $A$  and  $B$  (denoted  $r_A$  and  $r_B$ ). We denoted the gap preceding  $r_A$  as  $g_A$ , and the gap preceding  $r_B$  as  $g_B$ . We manipulated the lengths of the runs and the gaps.

The response variable was binomial—choosing  $r_A$  or  $r_B$ .<sup>1</sup> To quantify the run and gap principles, we calculated the *run-length difference*:  $\Delta r = \text{length}(r_A) - \text{length}(r_B)$  and the *gap-length difference*:  $\Delta g = \text{length}(g_A) - \text{length}(g_B)$ .

The best model<sup>2</sup> of the data is additive with  $\Delta r$  and  $\Delta g$  as predictors. Figure 16 shows the effect of  $\Delta r$  and  $\Delta g$  on the probability of choosing  $r_A$  as the clasp (in log-odds scale). The parallel lines demonstrate the additivity of the two principles, and for reasons as yet unknown the effect of gap length difference is much larger than the effect of run length difference.



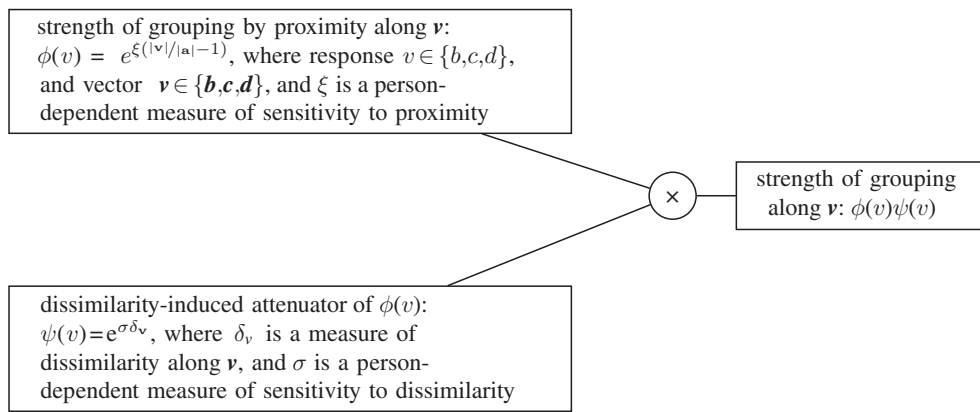


Figure 13. Multiplicative model of conjoined grouping that implies additivity of grouping by proximity and grouping by similarity when choice probabilities are represented as log-odds, as in figure 12.

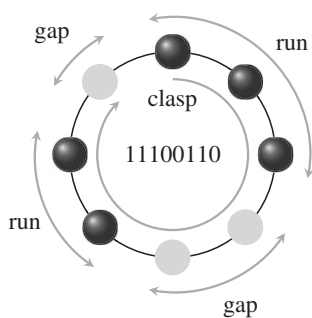


Figure 14. An auditory necklace 11100110 of length  $n = 8$ ; i.e. it is eight beats long. It has five notes and three rests, grouped into two sequences of notes called *runs*, and two sequences of rests called *gaps*.

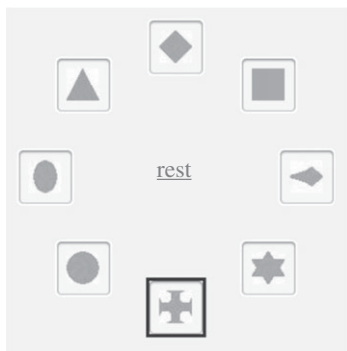


Figure 15. The response screen. At the moment depicted, the cross is highlighted.

