Crossmodal correspondences: A tutorial review

Charles Spence

Abstract In many everyday situations, our senses are bombarded by many different unisensory signals at any given time. To gain the most veridical, and least variable, estimate of environmental stimuli/properties, we need to combine the individual noisy unisensory perceptual estimates that refer to the same object, while keeping those estimates belonging to different objects or events separate. How, though, does the brain "know" which stimuli to combine? Traditionally, researchers interested in the crossmodal binding problem have focused on the roles that spatial and temporal factors play in modulating multisensory integration. However, crossmodal correspondences between various unisensory features (such as between auditory pitch and visual size) may provide yet another important means of constraining the crossmodal binding problem. A large body of research now shows that people exhibit consistent crossmodal correspondences between many stimulus features in different sensory modalities. For example, people consistently match high-pitched sounds with small, bright objects that are located high up in space. The literature reviewed here supports the view that crossmodal correspondences need to be considered alongside semantic and spatiotemporal congruency, among the key constraints that help our brains solve the crossmodal binding problem.

Keywords Multisensory integration · Crossmodal correspondence · Synaesthesia · Bayesian integration theory · Crossmodal binding problem

C. Spence (✉)
Crossmodal Research Laboratory,
Department of Experimental Psychology, University of Oxford,
South Parks Road,
Oxford OX1 3UD, UK
e-mail: charles.spence@psy.ox.ac.uk

"What is essential in the sensuous-perceptible is not that which separates the senses from one another, but that which unites them; unites them among themselves; unites them with the entire (even with the non-sensuous) experience in ourselves; and with all the external world that there is to be experienced." (Von Hornbostel, The Unity of the Senses, 1927/1950, p. 214)

For many years now, the majority of cognitive neuroscience research on the topic of multisensory perception has tended to focus on trying to understand, and increasingly to model (Alais & Burr, 2004; Ernst & Bülthoff, 2004; Roach, Heron, & McGraw, 2006), the spatial and temporal factors modulating multisensory integration (e.g., see Calvert, Spence, & Stein, 2004; Spence & Driver, 2004). Broadly speaking, it appears that multisensory integration is more likely to occur the closer that the stimuli in different modalities are presented in time (e.g., Jones & Jarick, 2006; Shore, Barnes, & Spence, 2006; van Wassenhove, Grant, & Poeppel, 2007). Spatial coincidence has also been shown to facilitate multisensory integration under some (Frens, Van Opstal, & Van der Willigen, 1995; Slutsky & Recanzone, 2001), but by no means all, conditions (see, e.g., Bertelson, Vroomen, Wiegeraad, & de Gelder, 1994; Innes-Brown & Crewther, 2009; Jones & Jarick, 2006; Jones & Munhall, 1997; Recanzone, 2003; Vroomen & Keetels, 2006).

What other factors influence multisensory integration in humans? There has been a recent resurgence of research interest in the effects of both semantic (Y. C. Chen & Spence, 2010; Doehrmann & Naumer, 2008; Grassi & Casco, 2010; Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004; Naumer & Kaiser, 2010) and synaesthetic (Evans & Treisman, 2010; Gallace & Spence, 2006; Makovac & Gerbino, 2010; Parise & Spence, 2008a, 2009) congruency on multisensory information processing.

Semantic congruency usually refers to those situations in which pairs of auditory and visual stimuli are presented that vary (i.e., match vs. mismatch) in terms of their *identity* and/or *meaning*. In laboratory studies of multisensory perception, semantic congruency effects are typically assessed by measuring the consequences of presenting matching or mismatching object pictures and environmental sounds (such as a woofing sound paired with a static picture of a dog or cat; Hein et al., 2007; Molholm, Ritter, Javitt, & Foxe, 2004) or of visually presenting letters with matching or mismatching speech sounds (e.g., van Atteveldt, Formisano, Goebel, & Blomert, 2004). A number of researchers have also studied semantic congruency effects by investigating the multisensory integration of gender-matched versus gender-mismatched audiovisual speech stimuli (e.g., Easton & Basala, 1982; Green, Kuhl, Meltzoff, & Stevens, 1991; Vatakis & Spence, 2007; Walker, Bruce, & O'Malley, 1995). By contrast, *synaesthetic congruency* refers to correspondences between more basic stimulus *features* (e.g., pitch, lightness, brightness, size) in different modalities. The term *synaesthetic congruency* usually refers to correspondences between putatively nonredundant stimulus attributes or dimensions that happen to be shared by many people. It has recently become popular for researchers to argue that stimuli that are either semantically or synaesthetically congruent will more likely be bound together, a notion that is sometimes referred to as the "unity effect" (e.g., Spence, 2007; Vatakis, Ghazanfar, & Spence, 2008; see also Welch & Warren, 1980).

In this tutorial article, I start by reviewing the historical evidence for the existence of crossmodal correspondences that emerged from early studies of sound symbolism and crossmodal matching. In the sections that follow, I then go on to review the evidence concerning the effects of crossmodal correspondences for participants' performance in both speeded classification tasks and unspeeded psychophysical tasks. Taken together, the evidence reviewed in these sections is consistent with the view that there may be several qualitatively different kinds of crossmodal correspondence—statistical, structural, and semantically mediated—and that they may have different developmental trajectories as well as different consequences for human perception and behaviour. Next, I evaluate the extent to which crossmodal correspondences can be modelled in terms of the coupling priors that are increasingly being incorporated into contemporary Bayesian decision theory. Finally, I consider the evidence concerning the likely neural substrates underlying crossmodal correspondences. Here, I focus both on the question of where such information may be represented in the human brain and on how the acquisition of novel crossmodal associations between stimuli presented in different modalities impacts on neural activity.

Crossmodal correspondence: A note regarding terminology

In this review, I will evaluate the evidence regarding the existence, and the consequences for human information processing, of synaesthetic congruency/crossmodal correspondences. Other terms that have been used to refer to similar phenomena over the years include *synaesthetic correspondences* (Braaten, 1993; Martino & Marks, 2000; Melara & O'Brien, 1987; Parise & Spence, 2009; P. Walker et al., 2010), *synaesthetic associations* (Parise & Spence, 2008a), *crossmodal equivalences* (Lewkowicz & Turkewitz, 1980), *crossmodal similarities* (Marks, 1987a, 1987b, 1989a, 1989b), and *natural crossmodal mappings* (Evans & Treisman, 2010). Such terms have been used to describe the many nonarbitrary associations that appear to exist between different basic physical stimulus attributes, or features, in different sensory modalities. These crossmodal associations or correspondences may be used by humans (and presumably other species as well; see Bee, Perrill, & Owen, 2000; Davies & Halliday, 1978; Fitch & Reby, 2001; Harrington, 1987; Morton, 1994) along with spatiotemporal correspondence and semantic congruency to help solve the *crossmodal binding problem* (i.e., knowing which of the many stimuli that happen to be presented in different modalities at any one time should be bound together; see Ernst, 2007; Spence, Ngo, Lee, & Tan, 2010).

Generally speaking, the terms *synaesthetic correspondence* and *synaesthetic association* have been used to describe only those correspondences between nonredundant sensory dimensions (such as between pitch in audition and brightness in vision). By contrast, the other terms, such as *crossmodal correspondence* or *crossmodal similarity*, have a broader coverage, including both synaesthetic correspondences and correspondences between redundantly coded stimulus features (i.e., those features that can potentially be perceived through different sensory modalities), such as the size or shape of an object, or the auditory and visual duration of an event. However, it should be noted that this distinction isn't always as easy to maintain as it might at first seem. Indeed, certain crossmodal correspondences that might initially appear to be nonredundant, such as between pitch and size, may, on closer inspection, actually turn out to reflect the redundant coding of object size (this issue will be discussed in more detail later).

While, in a literal sense, the use of the term *synaesthetic correspondences* in this context is clearly appropriate, meaning as it does, the "joining of the senses" (Melara & O'Brien, 1987; Wicker, 1968), in another sense it is inappropriate (or, at the very least, potentially misleading). For while some researchers have argued that such crossmodal correspondences should be conceptualized as a weak

form of synaesthesia (Martino & Marks, 2001), I believe (as we will see below) that synaesthesia may not necessarily be the most appropriate model for thinking about all such crossmodal phenomena (especially given that the experience of a concurrent stimulus, a core feature of full-blown synaesthesia, has never, at least as far as I am aware, been reported in the case of any crossmodal correspondences; see also Elias, Saucier, Hardie, & Sarty, 2003). Instead, more general terms such as *crossmodal correspondences* (Gilbert, Martin, & Kemp, 1996; Mondloch & Maurer, 2004) or *crossmodal associations* (Martino & Marks, 2001) may be more appropriate (and will be used here from now on), given that they are less pejorative with regard to the putative neural substrates underlying these effects. To be absolutely clear, then, the term *crossmodal correspondence* is used in this review to refer to a compatibility effect between attributes or dimensions of a stimulus (i.e., an object or event) in different sensory modalities (be they redundant or not). Such correspondences occur between polarized stimulus dimensions, such that a more-or-less extreme stimulus on a given dimension should be compatible with a more-or-less extreme value on the corresponding dimension. A key feature of (or assumption underlying) all such crossmodal correspondences is that they are shared by a large number of people (and some may, in fact, be universal).

At the outset, it is important to note that there are a number of different ways in which stimuli, objects, and/or events in different sensory modalities can be matched (or associated; see also Marks, 1978, pp. 4–7). At the most basic level, they may be related in terms of some common (*amodal*)[1] stimulus feature shared by a number, if not necessarily all, of the modalities (Marks, Szczesiul, & Ohlott, 1986). To date, there is already some limited evidence that when the different senses provide redundant information about the same amodal stimulus feature (such

as its temporal pattern), the likelihood of multisensory integration is increased (see Frings & Spence, 2010; Radeau & Bertelson, 1987; Thomas, 1941; though see also Spence, 2007). At the next level up, they may occur between different, seemingly unrelated (and in some cases *modal*) features present in two or more sensory modalities, as when people match high-pitched sounds with small and/or bright objects.

Crossmodal correspondences between stimuli may also be established at a more abstract level, such as in terms of their pleasantness, cognitive meaning, or activity (see Bozzi & Flores D'Arcais, 1967; Crisinel & Spence, 2010a; Hartshorne, 1934; Janković, 2010; Lyman, 1979; Osgood, Suci, & Tannenbaum, 1957). In a related vein, it has also been suggested that crossmodal correspondences can be established at the level of the effect that the stimuli have on the observer: For example, stimuli may be matched (or associated) if they both happen to increase an observer's level of alertness or arousal, or if they both happen to have the same effect on an observer's emotional state, mood, or affective state (see, e.g., Boernstein, 1936, 1970; Collier, 1996; Cowles, 1935; Poffenberger & Barrows, 1924; Simpson, Quinn, & Ausubel, 1956; see also Lewkowicz & Turkewitz, 1980). There is now reasonable evidence to support the claim that crossmodal correspondences may occur at all of these levels (i.e., from correspondences between low-level amodal stimulus properties such as duration, through to high-level cognitive correspondences based on stimulus meaning/valence).

Crossmodal correspondences: Early research on crossmodal matching

Psychologists have known about the existence of crossmodal correspondences for many years (see, e.g., Fox, 1935; Jespersen, 1922; Köhler, 1929; Newman, 1933; Sapir, 1929; Uznadze, 1924; Wertheimer, 1958, for early research). For example, more than 80 years ago, Edward Sapir highlighted the existence of a crossmodal association between the speech sounds /a/ and /i/ and object size. He observed that most people associate the nonsense words "mal" and "mil" with large and small objects, respectively (see Fig. 1a). In the same year, Köhler (1929, pp. 224–225) reported that when shown the two shapes illustrated in Fig. 1b, most people matched the globular rounded shape on the left with the nonsense word "Baluma" and the straight-edged angular shape on the right with the nonsense word "Takete," rather than vice versa. Results such as these have led to a fairly constant stream of research over the intervening years on the topic of *sound* (or *phonetic*) *symbolism*, an area that is still just as popular today (e.g., Imai, Kita, Nagumo, & Okada, 2008; Parise & Pavani,
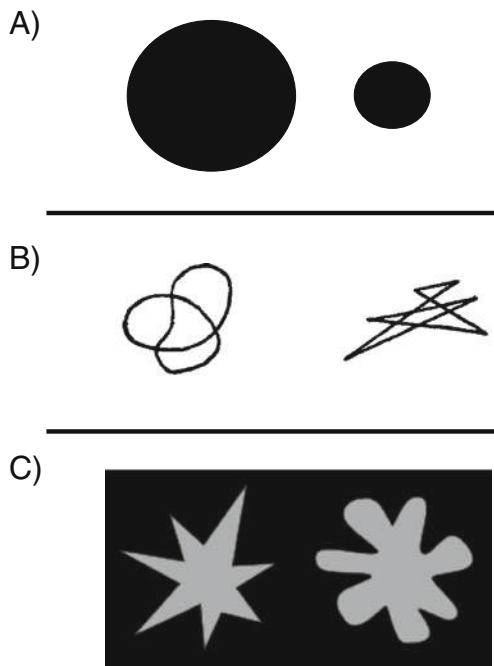
---

[1] *Amodal* features can be used to identify an aspect/attribute of an event or object in more than one (though, note, not necessarily in every) sensory modality. The duration and temporal patterning (rhythm) of events both constitute amodal stimulus features (see Frings & Spence, 2010; Guttman, Gilroy, & Blake, 2005; Marks, 1987b; Radeau & Bertelson, 1987; Thomas, 1941), as do the shape and size of objects (which can be discerned by vision, by touch/haptics, and to a certain extent by audition; see Spence & Zampini, 2006). Intensity is another amodal feature that can be used to describe objects (e.g., Lewkowicz & Turkewitz, 1980; Marks, 1987a; Marks et al., 1986; R. Walker, 1987). Other stimulus attributes that can perhaps also be treated as amodal features include movement, location, texture, and gender. By contrast, *modal* features/attributes identify an aspect of a stimulus that is specific (or peculiar) to a single sensory modality: The pitch of a sound, the colour of a light, and the sweetness and bouquet of a wine are all modal stimulus properties (Lewkowicz & Turkewitz, 1980; Spence, Levitan, Shankar, & Zampini, 2010). Crossmodal correspondences have now been demonstrated between many different combinations of modal and amodal stimulus features.

A)

B)

C)

Fig. 1 (**a, b**) Schematic figure illustrating the kinds of stimuli used by (**a**) Sapir (1929) and (**b**) Köhler (1929) to demonstrate the existence of reliable crossmodal associations between different auditory and visual dimensions. Sapir showed that people associate the nonsense word "mal" with large objects and the word "mil" with small objects. Meanwhile, Köhler (1929) demonstrated that people associate the nonsense word "Baluma" with the shape on the left and the word "Takete" with the shape on the right. In the 1947 version of Köhler's book, he finally settled on the words "maluma" and "takete", given concerns that "Baluma" sounded a bit too much like "balloon." (**c**) Recent interest in sound (or phonetic) symbolism has come, at least in part, from Ramachandran and Hubbard's (2001, 2003) finding that between 95% and 98% of the population agree on which of the shapes in (**c**) is the "bouba" (right) and which the "kiki" (left). [Panel B is redrawn from *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology* (p. 254–255). by W. Köhler, 1947, New York: Liveright, Copyright 1947 by Liveright Publications. Redrawn with permission.]

2011; Westbury, 2005; see Belli & Sagrillo, 2001; Hinton, Nichols, & Ohala, 1994; Nuckolls, 2003, for reviews). At least a part of the recent wave of popularity in this area can be attributed to a couple of publications by Ramachandran and Hubbard (2001, 2003), in which the authors replicated Köhler's basic results using slightly different words and shapes (see Fig. 1c) and christened it the *bouba/kiki effect*. It should be noted, though, that the earliest discussion of the nonarbitrary association between a word's sound and its meaning appears in Plato's Cratylus dialogue (see Plato, 1961).

The majority of studies in the area of sound symbolism have tended to restrict themselves to detailing the existence of particular crossmodal associations (e.g., Boyle & Tarte, 1980; Holland & Wertheimer, 1964; Lindauer, 1990; Taylor, 1963), checking for the universality of such associations across cultures/languages (e.g., Davis, 1961;

Gebels, 1969; Osgood, 1960; Rogers & Ross, 1975; Taylor & Taylor, 1962; see Hinton et al., 1994, for a review), and/or charting their emergence over the course of human development (e.g., Irwin & Newland, 1940; Maurer, Pathman, & Mondloch, 2006). Interestingly, the latest research suggests that neuropsychological factors may also impact on the expression of crossmodal correspondences: It has, for instance, been reported that children with autism spectrum disorder do not show the bouba/kiki effect (Oberman & Ramachandran, 2008; Ramachandran & Oberman, 2006). Meanwhile, Ramachandran and Hubbard (2003, p. 48) mention, albeit in passing, that damage to the angular gyrus (located within the temporal–parietal–occipital [TPO] region) results in the loss of the bouba/kiki effect, such that individuals "cannot match the shape with the correct sound." If confirmed, such results might be taken to suggest that crossmodal correspondences, at least those involving sound symbolism, can occur at quite a high level (see also Evans & Treisman, 2010; Nahm, Tranel, Damasio, & Damasio, 1993; Westbury, 2005).

More or less independently of this research on sound symbolism, psychophysicists started to investigate the ability of people to match the stimuli presented in different sensory modalities and the extent to which such effects were reliable across groups of participants. For example, S. S. Stevens and his colleagues at Harvard demonstrated that both adults and children (5-year-olds) reliably matched brightness with loudness crossmodally (e.g., Bond & Stevens, 1969; J. C. Stevens & Marks, 1965; see also Root & Ross, 1965). That is, both groups of participants paired light grey colour patches with louder sounds and darker grey colour patches with quieter sounds. People also match high-pitched tones with brighter surfaces (Marks, 1974; Wicker, 1968) and louder sounds with visual stimuli that have a higher contrast (Wicker, 1968). One of the longstanding suggestions here has been that such crossmodal matches may be based on the perceived brightness (or intensity) of the stimuli (see Külpe, 1893; von Hornbostel, 1931; though see also Cohen, 1934; Krantz, 1972).

Simpson et al. (1956) reported a systematic relation between hue and pitch in children, with high-pitched tones more likely to be matched with yellow (rather than with blue). However, in hindsight, it is unclear whether this result might not actually reflect a crossmodal matching of lightness and pitch, since the stimuli were not matched for perceived lightness or physical luminance in this early study, and yellow stimuli are often lighter than blue stimuli. Hence, before any weight is put on the crossmodal matching of hue and pitch, a more carefully controlled replication would be needed in which the lightness of the stimuli was carefully matched. Several researchers have also demonstrated that higher-pitched sounds tend to be associated with higher elevations in space (Mudd, 1963;

Pratt, 1930; Roffler & Butler, 1968). Moving beyond the examples of simple unidimensional sensory stimuli, it turns out that people can also reliably match more complex stimuli, such as music with pictures (e.g., Cowles, 1935; Karwoski, Odbert, & Osgood, 1942).

Developmental researchers have shown that children can match loud sounds with large shapes by 2 years of age (L. B. Smith & Sera, 1992), while the ability to match other dimensions crossmodally appears to develop somewhat more slowly (see Marks, 1984; Marks, Hammeal, & Bornstein, 1987). Meanwhile, research using more indirect measures (such as cardiac habituation/dishabituation, looking preferences, etc.) has provided suggestive evidence that infants may be aware of (or at least their behaviour can be shown to be sensitive to) certain crossmodal correspondences, such as that between auditory pitch and visual elevation/sharpness, within 3–4 months of birth (P. Walker et al., 2010). Some form of crossmodal correspondence between loudness and brightness was demonstrated by Lewkowicz and Turkewitz (1980) in 20-30 day-old infants (see also Braaten, 1993; Maurer et al., 2006; Mondloch & Maurer, 2004; Wagner, Winner, Cicchetti, & Gardner, 1981).

While the focus of this review is primarily on the nature of the crossmodal correspondences that exist between auditory and visual stimuli, it is important to note that crossmodal associations have now been documented between many different pairs of sensory modalities, such as vision and touch (e.g., Martino & Marks, 2000; G. A. Morgan, Goodson, & Jones, 1975; Simner & Ludwig, 2009; P. Walker, Francis, & Walker, in press), audition and touch (e.g., P. Walker & Smith, 1985; Yau, Olenczak, Dammann, & Bensmaia, 2009; see also von Békésy, 1959), and tastes/flavours and sounds (see Bronner, 2011; Bronner, Bruhn, Hirt, & Piper, 2008; Crisinel & Spence, 2009, 2010a, 2010b; Holt-Hansen, 1968, 1976; Mesz, Trevisan, & Sigman, 2011; Rudmin & Cappelli, 1983; Simner, Cuskley, & Kirby, 2010). Researchers have also highlighted crossmodal associations between colours and odours (Gilbert et al., 1996; Kemp & Gilbert, 1997; Spence, 2010), tastes (O'Mahony, 1983), and flavours (see Spence et al., 2010, for a review). Elsewhere, crossmodal associations have been documented between auditory pitch and smell (Belkin, Martin, Kemp, & Gilbert, 1997; Piesse, 1891; von Hornbostel, 1931), smells and shapes (Seo et al., 2010), and even shapes and tastes/flavours (Gal, Wheeler, & Shiv, 2011; Spence & Gallace, in press; see also Gallace, Boschin, & Spence, in press). It therefore appears likely that crossmodal correspondences exist between all possible pairings of sensory modalities. However, given that the majority of research to date has focused on the existence, and consequences, of the correspondences between auditory and visual stimuli, it is on those that we will focus here.

Interim summary

While researchers have generally adopted different experimental approaches, the results of a large body of research on both sound symbolism and crossmodal matching have converged on the conclusion that many nonarbitrary crossmodal correspondences exist between a variety of auditory and visual stimulus features/dimensions. These crossmodal correspondences have been documented both between simple stimulus dimensions, such as loudness and brightness, and between more complex stimuli, such as shapes/images and nonsense words/short musical clips. Having demonstrated the ubiquitous nature of such crossmodal correspondences, the next question to be addressed by researchers was whether or not these correspondences would impact on the efficacy of human information processing. In particular, from the early 1970s onward, psychologists started to investigate whether adults would find it easier to process certain combinations of auditory and visual stimuli if the dimensions on which the stimuli varied happened to share some sort of crossmodal correspondence. Therefore, the next section reviews those studies that have assessed the consequences of crossmodal correspondences between auditory and visual stimuli on information processing in neurologically normal (i.e., nonsynaesthetic) human adults. Note that these studies were all primarily concerned with assessing the efficiency of selective attention, given that the participants always had to discriminate the stimuli presented in one sensory modality while trying to ignore those presented in another modality.

Assessing the impact of crossmodal correspondences on human information processing using the speeded classification task

Bernstein and Edelstein (1971) conducted one of the first studies to demonstrate that people respond more slowly to visual stimuli when their elevation happens to be inconsistent with the relative pitch of a task-irrelevant sound. (Note here the similarity between the speeded classification task and the crossmodal Stroop effect; see, e.g., Cowan & Barron, 1987; MacLeod, 1991; Stroop, 1935.) Visual targets were presented diagonally on either side of fixation (either upper left and lower right, for some participants, or else lower left and upper right for the rest). The participants in Bernstein and Edelstein's study had to discriminate the location of the visual targets as rapidly as possible while a task-irrelevant auditory stimulus was presented either simultaneously with the visual stimulus or slightly (up to 45 ms) later. Crossmodal congruency effects were observed in those blocks of trials

in which the pitch of the sound (either 100 or 1000 Hz) varied randomly on a trial-by-trial basis, but not when the pitch of the sound was blocked. It turns out that the crossmodal correspondence between auditory pitch and visual elevation constitutes one of the more robust associations to have been reported to date (see Evans & Treisman, 2010). Indeed, this particular crossmodal correspondence has subsequently been replicated by a number of researchers (see Ben-Artzi & Marks, 1995; Melara & O'Brien, 1987; Patching & Quinlan, 2002; see also Maeda, Kanai, & Shimojo, 2004; Widmann, Kujala, Tervaniemi, Kujala, & Schröger, 2004); even 6-month-old infants appear to be sensitive to it (Braaten, 1993).

In the years following the publication of Bernstein and Edelstein's (1971) seminal study, Lawrence Marks and his colleagues at Yale University conducted many further speeded classification studies investigating crossmodal correspondences (see Marks, 2004, for a review). Marks reported that people find it harder (i.e., they are slower and less accurate) to classify the target stimuli presented in one sensory modality (e.g., vision) when the distractor stimuli presented in a task-irrelevant modality (e.g., audition) happen to vary along a dimension that shares some sort of crossmodal correspondence with the target dimension.[2] So, for example, Marks (see also Marks et al., 1987) demonstrated that people find it harder to classify the size of a visual stimulus (as either large or small) when the task-irrelevant sound presented on each trial is incongruent in pitch (e.g., when a high-pitched tone is presented at the same time as a large target) than when the distractor sound is congruent (e.g., when a low tone is presented with the large target; see Fig. 2).

Crossmodal correspondences have been demonstrated between both pitch and loudness in audition and lightness and brightness in vision (Hubbard, 1996; Marks, 1987a, 1989b; Martino & Marks, 1999; see also Melara, 1989a). Marks (1987a, Experiment 4) also used the speeded classification methodology to illustrate the consequences for human information processing of the pitch/shape correspondence first reported by Köhler (1929). The visual stimuli in Marks' study consisted of an upturned "V" and an upturned "U," while the auditory stimuli consisted of either a relatively low- or high-pitched tone. As expected, the higher-pitched sound facilitated participants' speeded responses to the more angular shape, while the lower tone facilitated their responses to the more rounded U-shape (see Fig. 2).

---

[2] Note that the assumption underlying all of this research is that the presence of a dimensional correspondence (congruence) will show up as an interaction in the speeded classification task (see Marks, Ben-Artzi, & Lakatos, 2003, p. 126).
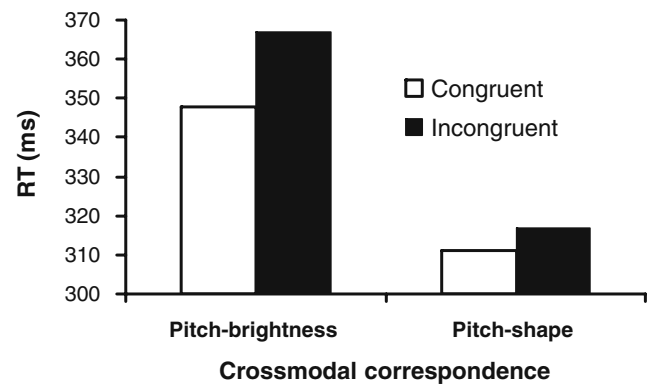

Fig. 2 Graph highlighting the mean RTs to discriminate visual stimuli paired with either crossmodally congruent or incongruent auditory stimuli in Marks's (1987a) study. The visual stimuli varied in either brightness (dim vs. bright) or angularity (rounded vs. sharp), while the auditory stimuli varied in pitch (low vs. high). Responses to crossmodally congruent pairs of stimuli (i.e., a dim visual target paired with a lower-pitched sound or a bright visual stimulus paired with a high-pitched sound, in the left example; a rounded shape with a low-pitched sound or an acute shape paired with a high-pitched tone on the right) were significantly faster than responses to crossmodally incongruent stimulus pairs (i.e., bright visual stimuli with low-pitched sounds or dim visual stimuli paired with high-pitched tones on the left; acute shapes paired with low-pitched tones or rounded shapes paired with high-pitched tones on the right). [From Fig. 4.3 of "Synesthesia" (p. 121–149), by L. E. Marks, in *Varieties of Anomalous Experience: Examining the Scientific Evidence*, edited by E. Cardeña, S. J. Lynn, and S. C. Krippner, 2000, Washington, DC: American Psychological Association. Copyright 2000 by the American Psychological Association. Redrawn with permission.]