**5. FINAL COMMENTS AND FUTURE DIRECTIONS**

We have previously laid out our views of the intricate relationships between visual perceptual organization and auditory perceptual organization [31]. There we argued for three kinds of relationships between the two modalities:

- a duality, based on the theory of indispensable attributes, and the differing functions of these modalities. The visual system’s primary function is to identify reflective surfaces, and its secondary function is to take sources of light into account. The auditory system’s primary function is to

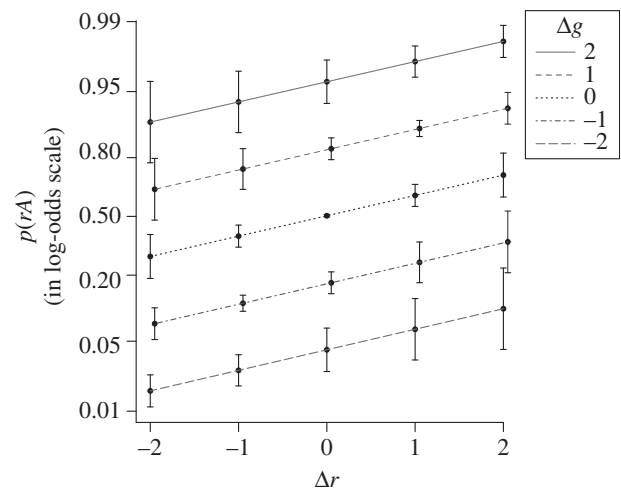


Figure 16.  $p(r_A)$  as a function  $\Delta r$  and  $\Delta g$ . Error bars span  $\pm 1$  s.e.

identify sources of sound, and its secondary function is to take reverberations from surfaces into account;

- a dependence, according to which auditory localization is primarily in the service of visual orientation; and
- audio-visual object formation, which is the result of binding (as described in the first part of this paper).

Here we have shown that when we conjoin Gestalt principles within a modality, they operate additively, i.e. do not form a new emergent property, a new Gestalt. In the terminology we developed earlier, they are *independent*, not *coherent*. More generally, we claim that binding and emergence are different phenomena: when an auditory and a visual input undergo binding, *a new property does not emerge*. The one exception is speech: when a person’s lips move and we hear a concurrent sound, we experience a common emergent property—meaning. In contrast, even though we experience the sight of a hammer hitting a nail and the sound of the impact as one, the percept retains two fully separable modal aspects; it is not a trans-modal experience.

To make this idea clear, we turn to Gibson, whose position was tantamount to the claim that *fire* was an emergent property of four kinds of stimulation:

A fire is a terrestrial event with flames and fuel. It is a source of four kinds of stimulation. 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 perceptual systems are equivalent. If the perception of fire were a compound of separate sensations of sound, smell, warmth and colour, they would have had to be associated in past experience to explain how one of them could evoke memories of the others. But if perception is simply the pickup of information [it] will be the same whatever system is activated.

... the problem of perception is not 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 [63, pp. 54–55])

A byproduct of our analysis may be a new understanding of why Gibson's approach to multimodal integration may not be fruitful.

These ideas may also clarify the issue of multimodal multistability. For multimodal multistability to occur, there must exist trans-modal Gestalts that can undergo multiple forms of perceptual organization. Because we have no evidence that trans-modal Gestalts exist outside of speech, we are sceptical about multimodal multistability as a general phenomenon.

## ENDNOTES

<sup>1</sup>Participants made more than 95 per cent of their responses to the first note of a run. We treated other responses—whether to a gap or to a note that was not the first in a run—as errors and excluded them from further analysis.

<sup>2</sup>A generalized linear mixed model.