Gallace and Spence (2006) conducted a number of experiments to demonstrate that the presentation of a task-irrelevant sound (of either relatively low or high pitch) can significantly influence participants' responses on a speeded visual size discrimination task. They presented two masked grey disks at fixation, one after the other (see Fig. 3). The participants had to respond either as to whether the second variable-sized disk was larger or smaller than the first, standard-sized disk (Experiments 1–3), or else to whether the two disks were the same size or not (Experiment 4). On the majority of trials, a sound was presented in synchrony with the second disk (otherwise, no sound was presented). The relative frequency of the sound (300 or 4500 Hz) was either congruent or incongruent with the size of the second disk (relative to the first). The participants in Gallace and Spence's study responded significantly more rapidly (and somewhat more accurately) on the congruent crossmodal trials (e.g., where a high-frequency sound was presented with a small disk) than on the incongruent trials (e.g., where a low-frequency sound was coupled with a small disk). Interestingly, just as in Bernstein and Edelstein's (1971) original research, when the high-, low-, and no-sound trials were presented
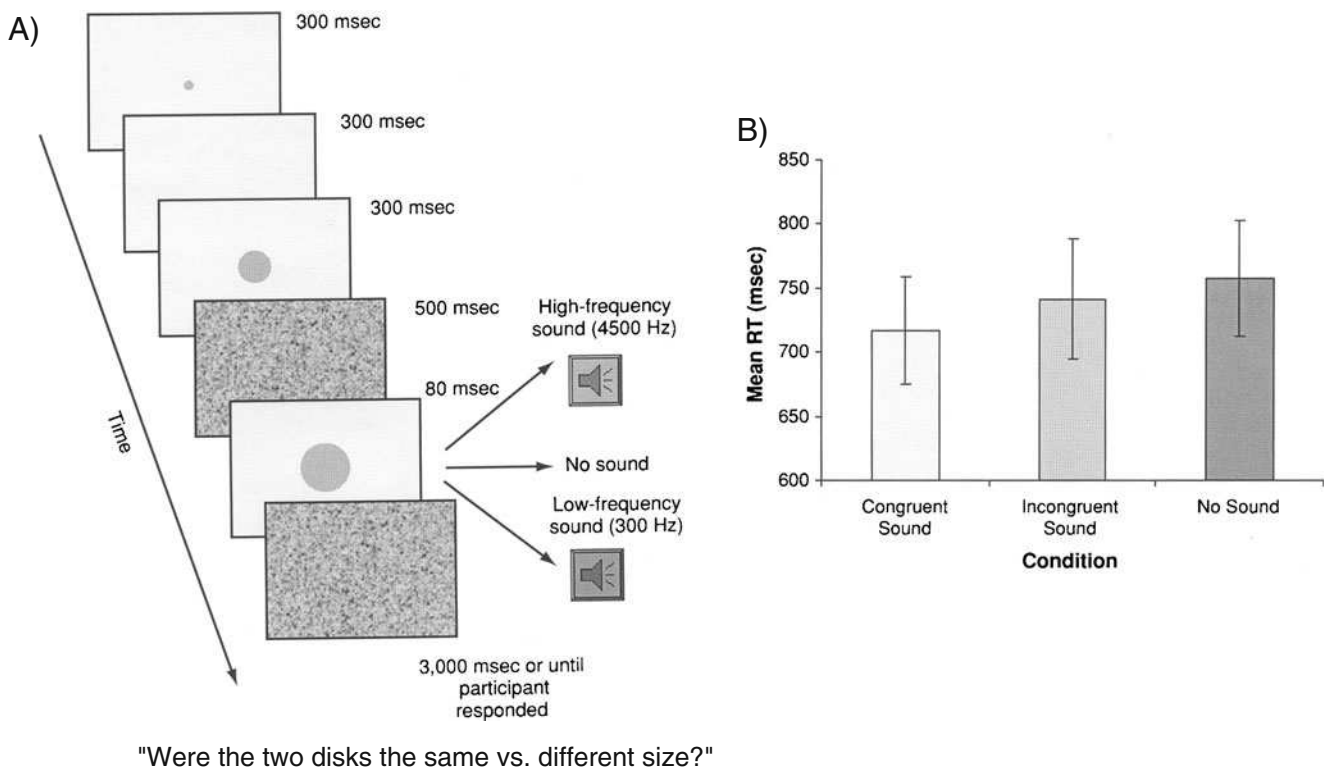
Fig. 3 **a** Schematic time line and stimuli from Gallace and Spence's (2006) experiments highlighting the consequences of the crossmodal correspondence between auditory pitch and visual size for participants' speeded discrimination responses. **b** Results from the speeded "same vs. different" visual size discrimination task. Congruent pairs of stimuli (e.g., a larger disk paired with the lower tone) gave rise to faster RTs. The error bars indicate the standard errors of the means. [From Figs. 1 and 5 of "Multisensory Synesthetic Interactions in the Speeded Classification of Visual Size," by A. Gallace and C. Spence, 2006, *Perception & Psychophysics*, *68*, pp. 1191–1203. Copyright 2006 by the Psychonomic Society. Redrawn with permission.]

in separate blocks, no such crossmodal congruency effect was observed.[3]

Gallace and Spence's (2006) results support previous suggestions that the crossmodal correspondence effects typically observed in the laboratory tend primarily to be relative (and not absolute, as is the case for many types of synaesthesia: Marks, 2000; though see Cytowic, 1993; Cytowic & Wood, 1982; see also E. L. Smith, Grabowecky,

---

[3] It is easy to conclude from the fact that crossmodal congruency effects in the laboratory are typically only observed when the stimuli vary on a trial-by-trial basis that such effects are unlikely to be important in real-world conditions of multisensory integration. However, bear in mind that the laboratory situation in which exactly the same stimulus is presented trial after trial (as when exactly the same tone is presented every few seconds in those blocks of trials where the stimuli are fixed; see Bernstein & Edelstein, 1971; Gallace & Spence, 2006) is completely unlike the situations we typically face in the real world, where multiple stimuli varying in many different attributes are likely to be presented at around the same time. Under the latter "real-world" conditions, when the brain has to decide, using the available information, which stimuli should be integrated and which should be kept separate (e.g., Bedford, 2001; Körding et al., 2007), crossmodal correspondences may play a more important role in crossmodal binding/grouping, since the presence of multiple stimuli will mean that some relative information is likely available.

& Suzuki, 2007). It should be noted, though, that the results of crossmodal matching research suggest that different stimulus dimensions may vary in this regard. For example, Marks et al. (1986) found that while the crossmodal matching of duration was nearly absolute, intensity matching showed some compromise between absolute equivalence and relative (contextual) comparison. One possibility to consider here, then, is that relative effects may be a more prominent attribute of semantic correspondences (or of crossmodal correspondences between modal stimulus dimensions, such as lightness and pitch), while the likelihood of showing some absolute effect presumably increases when one considers the case of crossmodal correspondences between pairs of dimensions that refer to the same underlying amodal stimulus property (such as duration).

Gallace and Spence (2006) reported that simply presenting the spoken words "high" and "low" had much the same effect on participants' performance as presenting high- or low-pitched sounds, thus suggesting that this particular form of crossmodal congruency is largely linguistically/semantically mediated (see also Long, 1977; Martino & Marks, 1999; Melara, 1989b; P. Walker & Smith, 1984). It therefore appears that dimensional crossmodal interactions between auditory and visual stimuli can occur between

pairs of sensory stimulus dimensions, between a sensory stimulus dimension and a polar pair of adjectives, or between two pairs of polar adjectives. Whatever the correct interpretation for Gallace and Spence's results turns out to be (i.e., whether they are semantically vs. perceptually mediated), they nevertheless demonstrate that the relative frequency of an irrelevant sound can influence the speed of people's judgements of the perceived size of a simultaneously presented visual stimulus, thus adding pitch–size to the list of auditory–visual crossmodal correspondences that have been demonstrated to date using the speeded classification task (see Table 1).

More recently, Evans and Treisman (2010) conducted nine speeded classification experiments highlighting the existence of bidirectional crossmodal correspondences between the features of auditory pitch and visual elevation, size, and spatial frequency, but not contrast. The participants in Evans and Treisman's first experiment were presented with a circular black-and-white visual grating positioned above or below fixation and/or with a high- or low-frequency tone (1500 and 1000 Hz, respectively). They either discriminated the pitch of the tone (low vs. high) or else, in other blocks of experimental trials, reported whether the grating had been presented from above or below fixation. The results showed that the participants responded significantly more rapidly (and somewhat more accurately) when the simultaneously presented auditory and visual stimuli were crossmodally congruent than when they were not (i.e., when they were crossmodally incongruent).

Evans and Treisman (2010) reported a similar pattern of results in another experiment utilizing an indirect task in which the participants were no longer explicitly asked about the elevation of the visual target or the pitch of the sound. Instead, they simply had to discriminate the identity of a computer-generated tone that simulated the sound of a piano or violin, or else judge the orientation (leftward vs. rightward) of a grating that just so happened to be randomly presented either above or below fixation. Evans and Treisman also demonstrated crossmodal associations between auditory pitch and visual size (thus replicating Gallace & Spence's, 2006, earlier findings) and between auditory pitch and visual spatial frequency (see Table 1). In the latter two cases, the direct and indirect versions of the speeded discrimination task gave rise to crossmodal effects that were of a similar magnitude, thus suggesting that the effects reflect genuine perceptual priming rather than merely some form of response compatibility effect.

It is, however, important to note that not all pairs of auditory and visual dimensions give rise to significant crossmodal congruency effects in the speeded classification task. So, for example, no crossmodal correspondence has so far been observed between pitch and hue (blue vs. red; Bernstein, Eason, & Schurman, 1971) or between loudness and lightness (Marks, 1987a). Similarly, Evans and Treisman

(2010) failed to demonstrate any crossmodal association between auditory pitch and visual contrast.

Distinguishing between different kinds of crossmodal correspondences

According to researchers, there are several different situations in which crossmodal correspondences may be observed: First, they may occur for pairs of stimulus dimensions that happen to be correlated in nature (such as the natural correlation between the size, or mass, of an object and its resonant frequency—the larger the object, the lower the frequency; see Coward & Stevens, 2004; Grassi, 2005; McMahon & Bonner, 1983). Second, they may occur because of neural connections that are present at birth (Mondloch & Maurer, 2004; see also Marks, 1978, 1987a; Wagner & Dobkins, 2009). As Marks (1978) put it, crossmodal correspondences may fall naturally out of the organization of the perceptual system (see also Marks et al., 1986). Such structural correspondences may also occur between pairs of *prothetic* (i.e., magnitude-related) dimensions such as loudness and size (see L. B. Smith & Sera, 1992),[4] given that magnitude (regardless of the particular dimension under consideration) appears to be represented in the same way by the brain (see Walsh, 2003). Third, crossmodal correspondences may occur when the terms that people use to describe the stimuli in the two dimensions overlap, as for the words "low" and "high," which are used to describe both the elevation of a visual stimulus and the pitch of a sound (see Gallace & Spence, 2006; Martino & Marks, 1999; Melara, 1989b; Mudd, 1963; Stumpf, 1883). As we will see later, these three kinds of crossmodal correspondence—statistical, structural, and semantically mediated—may have different consequences for human information processing.

Pairs of sensory dimensions that do not meet any of these conditions (such as the dimensions of pitch and hue tested by Bernstein et al., 1971) are thus unlikely to exhibit

---

[4] S. S. Stevens (1957, p. 154) highlighted a potentially important distinction between two kinds of perceptual continua, namely *prothetic* (magnitude) and *metathetic* (qualitative). The former are concerned with continua having to do with quantity (i.e., *how much*), while the latter describes continua where the stimuli are arranged in qualitative terms (i.e., *what kind* or *where*). Loudness, size, duration, and rate of change all constitute prothetic dimensions with a clear "more than" end (e.g., loud, bright, big) and another, "less than" end (e.g., quiet, dark, small; Walsh, 2003). By contrast, pitch constitutes a metathetic dimension, since a high-pitched tone is different in kind from a low-pitched tone, without necessarily being meaningfully related in a more than/less than way (shape is also metathetic; see L. B. Smith & Sera, 1992). The exact correspondence between these two classes of continua and their underlying neural representations isn't altogether clear (Lewkowicz & Turkewitz, 1980; Stevens, 1971). In adults, prothetic dimensions tend to possess a unitary and well-ordered psychophysics, while metathetic dimensions do not.

Table 1 Summary of crossmodal correspondences that have been shown to influence participants' performance on the speeded classification task together with the null results that have been reported to date

Auditory Dimension	Visual Dimension	Crossmodal Correspondence	High-Pitch/Loud Sound Corresponds to:	Studies
Pitch	Elevation	Yes	High elevation	Ben-Artzi and Marks (1995); Bernstein and Edelstein (1971); Evans and Treisman (2010); Melara and O'Brien (1987); Patching and Quinlan (2002)
	Brightness	Yes	Brighter stimulus	Marks (1987a)
	Lightness	Yes	Lighter stimulus	Marks (1987a); Martino and Marks (1999); Melara (1989a)
	Shape/angularity	Yes	More angular shape	Marks (1987a)
	Size	Yes	Smaller object	Evans and Treisman (2010); Gallace and Spence (2006)
	Spatial frequency	Yes	High spatial frequency	Evans and Treisman (2010)
	Direction of movement	Yes	Upward movement	Clark and Brownell (1976)
	Contrast	No	N/A	Evans and Treisman (2010)
	Hue	No	N/A	Bernstein, Eason, and Schurman (1971)
Loudness	Brightness	Yes	Brighter stimulus	Marks (1987a)
	Lightness	No	N/A	Marks (1987a)

Note. Other crossmodal correspondences demonstrated using other tasks are not mentioned here on the speeded classification task

any crossmodal correspondence.[5] Of course, speeded classification studies may also fail to provide evidence in support of the existence of crossmodal correspondences if there happen to be individual differences in the direction of the associations or matches that people make: This is the case for loudness–brightness, where Marks (1974) found that approximately half of the population tested matched loud sounds to a darker grey surface, while the rest thought the opposite mapping more appropriate instead (matching the louder sounds to lighter grey surfaces instead).

Interim summary

The results of the speeded classification studies reported in this section of the review illustrate that crossmodal correspondences influence the speed (and in some cases accuracy) of human information processing. They demonstrate that people find it harder to ignore distractors in one sensory modality if they happen to vary unpredictably along a dimension that just happens to share a crossmodal correspondence with the dimension along which they are making their speeded classification responses (see Table 1 for a summary of crossmodal correspondences evidenced by speeded classification studies). However, it is important to note that such findings do not necessarily mean that the stimuli presented in the different modalities have been *integrated* at a perceptual

level. Instead, such effects might operate at the level of decision-making/response selection. For, in all of the studies reported in this section, participants had to respond to the target stimuli presented in one modality, and the stimuli in the other modality were always task-irrelevant. These dimensional interactions in speeded classification (Kornblum, Hasbroucq, & Osman, 1990) therefore likely resulted from a failure of participants' *selective attention* (Marks, 2004).

That said, Parise and Spence (2008b) demonstrated that the crossmodal correspondence between visual size and auditory pitch can be demonstrated using a version of the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998)—that is, under conditions where participants had to make speeded discrimination responses to an unpredictable sequence of unimodal auditory and visual target stimuli. Parise and Spence (2008b) demonstrated that people found it much easier to respond to large visual stimuli and low tones with one response key and smaller visual stimuli and higher tones with another response key than they did when the mapping of stimuli to responses was reversed. Given that participants responded to every target in this paradigm, selective attention cannot account for the performance effects observed. Hence, while the results of the speeded classification task studies may be explained in terms of selective attention, not all effects of crossmodal correspondences on the speed of a participant's responding can be explained in this way.

Nevertheless, given the uncertainty over the appropriate level of explanation (decisional vs. perceptual) for the results of studies involving speeded target discrimination, together with claims that many of the effects of cross-

---

[5] Note here also that while pitch is a polar dimension, hue is a circular dimension (Marks, 1978), thus perhaps explaining why people do not match these dimensions crossmodally.

modal correspondence may be decisional in nature (see, e.g., Marks et al., 2003; Melara, 1989b; P. Walker & Smith, 1985), researchers have more recently started to investigate whether crossmodal correspondences influence multisensory integration using tasks where the influence of decisional/response selection on performance can be more easily ruled out.

Do crossmodal correspondences influence multisensory integration/perception?

The redundant-targets effect

J. O. Miller (1991, Experiment 1) conducted a study in which participants had to make a speeded response in a go/no-go task. The target stimuli consisted of a visual target presented from either above or below fixation and a high- or low-pitched tone. These targets could either be presented unimodally or as crossmodally congruent or incongruent bimodal targets. The participants made the same simple speeded response regardless of the target type (hence, redundancy at the level of response selection could be ruled out as an explanation of any crossmodal correspondence effects observed). The participants were instructed to refrain from responding whenever a visual stimulus was presented at fixation or a sound with an intermediate pitch was presented. Miller's results revealed that participants responded significantly more rapidly to congruent than to incongruent bimodal targets ($M = 585$ vs. 611 ms, respectively). The error rates in this study were negligible, thus allowing Miller to rule out a speed–accuracy account of the crossmodal congruency effect reported. What is more, the response time (RT) data violated the race model, thus arguing against the possibility that the two stimuli were processed independently, and instead favouring some form of genuine multisensory integration of the auditory and visual target signals. Another way in which researchers have attempted to reduce/minimize the effect of response selection/decisional biases on participants' performance has been through the use of unspeeded tasks, and it is to the results of such studies that we now turn.

Temporal integration

Parise and Spence (2009) demonstrated that crossmodal correspondences can modulate audiovisual spatiotemporal integration. The participants in their study had to make an unspeeded judgement regarding whether an auditory or visual stimulus had been presented second. The stimulus onset asynchrony (SOA) in this crossmodal temporal order judgement (TOJ) task was varied on a trial-by-trial basis using the method of constant stimuli. The auditory and visual stimuli

presented on each trial were chosen to be either crossmodally congruent or incongruent (see Fig. 4a). In one experiment, the crossmodal correspondence between auditory pitch and visual size was investigated. The visual stimulus consisted of a light grey circle (2.1° or 5.2° of visual angle presented at fixation against a white background), while the auditory stimulus consisted of a briefly presented tone (300 or 4500 Hz). All four possible combinations of auditory and visual stimuli (two crossmodally congruent and the other two incongruent) were presented equiprobably. Parise and Spence's (2009) hypothesis was that if crossmodal associations really do modulate audiovisual integration at a *perceptual* level, then people should find it harder to say which modality stimulus (either auditory or visual) has been presented second on crossmodally congruent as compared to crossmodally incongruent trials (cf. Vatakis et al., 2008; Vatakis & Spence, 2007, 2008, for the same logic applied to the integration of matching vs. mismatching audiovisual speech stimuli).

The results (see Figs. 4b and c) demonstrated that Parise and Spence's (2009) participants found it significantly harder to correctly resolve the temporal order of the auditory and visual stimuli (i.e., the just noticeable difference [JND] was higher) for pairs of stimuli that were crossmodally congruent than for pairs that were incongruent. A very similar pattern of results was also observed in another experiment in which the crossmodal correspondence between auditory pitch/waveform and visual shape was assessed instead. These results are therefore consistent with the view that more pronounced multisensory integration occurs for crossmodally congruent auditory and visual stimuli than for pairs of stimuli that happen to be incongruent. One way to think about the deleterious effect of synaesthetic congruency on the sensitivity of participants' crossmodal TOJs is that it results from the modulation of *temporal ventriloquism* (Morein-Zamir, Soto-Faraco, & Kingstone, 2003) by the unity effect— the claim being that synaesthetic congruency promotes temporal ventriloquism and hence impairs multisensory temporal resolution (Parise & Spence, 2008a). It is, however, worth noting that this particular interpretation has recently been questioned by Keetels and Vroomen (in press). Nevertheless, the key point remains that crossmodal congruency reliably modulates audiovisual temporal perception when assessed by means of performance on a TOJ task (see Parise & Spence, 2009). What is more, Parise and Spence (2008a) went on to demonstrate that audiovisual crossmodal correspondences also modulate the *spatial* (i.e., and not just the temporal) aspects of multisensory integration. That is, crossmodally congruent pairs of auditory and visual stimuli give rise to significantly larger spatial ventriloquism effects than do crossmodally incongruent stimulus pairings (see also Jackson, 1953).
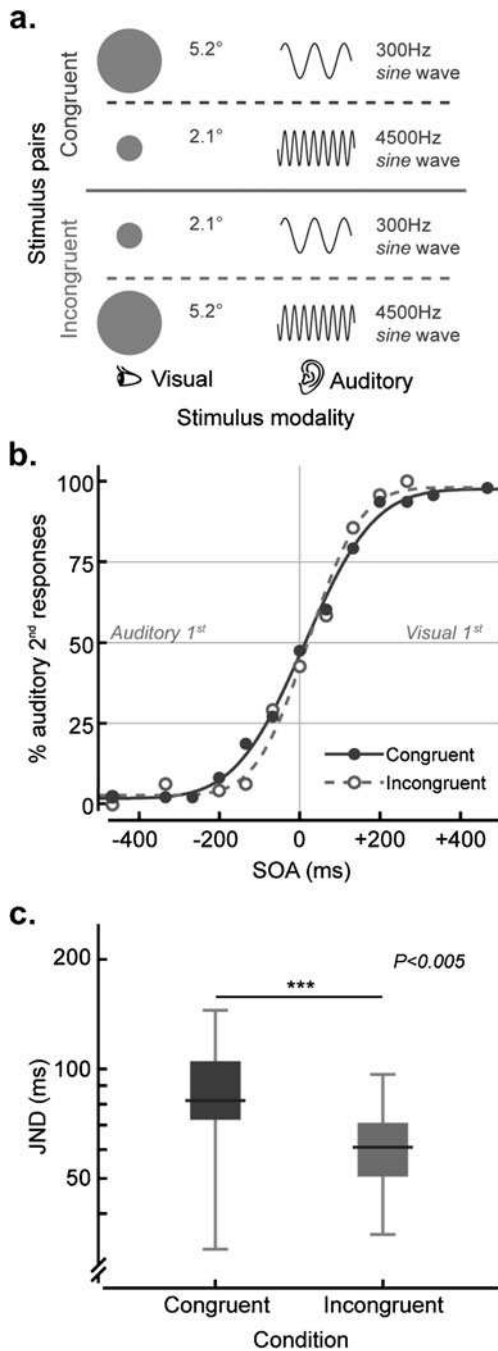
play as moving upward (rather than downward). Meanwhile, if the pitch of the sound decreases, the visual display will be more likely to be judged as moving downward instead. Maeda et al. also demonstrated that the sound only influenced participants' visual motion judgements when its onset occurred within 100 ms or so of the onset of the visual stimulus. Given that a high sensitivity to temporal coincidence is one of the signature features of multisensory integration effects (see Guest, Catmur, Lloyd, & Spence, 2002), this result supports the claim that crossmodal correspondences can have genuinely perceptual consequences. What is more, Maeda et al. also demonstrated that simply presenting the words "up" or "down" did not bias participants' responses concerning whether the ambiguous visual motion display appeared to be moving upwards or downwards. The latter result, which contrasts with some of the results reported earlier (see, e.g., Gallace & Spence, 2006), also argues against a purely semantic interpretation of Maeda et al.'s results. That said, given that it normally takes longer to process semantic than perceptual stimuli for meaning (see Y. C. Chen & Spence, 2011), it would have been interesting to see whether the presentation of directional words would have biased participants' performance if they had been presented *prior* to the visual motion stimulus (cf. Y. C. Chen & Spence, 2011). A conceptually similar set of findings, but involving the perception of auditory and visual motion in depth, was reported by Kitagawa and Ichihara (2002). That is, there also appears to be a crossmodal correspondence between looming (i.e., expanding) visual stimuli and increasing sound intensity (i.e., loudness).

Elsewhere, E. L. Smith et al. (2007) demonstrated that people's perception of the gender of an androgynous face can also be biased by the presentation of a task-irrelevant pure tone in the male or female fundamental-frequency range. The participants in this particular study

Elsewhere, researchers have demonstrated that crossmodal correspondences can modulate people's perception of the direction of motion of ambiguous visual motion displays. For example, Maeda et al. (2004) used a two-alternative forced choice (2AFC) procedure to show that when people are presented with a sound whose pitch ascends, they are significantly more likely to judge a simultaneously presented ambiguous visual motion dis-

were more likely to rate a face as looking male when a task-irrelevant "male" tone (which was lower in pitch) was presented at the same time. By contrast, presenting a higher-frequency "female" tone biased participants to report that the face looked female instead. Interestingly, subsequent experiments demonstrated that these effects were based on the *absolute* frequency of the tone, whereas participants' explicit judgements of the "gender" of the tone as being either "male" or "female" were based on the relative frequency of the tone instead.

Interim summary

The results of the research reported in this section demonstrate that crossmodal correspondences really do modulate multisensory integration/perception. Both the temporal and spatial aspects of multisensory binding are enhanced when crossmodally congruent (as compared to incongruent) pairs of auditory and visual stimuli are presented (Parise & Spence, 2009). When these are taken together with the results reported in the previous section, it would appear that crossmodal correspondences likely operate in a fairly automatic fashion (see Evans & Treisman, 2010; P. Walker & Smith, 1984), influencing both the speed and accuracy of a participant's responses (e.g., J. O. Miller, 1991; see also Makovac & Gerbino, 2010). These effects can be observed in tasks where the participants' responses are unspeeded (Kitazawa & Ichihara, 2002; Maeda et al., 2004; Parise & Spence, 2009; E. L. Smith et al., 2007) and in the absence of any explicit attentional manipulation (which is an integral feature of all speeded discrimination studies; see also Parise & Spence, 2008b).

The available evidence now supports the claim that dimensional interactions between auditory and visual stimuli sharing a particular crossmodal correspondence influence performance not only at a decisional level, but also at a more perceptual level (see also L. Chen & Zhou, 2010; Evans & Treisman, 2010; Kuze, 1995; A. Miller, Werner, & Wapner, 1958; O'Leary & Rhodes, 1984; Pratt, 1930; Roffler & Butler, 1968). Of course, the relative contributions of perceptual and decisional factors to the crossmodal interactions taking place between auditory and visual stimuli likely depend on the particular pairing of dimensions (not to mention the task; see Marks et al., 2003, p. 143) under consideration. While certain forms of crossmodal correspondence appear capable of modulating participants' *perceptions* of ambiguous visual motion displays (Kitagawa & Ichihara, 2002; Maeda et al., 2004), not to mention of androgynous faces (E. L. Smith et al., 2007), other forms of correspondence appear to operate at a more *decisional* level. For example, in contrast to the perceptual effects of crossmodal corre-

spondences just mentioned, Gallace and Spence (2006, their note 4) found that the presentation of either a crossmodally congruent or incongruent sound (varying in loudness) did not actually change the perceived size of the circle it was presented with (despite the fact that participants' RTs changed significantly). Thus, when this result is taken together with Parise and Spence's (2009) results, the most parsimonious suggestion regarding the effect of the crossmodal correspondence between size and pitch is that although congruency may alter the strength of the coupling between auditory and visual stimuli, this does not necessarily mean that congruency will influence the perceptual attributes of the component stimuli. In conclusion, the fact that seemingly perceptual effects can be demonstrated in certain tasks, and for certain crossmodal correspondences, should not necessarily be taken to mean that they will be demonstrated for other tasks or crossmodal correspondences.