## REFERENCES

- McGurk, H. & MacDonald, J. 1976 Hearing lips and seeing voices. *Nature* **264**, 746–748. (doi:10.1038/264746a0)
- Jack, C. E. & Thurlow, W. R. 1973 Effects of degree of visual association and angle of displacement on the 'ventriloquism' effect. *Percept. Mot. Skills* **37**, 967–979. (doi:10.2466/pms.1973.37.3.967)
- Thurlow, W. R. & Jack, C. E. 1973 Certain determinants of the 'ventriloquism effect'. *Percept. Mot. Skills* **36**, 1171–1184. (doi:10.2466/pms.1973.36.3c.1171)
- van Wassenhove, V., Grant, K. W. & Poeppel, D. 2007 Temporal window of integration in auditory-visual speech perceptions. *Neuropsychologia* **45**, 598–607. (doi:10.1016/j.neuropsychologia.2006.01.001)
- Bertelson, P., Vroomen, J. & de Gelder, B. 2003 Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychol. Sci.* **14**, 592–597. (doi:10.1046/j.0956-7976.2003.psci.1470.x)
- Bushara, K. O., Grafman, J. & Hallett, M. 2001 Neural correlates of auditory-visual stimulus onset asynchrony detection. *J. Neurosci.* **21**, 300–304.
- Colin, C., Radeau, M., Deltenre, P. & Morais, J. 2001 Rules of intersensory integration in spatial scene analysis and speech reading. *Psychol. Belg.* **41**, 131–144.
- Colin, C., Radeau, M. & Deltenre, P. 2005 Top-down and bottom-up modulation of audiovisual integration in speech. *Eur. J. Cogn. Psychol.* **17**, 541–560. (doi:10.1080/09541440440000168)
- Jones, J. A. & Callan, D. E. 2003 Brain activity during audiovisual speech perception: an fMRI study of the McGurk effect. *Cogn. Neurosci. Neuropsychol.* **14**, 1129–1133.
- Kanaya, S. & Yokosawa, K. 2011 Perceptual congruency of audio-visual speech affects ventriloquism with bilateral visual stimuli. *Psychon. Bull. Rev.* **18**, 123–128. (doi:10.3758/s13423-010-0027-z)
- Morsella, E. 2005 The function of phenomenal states: supramodular interaction theory. *Psychol. Rev.* **112**, 1000–1021. (doi:10.1037/0033-295X.112.4.1000)
- Köhler, W. 1929 *Gestalt psychology*. New York, NY: Liveridge.
- Köhler, W. 1947 *Gestalt psychology: an introduction to new concepts in modern psychology* (revised edition). New York, NY: Liveright.
- Ramachandran, V. & Hubbard, E. M. 2001 Synaesthesia: a window into perception, thought and language. *J. Conscious. Stud.* **8**, 3–34.
- Maurer, D., Pathman, T. & Mondloch, C. J. 2006 The shape of boubas: sound-shape correspondences in toddlers and adults. *Dev. Sci.* **9**, 316–322. (doi:10.1111/j.1467-7687.2006.00495.x)
- Parise, C. V. & Spence, C. 2009 'When birds of a feather flock together': synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE* **4**, e5664. (doi:10.1371/journal.pone.0005664)
- Gepshtein, S., Burge, J., Ernst, M. O. & Banks, M. S. 2005 The combination of vision and touch depends on spatial proximity. *J. Vis.* **5**, 1013–1023. (doi:10.1167/5.11.7)
- Schutz, M. & Kubovy, M. 2009 Causality and cross-modal integration. *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 1791–1810. (doi:10.1037/a0016455)
- Schutz, M. & Kubovy, M. 2009 Deconstructing a musical illusion: point-light representations capture salient properties of impact motions. *Can. Acoust.* **37**, 23–28.
- Armontrout, J. A., Schutz, M. & Kubovy, M. 2009 Visual determinants of a cross-modal illusion. *Attent. Percept. Psychophys.* **71**, 1618–1627. (doi:10.3758/APP.71.7.1618)
- Adams, P. A. & Haire, M. 1959 The effect of orientation on the reversal of one cube inscribed in another. *Am. J. Psychol.* **72**, 296–299. (doi:10.2307/1419384)
- Bradley, D. R. & Petry, H. M. 1977 Organizational determinants of subjective contour: the subjective Necker cube. *Am. J. Psychol.* **90**, 253–262. (doi:10.2307/1422047)
- Schumann, F. 1987 Contributions to the analysis of visual perception: first paper: some observations on the combination of visual impressions into units. In *The perception of illusory contours* (eds S. Petry & G. E. Meyer), pp. 21–34. New York, NY: Springer.
- Kanizsa, G. 1955 Margini quasi-percettivi in campi con stimolazione omogenea. *Riv. Psicol.* **49**, 7–30.
- Kanizsa, G. 1979 *Organization in vision: essays on Gestalt perception*. New York, NY: Praeger.
- Warren, R. M. & Gregory, R. L. 1958 An auditory analogue of the visual reversible figure. *Am. J. Psychol.* **71**, 612–613. (doi:10.2307/1420267)
- Sato, M., Basirat, A. & Schwartz, J. 2007 Visual contribution to the multistable perception of speech. *Percept. Psychophys.* **69**, 1360. (doi:10.3758/BF03192952)
- Strother, L. & Kubovy, M. 2003 Perceived complexity and the grouping effect in band patterns. *Acta Psychol.* **114**, 229–244. (doi:10.1016/j.actpsy.2003.06.001)
- Kubovy, M. 1981 Concurrent pitch segregation and the theory of indispensable attributes. In *Perceptual*