Now, while Parise and Spence's (2009) results demonstrated increased spatial and temporal integration for pairs of auditory and visual stimuli that share a crossmodal correspondence, other researchers have argued that the perceptual salience of crossmodally congruent (as compared to incongruent) pairs of auditory and visual stimuli may also be heightened (see, e.g., Evans & Treisman, 2010; Marks, 1987a; though see Marks et al., 2003; Melara, 1989b). Indeed, the latest electrophysiological research has demonstrated that when crossmodally matching (as compared to mismatching) pairs of stimuli are presented, the early neural evoked response (e.g., N1) may peak significantly earlier and have an enhanced amplitude (e.g., Kovic, Plunkett, & Westermann, 2009; see also Seo et al., 2010). Results such as these have been taken to support a perceptual enhancement account of at least some part of certain crossmodal correspondence effects. However, seemingly standing against the perceptual enhancement account are the findings of a well-controlled series of psychophysical studies reported by Marks et al. (2003). These researchers used a two-interval same–different procedure to demonstrate that the presentation of a visual stimulus had no effect on auditory sensitivity as assessed by performance in a pitch discrimination task. The participants in this particular study were presented with two bimodal stimuli on each trial, one after the other. On each trial, the participants had to decide whether the brightness of the visual component or the pitch of the auditory component, was the same or different for the two stimuli. Analysis of the results using signal detection theory revealed *no* evidence of a change in perceptual sensitivity (d'; nor, for that matter, much evidence of a shift in criterion c) when crossmodally congruent stimuli were presented, as compared to when incongruent pairings of auditory and visual stimuli were presented instead. What is more, no effect of

auditory stimulation on visual brightness discrimination performance was observed either. These results therefore led Marks et al. (2003, p. 125) to conclude that "cross-modal interactions result primarily from relatively late decisional processes (e.g., shifts in response criterion or 'bias')."

It is, at present, unclear how this discrepancy between the significant perceptual effects of crossmodal correspondence reported in certain studies (e.g., Kitagawa & Ichihara, 2002; Maeda et al., 2004; Parise & Spence, 2009; Radeau & Bertelson, 1987; E. L. Smith et al., 2007; Thomas, 1941) and the null results on perceptual sensitivity reported in others (e.g., Marks et al., 2003) should be resolved. It may simply be that while certain crossmodal correspondences (such as the pitch–size and pitch–angularity correspondences investigated by Parise & Spence, 2009) result in perceptual interactions, and possibly also decisional-level effects (but see also Gallace & Spence, 2006, note 4), others (such as the pitch–brightness correspondence studied by Marks et al., 2003) operate primarily at the decisional level instead. Alternatively, however, it may equally well be that although crossmodal correspondences may enhance the spatiotemporal aspects of multisensory integration (i.e., they may impact on the strength of coupling between a pair of unimodal stimuli), that does not mean that they will necessarily impact on the perceptual discriminability or salience of the stimuli concerned as well. That is, it is possible that crossmodal correspondences may facilitate crossmodal binding, while not necessarily enhancing a participant's perception of (or sensitivity to) the multisensory object or event thus formed (cf. Lippert, Logothetis, & Kayser, 2007; Spence & Ngo, in press). Of course, to better understand this dissociation, it will be useful in future research to investigate specific crossmodal correspondences using a variety of different experimental paradigms (rather than investigating different correspondences using different tasks, which has largely been the case thus far).

What is the relation between crossmodal correspondences and synaesthesia?

As noted earlier, describing crossmodal correspondences as "synaesthetic correspondences" is in some sense pejorative, for it would seem to imply that there is a meaningful link between the crossmodal correspondence effects seen in "normal" people and the crossmodal confusions that are often seen in "full-blown" synaesthetes (see Cytowic & Eagleman, 2009, for a recent review of the latter). Indeed, Martino and Marks (2001) argued that crossmodal correspondences share many similarities with full-blown synaesthesia (see also Marks, 1989a). They go further, though, by suggesting that the two phenomena may rely on many of the same

underlying neural mechanisms,[6] a view that has been echoed more recently by several other researchers (e.g., Sagiv & Ward, 2006; Ward, Huckstep, & Tsakanikos, 2006). In fact, it is currently popular for researchers to argue that we all lie somewhere along a continuum from normal (i.e., non-synaesthetic) to full-blown synaesthetic behaviour (e.g., Martino & Marks, 2001; Simner et al., 2005; Ward et al., 2006; see also Cohen Kadosh & Henik, 2007; Rader & Tellegen, 1987). However, I would argue that it is not necessarily appropriate to think of *all* crossmodal correspondences as bearing much of a relation to synaesthesia (see also Elias et al., 2003).

On the one hand, one of the reasons for thinking that full-blown synaesthesia might be related to the crossmodal correspondences experienced by nonsynaesthetes is the striking similarity in the nature of many of the correspondences that have been observed in the two groups over the years (see Marks, 2000). For example, both synaesthetes and nonsynaesthetes tend to associate high-pitched sounds with jagged sharp visual images, low-pitched tones with smooth rounded forms (Karwoski & Odbert, 1938), and high-pitched sounds with small, bright percepts (Marks, 1974, 1987a; Marks et al., 1987). On the other hand, if synaesthetes really were like nonsynaesthetes, except for a difference in the strength of their crossmodal correspondences, then one might expect to see enhanced multisensory integration for crossmodally congruent stimuli in perceptual tasks and/or greater interference in speeded classification tasks in synaesthetes as compared to nonsynaesthetes. However, there is currently little evidence to support such a suggestion. In fact, preliminary testing of this notion here in Oxford, using a version of the speeded classification task reported by Gallace and Spence (2006), failed to highlight any such differences between synaesthetes and nonsynaesthetes. It is possible, though, that tasks that rely more on multisensory perceptual interactions than on a failure of selective attention (as was the case in Gallace and Spence's speeded classification studies) might give rise to a different pattern of results.

P. Walker et al. (2010, p. 21) recently suggested that crossmodal correspondences reflect "an unlearned aspect of perception." I will argue shortly that such correspondences,

---

[6] In their review article, Martino and Marks (2001, p. 61) distinguished between strong and weak forms of synaesthesia. *Strong synaesthesia*, they argue, is characterized by vivid percepts in one sensory modality being induced by stimulation in another (though note that many cases of synaesthesia are actually intramodal, as in colour–grapheme synaesthesia, probably the most common type of synaesthesia; see Day, 2005). *Weak synaesthesia*, they suggest, is "characterized by cross-sensory correspondences expressed through language, perceptual similarity, and perceptual interactions during information processing" (see also Marks, 1975, 2000). Despite there being some dissimilarities between these two phenomena, Martino and Marks (2001) claim that the two forms of synaesthesia draw on similar neuronal mechanisms.

which can be categorized as structural correspondences, may indeed bear fruitful comparison with synaesthesia. However, I believe that other kinds of correspondence may need a different kind of theoretical interpretation. Indeed, an alternative way to think about those crossmodal correspondences that result from the internalization of the statistics of the natural environment comes from the latest research on Bayesian integration theory, and it is to that we turn next.

Explaining crossmodal correspondences in terms of Bayesian priors

It is currently becoming increasingly popular to model multisensory integration in terms of Bayesian integration theory (cf. Ernst, 2006; Ernst & Bülthoff, 2004). The suggestion is that humans (and other animals; see M. L. Morgan, DeAngelis, & Angelaki, 2008) may combine stimuli in a statistically optimal manner by combining prior knowledge and sensory information and weighting each of them by their relative reliabilities. It would seem only natural, therefore, to consider how the notion of crossmodal correspondences might be modelled within such a framework as a form of prior knowledge. According to Marc Ernst (2006), the strength of crossmodal coupling is a function of our sensory system's prior knowledge that certain stimuli "go together" crossmodally: Such prior knowledge concerning the mapping between sensory signals can be modelled by a *coupling prior* (see also Roach et al., 2006; Shams & Beierholm, 2010; Shams, Ma, & Beierholm, 2005), representing the expected (i.e., a priori) joint distribution of the signals. In the case of bimodal integration, the prior distribution can be considered as a 2-D Gaussian distribution with infinite variance along the positive diagonal (the identity line). Variance along the negative diagonal then depends on the specific combinations of cues under consideration (see Fig. 5). It is the latter variance that we are interested in here. The coupling prior influences the strength of coupling in inverse proportion to its variance. A variance approaching infinity (i.e., a flat prior) means that the signals presented in different sensory modalities are treated as independent and there is no interaction between them; conversely, a variance approaching 0 indicates that the signals will be completely fused into an integrated multisensory percept instead. Intermediate values likely result in a coupling of the unimodal signals without necessarily giving rise to sensory fusion.[7]

---

[7] Note here also that the effect of increasing the temporal disparity between the auditory and visual signals (as in Bernstein & Edelstein's, 1971, speeded classification study) on perception/performance can be simply modelled as an increase in the variance of the coupling prior. That is, the weight of the coupling prior declines with increasing temporal separation (see Ernst, 2007).

Three schematic examples showing how visual and auditory signals are combined given different coupling priors (columns).
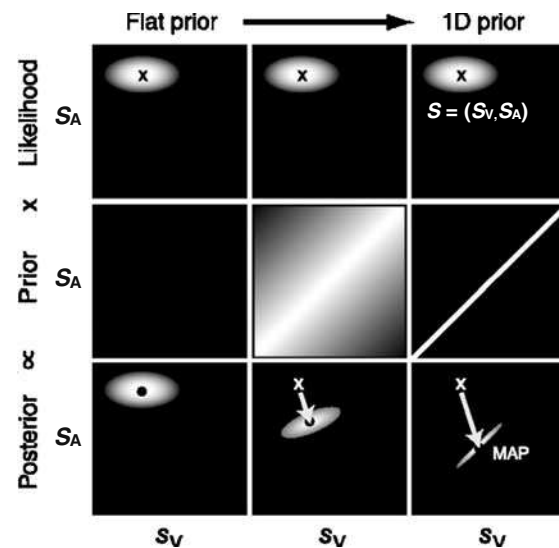

Fig. 5 Three schematic examples illustrating how visual and auditory signals with different priors (columns; with the variance of the coupling prior decreasing from left to right) can be combined. S_A = auditory stimulus; S_V = visual stimulus; $S = (S_V, S_A)$ = a physical stimulus having a visual and an auditory property. Top row: Likelihood distributions with standard deviation σ_V double σ_A; "x" denotes the actual stimulus. Middle row: Prior distributions—on the left, flat prior $\sigma_1^2 = \infty$, $\sigma_2^2 = \infty$; in the middle, $\sigma_1^2 = \infty$, $\infty > \sigma_2^2 > 0$; on the right, $\sigma_1^2 = \infty$, $\sigma_2^2 = 0$. Bottom row: Posterior distributions, which are the product of the likelihood and prior distributions. The maximum a posteriori (MAP) estimate is indicated by the •. The arrows indicate the bias in the MAP estimate relative to the physical stimulus (x). [From Fig. 1 of "Learning to Integrate Arbitrary Signals From Vision and Touch," by M. O. Ernst, 2007, *Journal of Vision*, 7 (5), 7:1–14. Copyright 2007 by the Association for Research in Vision and Ophthalmology. Adapted with permission.]

According to Bayesian models of multisensory integration, the reliability of a person's estimate regarding intersensory conflict is proportional to the strength of the coupling between the signals being integrated (see Ernst, 2007). The stronger the coupling, the more likely it is that the original unimodal signals will be fused completely into a single integrated multisensory percept. Behaviourally, this will show up as a reduction in the reliability of a person's conflict estimate (i.e., as a higher discrimination threshold). In fact, if fusion is complete, the conflict should disappear perceptually. By contrast, weaker coupling gives rise to only partial integration, with the perceptual system still retaining access to reliable conflict estimates (thus resulting in a lower discrimination threshold).

The effect of crossmodal correspondences on human information processing and multisensory integration can be interpreted in terms of differences in the strength (i.e., variance) of the coupling prior: A smaller variance for

crossmodally congruent stimulus pairs than for crossmodally incongruent pairs. According to this Bayesian view (Parise & Spence, 2009), the existence of crossmodal correspondences, at least those based on natural statistical correlations, helps the perceptual system integrate the appropriate stimuli from different sensory modalities, and thus helps resolve the crossmodal binding problem (see Ernst, 2007; Spence, Ngo, et al., 2010). That said, thinking about crossmodal correspondences in terms of variations in the strength of the coupling prior seems a very different kind of explanation from the weak-synaesthesia account posited by Martino and Marks (2000; see also Karwoski & Odbert, 1938). While the former account suggests a mechanism acting on the likelihood that certain combinations of stimuli will be bound together, the latter argues for the existence of additional neural connections (Rouw & Scholte, 2007; or a breakdown of modularity, as in Baron-Cohen, Harrison, Goldstein, & Wyke, 1993) between the parts (modules) of the brain coding the information from different senses; and even, sometimes, for different attributes/features within a sense (see Day, 2005). What is more, it is not always clear whether the coupling prior should be thought of as affecting the perceptual or the decision-making level of information processing, or both.

Modifying coupling priors, weak synaesthesia, and the varieties of crossmodal correspondence

Researchers have recently shown that the variance of the coupling prior (and hence the strength of crossmodal coupling) can be modulated by a person's knowledge that various unimodal stimuli originate from a single object (Helbig & Ernst, 2007; see also E. A. Miller, 1972; Spence, 2007; Welch, 1972) or event (Bresciani, Dammeier, & Ernst, 2006). What is more, repeated exposure to the statistical co-occurrence of particular pairs of stimuli can also influence the coupling prior, even for pairs of stimulus dimensions that happen, initially, to be unrelated (e.g., Ernst, 2007; see also Teramoto, Hidaka, & Sugita, 2010). So, for example, Ernst (2007) demonstrated that exposing people to an arbitrary correspondence between the luminance of a visual object and its felt stiffness, a haptically ascertained stimulus property that is not correlated with luminance in the natural environment, can give rise to a change in the coupling prior. The participants in Ernst's (2007) study were trained with multisensory stimuli in which an artificial correlation had been introduced between the stimulus dimensions: For some of the participants, the stiffer the object, the brighter it appeared, while this mapping was reversed for other participants.

The results highlighted a significant change in participants' discrimination performance when their responses to congruent and incongruent pairs of haptic stimuli were compared before and after training. These changes were attributable to changes in the distribution of the coupling prior. Given that all of the training in Ernst's (2007) study took place within a single day (in a session lasting between 1.5 and 2.5 h), one can easily imagine how much stronger the coupling prior might be for pairs of stimuli that have been correlated over the course of a person's lifetime. However, the fact that extensive training (even over tens of thousands of trials) with specific crossmodal pairings of auditory and visual stimuli does not give rise to synaesthetic concurrents (e.g., Howells, 1944; Kelly, 1934; see also Kusnir & Thut, 2010), despite presumably changing the variance of participants' coupling priors, again argues against the appropriateness of synaesthesia as a model for this particular kind of (statistical) correspondence.