- organization (eds M. Kubovy & J. R. Pomerantz), pp. 55–98. Hillsdale, NJ: Lawrence Erlbaum.
- 30 Kubovy, M. 1988 Should we resist the seductiveness of the space:time::vision:audition analogy? *J. Exp. Psychol. Hum. Percept. Perform.* **14**, 318–320. (doi:10.1037/0096-1523.14.2.318)
  - 31 Kubovy, M. & Schutz, M. 2010 Audio-visual objects. *Rev. Philos. Psychol.* **1**, 41–61. (doi:10.1007/s13164-009-0004-5)
  - 32 Kubovy, M. & Van Valkenburg, D. 2001 Auditory and visual objects. *Cognition* **80**, 97–126. (doi:10.1016/S0010-0277(00)00155-4)
  - 33 Cook, L. A. & Van Valkenburg, D. L. 2009 Audio-visual organisation and the temporal ventriloquism effect between grouped sequences: evidence that unimodal grouping precedes cross-modal integration. *Perception* **38**, 1220–1233. (doi:10.1068/p6344)
  - 34 Pressnitzer, D. & Hupé, J.-M. 2006 Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Curr. Biol.* **16**, 1351–1357. (doi:10.1016/j.cub.2006.05.054)
  - 35 Bregman, A. S. & Campbell, J. 1971 Primary auditory stream segregation and perception of order in rapid sequences of tones. *J. Exp. Psychol.* **89**, 244–249. (doi:10.1037/h0031163)
  - 36 van Noorden, L. P. A. S. 1975 Temporal coherence in the perception of tone sequences. PhD thesis, Technical University of Eindhoven, Eindhoven, The Netherlands.
  - 37 Wallach, H. 1935 Über visuell wahrgenommene Bewegungsrichtung. *Psychol. Forschung* **20**, 325–380. (doi:10.1007/BF02409790)
  - 38 Meng, M. & Tong, F. 2004 Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures. *J. Vis.* **4**(7), 2. (doi:10.1167/4.7.2)
  - 39 Binda, P., Lunghi, C. & Morrone, C. 2010 Touch disambiguates rivalrous perception at early stages of visual analysis. *J. Vis.* **10**(7), 854. (doi:10.1167/10.7.854)
  - 40 Lunghi, C., Binda, P. & Morrone, M. C. 2010 Touch disambiguates rivalrous perception at early stages of visual analysis. *Curr. Biol.* **20**, R143–R144. (doi:10.1016/j.cub.2009.12.015)
  - 41 Maruya, K., Yang, E. & Blake, R. 2007 Voluntary action influences visual competition. *Psychol. Sci.* **18**, 1090–1098. (doi:10.1111/j.1467-9280.2007.02030.x)
  - 42 Conrad, V., Bartels, A., Kleiner, M. & Noppeney, U. 2010 Audiovisual interactions in binocular rivalry. *J. Vis.* **10**(10), 27. (doi:10.1167/10.10.27)
  - 43 Kang, M.-S. & Blake, R. 2005 Perceptual synergy between seeing and hearing revealed during binocular rivalry. *Psychologija* **32**, 7–15.
  - 44 Wertheimer, M. 1923 Untersuchungen zur Lehre von der Gestalt, II. *Psychol. Forschung* **4**, 301–350, 1923. [Transl. extracted from Ellis, pp. p71–88.]
  - 45 Ash, M. G. 1998 *Gestalt psychology in German culture, 1890–1967: Holism and the quest for objectivity*. Cambridge, UK: Cambridge University Press.
  - 46 Schumann, F. 1987 Beiträge zur psychologie der gesichtswahrnehmung. *Z Psychol.* **23**, 1–32, 1900. [Transl. by Schumann.]
  - 47 Müller, G. E. 1903 Die Gesichtspunkte und die Tatsachen der psychophysischen Methodik. In *Ergebnisse der physiologie*, vol. II, part 2 (eds L. Asher & K. Spiro), pp. 267–516. Wiesbaden, Germany: J.F. Bergmann.