It would seem eminently plausible that a number of the crossmodal correspondences reported in this review can be explained in terms of the correlation between certain pairs of attributes in the natural environment (McMahon & Bonner, 1983; Simpson et al., 1956; see also Kadosh, Henik, & Walsh, 2007). It would obviously make sense for our brains to take advantage of the regularities that exist in the world around us (cf. Shepard, 1994; von Kriegstein & Giraud, 2006) when deciding which of the many possible unisensory stimuli to integrate. For example, in nature, the resonant frequency of an object is related to its size (the larger the object, the lower the frequency it makes when struck, dropped, sounded, etc.; Carello, Anderson, & Kunkler-Peck, 1998; Coward & Stevens, 2004; Grassi, 2005; Marks, 2000; McMahon & Bonner, 1983; D. R. R. Smith, Patterson, Turner, Kawahara, & Irino, 2005; Spence & Zampini, 2006; Stumpf, 1883). The frequency of acoustic resonance of a body also depends on its mass, its tension, and so forth. There is even information in the ambient acoustic array to specify the approximate shape of a resonating body (Spence & Zampini, 2006). Furthermore, larger objects (e.g., animals) normally emit louder sounds (Bee et al., 2000; Davies & Halliday, 1978; Fitch & Reby, 2001; Harrington, 1987). Therefore, it soon becomes apparent just how many statistical correlations might actually be out there between the stimuli we perceive through the auditory, visual, and tactile modalities. What is more, there is presumably also a correlation between the size of objects and their likely elevation—elephants were never going to fly (cf. Berlin, 1994)! The existence of such natural correlations might therefore help explain the reliable crossmodal correspondences that have been documented both between pitch and visual size (Evans & Treisman, 2010; Gallace & Spence, 2006) and between pitch and visual elevation (Ben-Artzi & Marks, 1995; Bernstein & Edelstein, 1971; Evans & Treisman, 2010; Melara &

O'Brien, 1987; J. O. Miller, 1991; O'Leary & Rhodes, 1984; Patchling & Quinlan, 2002).[8]

Marks and his colleagues were unable to come up with any plausible natural correlations that might explain the existence of the crossmodal correspondences that have been reported between either loudness or pitch and brightness (see Marks et al., 2003, p. 127). However, that does not necessarily mean that such natural correlations don't exist: Think of thunder and lightning, or an explosion. The loudness and brightness of these events are both proportional to the energy contained in the event itself, and therefore are necessarily correlated. Just how many of these natural correlations there are is obviously a question for conjecture (not to mention future research). Taking a Gibsonian perspective, though, one might even put the argument in terms of the perceptual system *directly* perceiving certain invariants of the environmental array that can be picked up by the information available to a number of senses simultaneously (Gibson, 1966; Spence & Zampini, 2006; Stoffregen & Bardy, 2001). Certainly, those crossmodal correspondences that result from the pick-up of statistical correlations in the environment can most easily be interpreted in terms of such invariants.

As hinted at earlier, it may be most appropriate to consider the likelihood that multiple mechanisms support the many different examples of crossmodal correspondences that have been reported to date (see Table 2). While certain crossmodal correspondences may reflect the internalization of the statistics of the natural environment (and may perhaps best be modelled in terms of coupling priors according to Bayesian integration theory; Ernst, 2007), others may result from the peculiarities of the neural systems we happen to use to code sensory information (see Marks, 1978, 1987a; Mondloch & Maurer, 2004; S. S. Stevens, 1957; Walsh, 2003). It is the latter crossmodal correspondences, then, that bear the closest comparison with synaesthesia proper, given claims that we are all born synaesthetic (Maurer, 1997; Maurer & Mondloch, 2005; Wagner & Dobkins, 2009).

The possibility that there may be several kinds of crossmodal correspondence (some learned, others possibly innate) is certainly consistent with the view that while a number of crossmodal correspondences may be present from birth (see Lewkowicz & Turkewitz, 1980; Mondloch & Maurer, 2004; P. Walker et al., 2010), others, especially those that appear to be semantically mediated, only emerge

after the advent of language (see Marks, 1984; Marks et al., 1987; L. B. Smith & Sera, 1992; see also Cohen Kadosh, Henik, & Walsh, 2009). In future research, it will therefore be interesting to determine whether the stage of development at which specific crossmodal correspondences emerge bears any relation to the mechanisms underlying such crossmodal correspondence effects in adults (i.e., whether they are semantically vs. perceptually mediated; cf. L. B. Smith & Sera, 1992).

However, with regard to the question of the innateness of certain crossmodal correspondences, it is important to remember that the youngest infants to have been tested in such tasks were 20–30 days old in Lewkowicz and Turkewitz's (1980) study, and 3–4 months old in P. Walker et al.'s (2010) study. Given the speed of learning observed in Ernst's (2007) study (which occurred over a period lasting only an hour or two; see also Conway & Christiansen, 2006), and given infants' sensitivity to at least certain of the statistical regularities in the environment (e.g., as when computing conditional probabilities in speech stimuli; Aslin, Saffran, & Newport, 1998), it could be argued that there is sufficient time for their plastic perceptual systems to learn these statistical crossmodal correspondences during early development (not to mention while in the womb). Therefore, the question of the innateness of crossmodal correspondences, at least for now, remains unresolved.

Assessing the neural substrates underlying crossmodal correspondences

Thus far in this tutorial review, little consideration has been given to any consideration of the neural underpinnings of crossmodal correspondence effects. Indeed, much of the research on the Bayesian approach to the study of multisensory integration discussed in the previous section can be criticized for failing to specify how the "ideal observer" approach is (or even might be) instantiated in the brain (though see M. L. Morgan et al., 2008, for an exception). That said, it would seem sensible to pause and consider, in closing, whether different neural substrates might underlie the different classes of crossmodal correspondence that have been outlined here. Should this prove to be the case, neuroimaging studies might, in the future, provide a fruitful means of distinguishing between the various types of crossmodal correspondence. This could be particularly helpful given that, as we have seen already, the various kinds of crossmodal correspondences often give rise to very similar patterns of behavioural responding (e.g., in terms of patterns of crossmodal facilitation vs. inhibition; though see Elias et al., 2003).

There would appear to be at least two key questions here with regard to the neural underpinnings of crossmodal correspondences: First, where are crossmodal correspondences

---

[8] Of course, the linguistic mediation account mentioned earlier (see Martino & Marks, 1999) provides just as convincing a post-hoc explanation for the latter association. The fact that we use the same pair of adjectives ("high" and "low") to describe the ends of both the pitch and elevation dimensions means that a semantic interaction account of the congruency effects in speeded classification studies where pitch and elevation vary is plausible.

Table 2 Summary of the three principal types of crossmodal correspondence

Type of Crossmodal Correspondence	Example	Time Course and Explanation	Consequences
Structural	Loudness–brightness	Possibly innate, but may also depend on maturation of neural structures for stimulus coding	Perceptual & decisional
Statistical	Pitch–elevation Pitch–size Loudness–size	Learned: Coupling priors established on the basis of experience with regularities of the environment	Perceptual & decisional
Semantic	Pitch–elevation Pitch–spatial frequency	Learned: Emerge following language development as certain terms come to be associated with more than one perceptual continuum	Primarily decisional

Note. Other crossmodal correspondences between auditory and visual stimulus dimensions that are currently harder to place include the crossmodal correspondence between pitch and brightness

represented in the human brain? And second, what neural changes are associated with the establishment of new cross-modal correspondences, or associations between specific auditory and visual features? With regard to the former question, polysensory areas in the temporal cortex (e.g., superior temporal sulcus) have already been shown to respond more vigorously to crossmodally congruent, as compared to incongruent, pairings of simultaneously presented auditory and visual object action stimuli (e.g., Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp, Lee, Argall, & Martin, 2004; Naumer & Kaiser, 2010). It is currently an open question whether the same area would be modulated by the crossmodal congruency of more basic stimulus features as well. Meanwhile, preliminary neuropsychological evidence suggests that damage to the angular gyrus (which lies within the temporo–parieto–occipital region) can interfere with a person's ability to match stimuli crossmodally, as assessed by the bouba/kiki test (see Ramachandran & Hubbard, 2003).

With regard to the second question, several recent studies have documented changes in neural activity as two initially unrelated stimuli (one auditory, the other visual) become reliably associated in the mind of an observer (e.g., Baier, Kleinschmidt, & Müller, 2006; Zangenehpour & Zatorre, 2010). Just as in Ernst's (2007) study discussed earlier, these changes have been observed after surprisingly short exposure periods (i.e., less than 45 minutes' exposure in Zangenehpour and Zatorre's study resulted in effects that were still present a day later). For example, the participants in Baier et al.'s study were presented with objects and their characteristic sounds. They were given a cue prior to each stimulus pair indicating whether they should respond to the auditory or the visual component of the target. Significant changes in the cue-induced preparatory neural activity were observed. Prior to training, the presentation of the cue resulted in enhanced activity in the task-relevant sensory system, while suppression was observed in the other, task-irrelevant sensory cortex (and associated thalamus). How-

ever, once the participant had learned the association between the pair of stimuli (through repeated pairing), activity levels in the cortices associated with *both* modalities were enhanced during the cue period, regardless of the target modality that happened to be task-relevant (see also Zangenehpour & Zatorre, 2010). Such results support the view that there may be an enhanced spread of attention between stimulus features that share some form of cross-modal correspondence (or learned association; see also Fiebelkorn, Foxe, & Molholm, 2010).

Meanwhile, in another study, von Kriegstein and Giraud (2006) demonstrated that brief exposure to ecologically valid couplings of auditory and visual speech stimuli lead very rapidly to enhanced functional coupling (interestingly, though no such effect was observed, at this very short time frame, for randomly coupled pairs of stimuli, as when a voice was paired with a ring tone). As a consequence, the subsequent presentation of one component of a stimulus pair (e.g., voice) can come to elicit patterns of neural activity that would normally be associated with stimulation in the other sensory modality (see also von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005). In future neuro-imaging studies, it will be interesting to look at the neural changes associated with learning novel crossmodal associations over a much longer time period than has been used in the studies reviewed here.

The evidence emerging from the neuroimaging studies published to date therefore supports the view that the establishment of novel learned associations between auditory and visual stimuli (presumably shaped by previous associative learning) results in enhanced crossmodal connectivity (Zangenehpour & Zatorre, 2010). Consequently, one unimodal component of a multisensory pair is sufficient to activate the association representing the other unimodal component. Such functional changes may, in turn, be expected to facilitate both multisensory perception and multisensory integration, as well as supporting the

spread of attention and enhancing memory (see Fiebelkorn et al., 2010; Murray et al., 2004; E. L. Smith et al., 2007).

Conclusions

Traditionally, researchers interested in the topic of multisensory perception have tended to focus their efforts on studying the influence of spatiotemporal factors on multisensory integration (see Calvert et al., 2004; Spence & Driver, 2004; Wallace et al., 2004). However, the last few years have seen a rapid growth of interest in the other stimulus-related factors that may also modulate multisensory integration and the efficiency of human information processing (see Naumer & Kaiser, 2010, for a review). Both semantic congruency and cross-modal correspondences have now been demonstrated to modulate multisensory integration at both the decisional and perceptual levels (e.g., Y. C. Chen & Spence, 2010; Maeda et al., 2004; Parise & Spence, 2009; E. L. Smith et al., 2007). To date, crossmodal correspondences have been shown to influence people's performance in a wide range of different paradigms, including direct crossmodal matching, speeded classification tasks, speeded simple detection tasks, the Implicit Association Test, unspeeded TOJs, spatial localiza-tion, and perceptual discrimination tasks. The claim is that crossmodal correspondences may be established automatically (Evans & Treisman, 2010; P. Walker & Smith, 1984) and that they can affect late decisional processes as well as, under certain conditions, perceptual integration (Maeda et al., 2004; Marks et al., 2003; Parise & Spence, 2009; E. L. Smith et al., 2007). Taken together, these various findings highlight the importance of crossmodal correspondences to human infor-mation processing. To make further progress, though, I would argue that researchers will need to make a much clearer distinction between the various kinds of crossmodal corre-spondence that have been reported to date. This may be especially important given that they may reflect the influence of different underlying neural substrates, and may even have qualitatively different effects on human perception and performance (cf. Westbury, 2005). The evidence that has been reviewed here would appear to support the existence of at least three different kinds of crossmodal correspondence (see Table 2).

Structural correspondences The first class of crossmodal correspondence result from the peculiarities of the neural systems we use to code sensory information. Marks (1978) describes these as intrinsic attributes of the perceptual system's organization. One such idiosyncrasy highlighted early on by S. S. Stevens (1957) is that increases in stimulus intensity (regardless of the modality of the stimulus) generally appear to be represented by increased neural firing. He suggested that this might provide a

putative neural correspondence that could potentially underpin the psychological or phenomenological corre-spondence between loudness and brightness (see also Marks, 1978). Another kind of structural correspondence may relate to recent suggestions that there is a generalized system in the inferior parietal cortex for representing magnitude (see Walsh's, 2003, A Theory of Magnitude: ATOM). Such a common mechanism for coding magnitude could then presumably also provide the neural substrate for other crossmodal correspondences between pairs of pro-thetic (i.e., magnitude-based) dimensions or continua. One can think of crossmodal correspondences as simply the by-products of the architecture/mechanisms underlying the operation of the cognitive system. Alternatively, two sensory features might simply be coded in nearby brain areas, and therefore be more likely to be associated crossmodally (see Ramachandran & Hubbard, 2001). Or else, due to the principle of neural economy, the brain might use similar mechanisms (although possibly in far-apart brain areas) to process features from different sensory modalities, which might, as a consequence, happen to become associated. Thus, numerous possible causes can be postulated to help explain the existence of such structural correspondences.