  - 48 Ellis, W. D. (ed.) 1938 *A source book of Gestalt psychology*. New York, NY: Harcourt, Brace and Company.
  - 49 Kubovy, M. 1994 The perceptual organization of dot lattices. *Psychon. Bull. Rev.* **1**, 182–190. (doi:10.3758/BF03200772)
  - 50 Bravais, A. 1949 *Crystallographic studies: on the systems formed by points regularly distributed on a plane or in space*. Crystallographic Society of America, N.P. Original work published 1866. [Transl. from *Études cristallographiques: mémoire sur les systèmes formés par des points distribués régulièrement sur un plan ou dans l'espace*, Paris, France: Gauthier-Villars.]
  - 51 Oyama, T. 1961 Perceptual grouping as a function of proximity. *Percept. Mot. Skills* **13**, 305–306.
  - 52 Kubovy, M. & Wagemans, J. 1995 Grouping by proximity and multistability in dot lattices: a quantitative gestalt theory. *Psychol. Sci.* **6**, 225–234. (doi:10.1111/j.1467-9280.1995.tb00597.x)
  - 53 Kubovy, M., Holcombe, A. O. & Wagemans, J. 1998 On the lawfulness of grouping by proximity. *Cogn. Psychol.* **35**, 71–98. (doi:10.1006/cogp.1997.0673)
  - 54 Kubovy, M. & van den Berg, M. 2008 The whole is equal to the sum of its parts: a probabilistic model of grouping by proximity and similarity in regular patterns. *Psychol. Rev.* **115**, 131–154. (doi:10.1037/0033-295X.115.1.131)
  - 55 Deutsch, D. 1980 The processing of structured and unstructured tonal sequences. *Percept. Psychophys.* **28**, 381–389. (doi:10.3758/BF03204881)
  - 56 Longuet-Higgins, H. C. & Lee, C. S. 1982 The perception of musical rhythms. *Perception* **11**, 115–128. (doi:10.1068/p110115)
  - 57 Martin, J. G. 1972 Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychol. Rev.* **79**, 487–509. (doi:10.1037/h0033467)
  - 58 Royer, F. L. & Garner, W. R. 1966 Response uncertainty and perceptual difficulty of auditory temporal patterns. *Percept. Psychophys.* **1**, 41–47. (doi:10.3758/BF03207820)
  - 59 Garner, W. R. & Gottwald, R. L. 1968 The perception and learning of temporal patterns. *Q. J. Exp. Psychol.* **20**(Pt. II), 97–109. (doi:10.1080/14640746808400137)
  - 60 Royer, F. L. & Garner, W. R. 1970 Perceptual organization of nine-element auditory temporal patterns. *Percept. Psychophys.* **7**, 115–120. (doi:10.3758/BF03210146)
  - 61 Preusser, D., Garner, W. R. & Gottwald, R. L. 1970 Perceptual organization of two-element temporal patterns as a function of their component one-element patterns. *Am. J. Psychol.* **83**, 151–170. (doi:10.2307/1421321)
  - 62 Boker, S. M. & Kubovy, M. 1998 The perception of segmentation in sequences: local information provides the building blocks for global structure. In *Timing of behavior: neural, computational, and psychological perspectives*, (eds D. A. Rosenbaum & C. E. Collyer), pp. 109–123. Cambridge, MA: MIT Press.
  - 63 Gibson, J. J. 1966 *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.