Statistical correspondences The second class of cross-modal correspondence for which there is robust support reflect an adaptive response by our brains to the regularities of the world in which we live. Such statistical correspondences reflect the internalization of the natural correlations between stimulus attributes that exist in the environment (see Marks, 2000; R. Walker, 1987). It would appear that such crossmodal correspond-ences can be fruitfully framed in terms of coupling priors in Bayesian integration theory (e.g., Ernst, 2006, 2007; Parise & Spence, 2009). Crossmodal correspondences based on such statistical regularities are also more likely to be universal than are semantically mediated correspond-ences (given that, e.g., the resonance properties of objects are determined by physics, not culture, and so will be the same the world over). The learned associations highlighted by recent neuroimaging studies (e.g., Baier et al., 2006; Zangenehpour & Zatorre, 2010; see also Teramoto et al., 2010) would also seem to fit under the heading of statistical correspondences. However, that said, such correspondences are clearly not going to be universal, in that they reflect the response of the organism (mediated by associative learning) to the idiosyncratic statistics of the multisensory environ-ment in which it has recently found itself.

Semantically mediated correspondences The third class of crossmodal correspondence result when common linguistic terms are used to describe the stimuli falling along different

continua (Gallace & Spence, 2006; Long, 1977; Martino & Marks, 1999; Melara, 1989b; P. Walker & Smith, 1984). One of the most oft-cited examples of this class of correspondence is that documented between pitch and elevation. According to Stumpf (1883), pretty much every language uses the same words, "low" and "high," to describe stimuli that vary in pitch. Given that we use the same adjectives to describe the elevation of visual stimuli, linguistically mediated correspondences may underlie a third class of crossmodal correspondence (e.g., between pitch and elevation). Martino and Marks (1999) have developed a semantic coding hypothesis to account for such correspondences, which emerge only after the onset of language (see Marks, 1984; Marks et al., 1987; L. B. Smith & Sera, 1992) and appear to operate almost exclusively on later decisional stages of information processing (rather than on perception). Semantic correspondences appear to be almost exclusively contextually determined (i.e., there are no absolute matches between stimuli in the different modalities).[9]

There are, of course, limitations to this tripartite system for coding crossmodal correspondences. First, it is not clear, at present, what is the most appropriate explanation for certain of the crossmodal correspondences that have been reported to date. I would argue that the correspondence between auditory pitch and visual brightness currently falls in this category. However, further assessing the developmental time course and perceptual/decisional consequences of manipulating the congruency of stimuli in terms of this correspondence might provide some hints as to which category it should be placed in. Neuroimaging/ electrophysiological studies might also help (cf. Kovic et al., 2009). Put simply, one might expect to see effects of crossmodal correspondences localized to sensory cortices for perceptual effects, to prefrontal areas for decisional effects, and to language areas for more linguistically based effects.

Second, the crossmodal correspondence between pitch and elevation can seemingly be accounted for by at least two different kinds of correspondence. On the one hand, our brains may have internalized the natural statistics of the environment in terms of coupling priors (Ernst, 2007; Parise & Spence, 2009). However, the fact that we use the same words for elevation and pitch means that the semantic mediation hypothesis (e.g., Martino & Marks, 1999) can also be used to explain this correspondence. This ambiguity speaks both to the complexity of the endeavour (i.e., trying to categorize all of the crossmodal correspondences that have been reported to date) and to the close and possibly

nonarbitrary link between language and perception that gave rise to the field of sound symbolism all those years ago (see Hinton et al., 1994). Nevertheless, should researchers be able to demonstrate enhanced multisensory integration (i.e., at a perceptual level) between, say, pitch and elevation when congruent versus incongruent pairs of stimuli are presented, that might provide support for the statistical correspondence account. Of course, a more direct means of testing for such regularities directly in the environment might be to record a very large amount of audiovisual footage from the real world and look for correlations between pitch and elevation (cf. Hancock, Baddeley, & Smith, 1992; Pan, Yang, Faloutsos, & Duygulu, 2004).

The nonexclusivity of the various kinds of correspondence becomes all the more apparent when one considers the case of pairs of dimensions that refer to the same crossmodal property of an object (such as its size) or event (such as its duration). Amodal stimulus attributes inferred from different sensory inputs will be correlated statistically in natural environments: For example, the auditory and visual duration of an event are likely to be correlated, as are the seen, felt, and heard size of an object. However, people are also likely to use the same words to describe them: Think "small" and "big," or "short" and "long," terms that are equally applicable to the seen or felt size of an object or to the duration of an event. Thus, crossmodal correspondences based on the redundant coding of the same amodal stimulus attribute are also likely to be explainable in terms of both statistical and semantically mediated correspondences, at least in those who are old enough to speak. Relevant here is Meltzoff and Borton's (1979) finding that 29-day-old infants are already able to match the shape of a pacifier seen visually to that explored orally beforehand. Such results, demonstrating the existence of crossmodal correspondences very early in life, suggest but by no means prove that structural correspondences may also play some role here.

Finally, it is important to note that this tripartite categorization is by no means meant to be exhaustive. There may well be other classes of crossmodal correspondence, such as, for example, between stimulus dimensions that are based on the effect (be it arousing, affective, etc.) that a stimulus has on the observer. Nevertheless, despite these caveats and concerns, it seems clear that crossmodal correspondences, in all their forms, play an important, if currently underappreciated, role in constraining the crossmodal binding problem.

---

[9] One unresolved but intriguing question here concerns why so many languages use the same terms to describe both pitch and elevation. It would seem plausible that such semantic correspondences may build on either structural or statistical correspondences, though note that the direction of causality has yet to be established (see also L. B. Smith & Sera, 1992).

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*, 257–262.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324.

Baier, B., Kleinschmidt, A., & Müller, N. (2006). Cross-modal processing in early visual and auditory cortices depends on the statistical relation of multisensory information. *The Journal of Neuroscience, 26*, 12260–12265.

Baron-Cohen, S., Harrison, J., Goldstein, L. H., & Wyke, M. (1993). Coloured speech perception: Is synaesthesia what happens when modularity breaks down? *Perception, 22*, 419–426.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004). Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience, 7*, 1190–1192.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41*, 809–823.

Bedford, F. L. (2001). Towards a general law of numerical/object identity. *Cahiers de Psychologie Cognitive, 20*, 113–175.

Bee, M. A., Perrill, S. A., & Owen, P. C. (2000). Male green frogs lower the pitch of acoustic signals in defense of territories: A possible dishonest signal of size? *Behavioral Ecology, 11*, 169–177.

Belkin, K., Martin, R., Kemp, S. E., & Gilbert, A. N. (1997). Auditory pitch as a perceptual analogue to odor quality. *Psychological Science, 8*, 340–342.

Belli, F., & Sagrillo, F. (2001). *Qual è Takete? Qual è Maluma? La psicolinguistica applicata alla comunicazione pubblicitaria [What is Takete? What is Maluma? Psycholinguistics applied to advertising]* (2nd ed.). Milan: Franco Angeli.

Ben-Artzi, E., & Marks, L. E. (1995). Visual–auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics, 57*, 1151–1162.

Berlin, B. (1994). Evidence for pervasive synthetic sound symbolism in ethnozoological nomenclature. In L. Hinton, J. Nicholls, & J. J. Ohala (Eds.), *Sound symbolism* (pp. 76–93). New York: Cambridge University Press.

Bernstein, I. H., Eason, T. R., & Schurman, D. L. (1971). Hue–tone interaction: A negative result. *Perceptual and Motor Skills, 33*, 1327–1330.

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology, 87*, 241–247.

Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (1994). Exploring the relation between McGurk interference and ventriloquism. *Proceedings of the 1994 International Conference on Spoken Language Processing, 2*, 559–562.

Boernstein, W. S. (1936). On the functional relations of the sense organs to one another and to the organism as a whole. *The Journal of General Psychology, 15*, 117–131.

Boernstein, W. S. (1970). Perceiving and thinking: Their interrelationship and organismic organization. *Annals of the New York Academy of Sciences, 169*, 673–682.

Bond, B., & Stevens, S. S. (1969). Cross-modality matching of brightness to loudness by 5-year-olds. *Perception & Psychophysics, 6*, 337–339.

Boyle, M. W., & Tarte, R. D. (1980). Implications for phonetic symbolism: The relationship between pure tones and geometric figures. *Journal of Psycholinguistic Research, 9*, 535–544.

Bozzi, P., & Flores D'Arcais, G. (1967). Ricerca sperimentale sui rapporti intermodali fra qualità espressive [Experimental research on the intermodal relationships between expressive qualities]. *Archivio di Psicologia, Neurologia e Psichiatria, 28*, 377–420.

Braaten, R. (1993). *Synesthetic correspondence between visual location and auditory pitch in infants*. Paper presented at the 34th Annual Meeting of the Psychonomic Society.

Bresciani, J.-P., Dammeier, F., & Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *Journal of Vision, 6*(5), 2:554–564

Bronner, K. (2011). *What is the sound of citrus? Research on the correspondences between the perception of sound and taste/flavour*.

Bronner, K., Bruhn, H., Hirt, R., & Piper, D. (2008). *Research on the interaction between the perception of music and flavour*. Poster presented at the 9th Annual Meeting of the International Multisensory Research Forum (IMRF), Hamburg, Germany, July.

Calvert, G. A., Spence, C., & Stein, B. E. (Eds.). (2004). *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Carello, C., Anderson, K. L., & Kunkler-Peck, A. J. (1998). Perception of object length by sound. *Psychological Science, 9*, 211–214.

Chen, L., & Zhou, X. (2010). *Audiovisual synesthetic correspondence modulates visual apparent motion*. Poster presented at the 11th International Multisensory Research Forum meeting, Liverpool, U.K., June.

Chen, Y.-C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition, 114*, 389–404.

Chen, Y.-C., & Spence, C. (2011). *Multiple levels of modulation by naturalistic sounds and spoken words on visual picture categorization*. Manuscript submitted for publication.

Clark, H. H., & Brownell, H. H. (1976). Position, direction, and their perceptual integrality. *Perception & Psychophysics, 19*, 328–334.

Cohen, N. E. (1934). Equivalence of brightness across modalities. *The American Journal of Psychology, 46*, 117–119.

Cohen Kadosh, R., & Henik, A. (2007). Can synaesthesia research inform cognitive science? *Trends in Cognitive Sciences, 11*, 177–184.

Cohen Kadosh, R., Henik, A., & Walsh, V. (2009). Synaesthesia: Learned or lost? *Developmental Science, 12*, 484–491.

Collier, G. L. (1996). Affective synaesthesia: Extracting emotion space from simple perceptual stimuli. *Motivation and Emotion, 20*, 1–32.

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities: Pitting abstract against stimulus-specific representations. *Psychological Science, 17*, 905–912.

Cowan, N., & Barron, A. (1987). Cross-modal, auditory–visual Stroop interference and possible implications for speech memory. *Perception & Psychophysics, 41*, 393–401.

Coward, S. W., & Stevens, C. J. (2004). Extracting meaning from sound: Nomic mappings, everyday listening, and perceiving object size from frequency. *Psychological Record, 54*, 349–364.

Cowles, J. T. (1935). An experimental study of the pairing of certain auditory and visual stimuli. *Journal of Experimental Psychology, 18*, 461–469.

Crisinel, A.-S., & Spence, C. (2009). Implicit association between basic tastes and pitch. *Neuroscience Letters, 464*, 39–42.

Crisinel, A.-S., & Spence, C. (2010a). As bitter as a trombone: Synesthetic correspondences in nonsynesthetes between tastes/flavors and musical notes. *Attention, Perception, & Psychophysics, 72*, 1994–2002.

Crisinel, A.-S., & Spence, C. (2010b). A sweet sound? Exploring implicit associations between basic tastes and pitch. *Perception, 39*, 417–425.

Cytowic, R. E. (1993). *The man who tasted shapes*. New York: G. P. Putnam's Sons.

Cytowic, R. E., & Eagleman, D. M. (2009). *Wednesday is indigo blue: Discovering the brain of synesthesia*. Cambridge, MA: MIT Press.

Cytowic, R. E., & Wood, F. B. (1982). Synaesthesia II: Psychophysical relations in the synaesthesia of geometrically shaped taste and colored hearing. *Brain and Cognition, 1*, 36–49.

Davies, N. B., & Halliday, T. L. (1978). Deep croaks and fighting assessment in toads *Bufo bufo*. *Nature, 274*, 683–685.

Davis, R. (1961). The fitness of names to drawings: A cross-cultural study in Tanganyika. *British Journal of Psychology, 52*, 259–268.

Day, S. (2005). Some demographic and socio-cultural aspects of synesthesia. In L. C. Robertson & N. Sagiv (Eds.), *Synesthesia: Perspectives from cognitive neuroscience* (pp. 11–33). New York: Oxford University Press.

Doehrmann, O., & Naumer, M. J. (2008). Semantics and the multisensory brain: How meaning modulates processes of audio-visual integration. *Brain Research, 1242*, 136–150.

Easton, R. D., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics, 32*, 562–570.

Elias, L. J., Saucier, D. M., Hardie, C., & Sarty, G. E. (2003). Dissociating semantic and perceptual components of synaesthesia: Behavioural and functional neuroanatomical investigations. *Cognitive Brain Research, 16*, 232–237.

Ernst, M. O. (2006). A Bayesian view on multimodal cue integration. In G. Knoblich, I. M. Thornton, M. Grosjean, & M. Shiffrar (Eds.), *Human body perception from the inside out* (pp. 105–131). Oxford: Oxford University Press.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision, 7*(5), 7:1–14.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*, 162–169.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*(1), 6:1–12.

Fiebelkorn, I. C., Foxe, J. J., & Molholm, S. (2010). Dual mechanisms for the cross-sensory spread of attention: How much do learned associations matter? *Cerebral Cortex, 20*, 109–120.

Fitch, W. T., & Reby, D. (2001). The descended larynx is not uniquely human. *Proceedings of the Royal Society B, 268*, 1669–1675.

Fox, C. W. (1935). An experimental study of naming. *The American Journal of Psychology, 47*, 545–579.

Frens, M. A., Van Opstal, A. J., & Van der Willigen, R. F. (1995). Spatial and temporal factors determine audio-visual interactions in human saccadic eye movements. *Perception & Psychophysics, 57*, 802–816.

Frings, C., & Spence, C. (2010). Crossmodal congruency effects based on stimulus identity. *Brain Research, 1354*, 113–122.

Gal, D., Wheeler, S. C., & Shiv, B. (2011). *Cross-modal influences on gustatory perception*. Manuscript submitted for publication. Available at http://ssrn.com/abstract=1030197

Gallace, A., Boschin, E., & Spence, C. (in press). On the taste of "Bouba" and "Kiki": An exploration of word–food associations in neurologically normal participants. *Cognitive Neuroscience*.

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics, 68*, 1191–1203.

Gebels, G. (1969). An investigation of phonetic symbolism in different cultures. *Journal of Verbal Learning and Verbal Behavior, 8*, 310–312.

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

Gilbert, A. N., Martin, R., & Kemp, S. E. (1996). Cross-modal correspondence between vision and olfaction: The color of smells. *The American Journal of Psychology, 109*, 335–351.

Grassi, M. (2005). Do we hear size or sound: Balls dropped on plates. *Perception & Psychophysics, 67*, 274–284.

Grassi, M., & Casco, C. (2010). Audiovisual bounce-inducing effect: When sound congruence affects grouping in vision. *Attention, Perception, & Psychophysics, 72*, 378–386.

Green, K., Kuhl, P., Meltzoff, A., & Stevens, E. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics, 50*, 524–536.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Guest, S., Catmur, C., Lloyd, D., & Spence, C. (2002). Audiotactile interactions in roughness perception. *Experimental Brain Research, 146*, 161-171

Guttman, S. E., Gilroy, L. A., & Blake, R. (2005). Hearing what the eyes see: Auditory encoding of visual temporal sequences. *Psychological Science, 16*, 228–235.

Hancock, P. J. B., Baddeley, R. J., & Smith, L. S. (1992). The principal components of natural images. *Network, 3*, 61–70.

Harrington, F. H. (1987). Aggressive howling in wolves. *Animal Behaviour, 35*, 7–12.

Hartshorne, C. (1934). *The philosophy and psychology of sensation*. Chicago: University of Chicago Press.

Hein, G., Doehrmann, O., Müller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *The Journal of Neuroscience, 27*, 7881–7887.

Helbig, H. B., & Ernst, M. O. (2007). Knowledge about a common source can promote visual–haptic integration. *Perception, 36*, 1523–1533.

Hinton, L., Nichols, J., & Ohala, J. J. (Eds.). (1994). *Sound symbolism*. Cambridge: Cambridge University Press.

Holland, M. K., & Wertheimer, M. (1964). Some physiognomic aspects of naming, or maluma, and takete revisited. *Perceptual and Motor Skills, 19*, 111–117.

Holt-Hansen, K. (1968). Taste and pitch. *Perceptual and Motor Skills, 27*, 59–68.

Holt-Hansen, K. (1976). Extraordinary experiences during cross-modal perception. *Perceptual and Motor Skills, 43*, 1023–1027.

Howells, T. (1944). The experimental development of color–tone synesthesia. *Journal of Experimental Psychology, 34*, 87–103.

Hubbard, T. L. (1996). Synesthesia-like mappings of lightness, pitch, and melodic interval. *The American Journal of Psychology, 109*, 219–238.

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition, 109*, 54–65.

Innes-Brown, H., & Crewther, D. (2009). The impact of spatial incongruence on an auditory–visual illusion. *PLoS ONE, 4*, e6450.

Irwin, F. W., & Newland, E. (1940). A genetic study of the naming of visual figures. *The Journal of Psychology, 9*, 3–16.

Jackson, C. V. (1953). Visual factors in auditory localization. *The Quarterly Journal of Experimental Psychology, 5*, 52–65.

Janković, D. (2010). *Evaluative processing is not sensory modality specific*. Poster presented at the 11th International Multsensory Research Forum meeting, Liverpool, U.K., June.

Jespersen, O. (1922). The symbolic value of the vowel *i*. *Philologica, 1*, 1–19.

Jones, J. A., & Jarick, M. (2006). Multisensory integration of speech signals: The relationship between space and time. *Experimental Brain Research, 174*, 588–594.

Jones, J. A., & Munhall, K. G. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics, 25*, 13–19.

Kadosh, R. C., Henik, A., & Walsh, V. (2007). Small is bright and big is dark in synaesthesia. *Current Biology, 17*, R834–R835.

Karwoski, T. F., & Odbert, H. S. (1938). Color–music. *Psychological Monographs, 50*(2, Whole No. 22).

Karwoski, T. F., Odbert, H. S., & Osgood, C. E. (1942). Studies in synesthetic thinking: II. The rôle of form in visual responses to music. *Journal of General Psychology, 26*, 199–222.

Keetels, M., & Vroomen, J. (in press). No effect of synesthetic congruency on temporal ventriloquism. *Attention, Perception, & Psychophysics*.

Kelly, E. L. (1934). An experimental attempt to produce artificial chromaesthesia by the technique of the conditioned response. *Journal of Experimental Psychology, 17*, 315–341.

Kemp, S. E., & Gilbert, A. N. (1997). Odor intensity and color lightness are correlated sensory dimensions. *The American Journal of Psychology, 110*, 35–46.

Kitagawa, N., & Ichihara, S. (2002). Hearing visual motion in depth. *Nature, 416*, 172–174.

Köhler, W. (1929). *Gestalt psychology*. New York: Liveright.

Köhler, W. (1947). *Gestalt psychology: An introduction to new concepts in modern psychology*. New York: Liveright.

Körding, K. P., Beierholm, U., Ma, W. J., Tenenbaum, J. B., Quartz, S., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE, 2*, e943.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus–response compatibility—a model and taxonomy. *Psychological Review, 97*, 253–270.

Kovic, V., Plunkett, K., & Westermann, G. (2009). The shape of words in the brain. *Cognition, 114*, 19–28.

Krantz, D. H. (1972). A theory of magnitude estimation and cross-modality matching. *Journal of Mathematical Psychology, 9*, 168–199.

Külpe, O. (1893). *Grundriss der Psychologie [Fundamentals of psychology]*. Leipzig: Englemann.

Kusnir, F., & Thut, G. (2010, June). *When letters evoke colours: Probing for induction of synaesthetic behaviour in non-synaesthetes through explicit versus implicit grapheme–colour associations*. Poster presented at the 11th International Multisensory Research Forum meeting, Liverpool, U.K.

Kuze, J. (1995). The effect of tone frequency on successive comparison of brightness. *Psychologia, 38*, 50–57.

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research, 158*, 405–414.

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory–visual intensity matching. *Developmental Psychology, 16*, 597–607.

Lindauer, M. S. (1990). The meanings of the physiognomic stimuli taketa and maluma. *Bulletin of the Psychonomic Society, 28*, 47–50.

Lippert, M., Logothetis, N. K., & Kayser, C. (2007). Improvement of visual contrast detection by a simultaneous sound. *Brain Research, 1173*, 102–109.

Long, J. (1977). Contextual assimilation and its effect on the division of attention between nonverbal signals. *The Quarterly Journal of Experimental Psychology, 29*, 397–414.

Lyman, B. (1979). Representation of complex emotional and abstract meanings by simple forms. *Perceptual and Motor Skills, 49*, 839–842.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*, 163–203.

Maeda, F., Kanai, R., & Shimojo, S. (2004). Changing pitch induced visual motion illusion. *Current Biology, 14*, R990–R991.

Makovac, E., & Gerbino, W. (2010). Sound–shape congruency affects the multisensory response enhancement. *Visual Cognition, 18*, 133–137.

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American Journal of Psychology, 87*, 173–188.

Marks, L. E. (1975). On colored-hearing synesthesia: Cross-modal translations of sensory dimensions. *Psychological Bulletin, 82*, 303–331.

Marks, L. (1978). *The unity of the senses: Interrelations among the modalities*. New York: Academic Press.

Marks, L. E. (1984). Synaesthesia and the arts. In W. R. Crozier & A. J. Chapman (Eds.), *Cognitive processes in the perception of art* (pp. 427–447). Amsterdam: Elsevier.

Marks, L. E. (1987a). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 384–394.

Marks, L. E. (1987b). On cross-modal similarity: Perceiving temporal patterns by hearing, touch, and vision. *Perception & Psychophysics, 42*, 250–256.

Marks, L. E. (1989a). For hedgehogs and foxes: Individual differences in the perception of cross-modal similarity. In G. Ljunggren & S. Dornic (Eds.), *Psychophysics in action* (pp. 55–65). Berlin: Springer.

Marks, L. E. (1989b). On cross-modal similarity: The perceptual structure of pitch, loudness, and brightness. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 586–602.

Marks, L. E. (2000). Synesthesia. In E. Cardeña, S. J. Lynn, & S. C. Krippner (Eds.), *Varieties of anomalous experience: Examining the scientific evidence* (pp. 121–149). Washington: American Psychological Association.

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 85–105). Cambridge: MIT Press.

Marks, L. E., Ben-Artzi, E., & Lakatos, S. (2003). Cross-modal interactions in auditory and visual discrimination. *International Journal of Psychophysiology, 50*, 125–145.

Marks, L. E., Hammeal, R. J., & Bornstein, M. H. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development, 52*(1, Whole No. 215), 1–102.

Marks, L. E., Szczesiul, R., & Ohlott, P. (1986). On the cross-modal perception of intensity. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 517–534.

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception, 28*, 903–923.

Martino, G., & Marks, L. E. (2000). Cross-modal interaction between vision and touch: The role of synesthetic correspondence. *Perception, 29*, 745–754.

Martino, G., & Marks, L. E. (2001). Synesthesia: Strong and weak. *Current Directions in Psychological Science, 10*, 61–65.

Maurer, D. (1997). Neonatal synaesthesia: Implications for the processing of speech and faces. In S. Baron-Cohen & J. E. Harrison (Eds.), *Synaesthesia: Classic and contemporary readings* (pp. 224–242). Oxford: Blackwell.

Maurer, D., & Mondloch, C. J. (2005). Neonatal synaesthesia: A reevaluation. In L. C. Robertson & N. Sagiv (Eds.), *Synaesthesia: Perspectives from cognitive neuroscience* (pp. 193–213). Oxford: Oxford University Press.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental Science, 9*, 316–322.

McMahon, T. A., & Bonner, J. T. (1983). *On size and life*. New York: Scientific American.

Melara, R. D. (1989a). Dimensional interaction between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 69–79.

Melara, R. D. (1989b). Similarity relations among synesthetic stimuli and their attributes. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 212–231.

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General, 116*, 323–336.

Meltzoff, A. N., & Borton, R. W. (1979). Intermodal matching by human neonates. *Nature, 282*, 403–404.

Mesz, B., Trevisan, M., & Sigman, M. (2011). *The taste of music*. Manuscript submitted for publication.

Miller, A., Werner, H., & Wapner, S. (1958). Studies in physiognomic perception: V. *Effect of ascending and descending gliding tones on autokinetic motion. Journal of Psychology, 46*, 101–105.

Miller, E. A. (1972). Interaction of vision and touch in conflict and nonconflict form perception tasks. *Journal of Experimental Psychology, 96*, 114–123.

Miller, J. O. (1991). Channel interaction and the redundant targets effect in bimodal divided attention. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 160–169.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual–auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex, 14*, 452–465.

Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch–object correspondences in your children. *Cognitive, Affective & Behavioral Neuroscience, 4*, 133–136.

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: Examining temporal ventriloquism. *Cognitive Brain Research, 17*, 154–163.

Morgan, G. A., Goodson, F. E., & Jones, T. (1975). Age differences in the associations between felt temperatures and color choices. *The American Journal of Psychology, 88*, 125–130.

Morgan, M. L., DeAngelis, G. C., & Angelaki, D. E. (2008). Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron, 59*, 662–673.

Morton, E. S. (1994). Sound symbolism and its role in non-human vertebrate communication. In L. Hinton, J. Nicholls, & J. J. Ohala (Eds.), *Sound symbolism* (pp. 348–365). Cambridge: Cambridge University Press.

Mudd, S. A. (1963). Spatial stereotypes of four dimensions of pure tone. *Journal of Experimental Psychology, 66*, 347–352.

Murray, M. M., Michel, C. M., Grave de Peralta, R., Ortigue, S., Brunet, D., Gonzalez Andino, S., & Schnider, A. (2004). Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *Neuroimage, 21*, 125–135.

Nahm, F. K. D., Tranel, D., Damasio, H., & Damasio, A. R. (1993). Cross-modal associations and the human amygdale. *Neuropsychologia, 31*, 727–744.

Naumer, M. J., & Kaiser, J. (Eds.). (2010). *Multisensory object perception in the primate brain*. New York: Springer.

Newman, S. S. (1933). Further experiments in phonetic symbolism. *The American Journal of Psychology, 45*, 53–75.

Nuckolls, J. (2003). The case for sound symbolism. *Annual Review of Anthropology, 28*, 225–252.

Oberman, L. M., & Ramachandran, V. S. (2008). Preliminary evidence for deficits in multisensory integration in autism spectrum disorders: The mirror neuron hypothesis. *Social Neuroscience, 3*, 348–355.

O'Leary, A., & Rhodes, G. (1984). Cross-modal effects on visual and auditory object perception. *Perception & Psychophysics, 35*, 565–569.

O'Mahony, M. (1983). Gustatory responses to nongustatory stimuli. *Perception, 12*, 627–633.

Osgood, C. E. (1960). The cross-cultural generality of visual–verbal synesthetic tendencies. *Behavioral Science, 5*, 146–169.

Osgood, C., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Pan, J.-Y., Yang, H.-J., Faloutsos, C., & Duygulu, P. (2004). Automatic multimedia crossmodal correlation discovery. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 653–658). Seattle: ACM Press.

Parise, C. V., & Pavani, F. (2011). *I see it large, I say it louder. Evidence of sound symbolism in simple vocalizations*. Manuscript submitted for publication.

Parise, C., & Spence, C. (2008a). Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters, 442*, 257–261.

Parise, C., & Spence, C. (2008b). *Synesthetic correspondence modulates audiovisual temporal integration*. Poster presented at the 9th IMRF meeting, Hamburg, Germany, July.

Parise, C., & Spence, C. (2009). "When birds of a feather flock together": Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE, 4*, e5664.

Patching, G. R., & Quinlan, P. T. (2002). Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 755–775.

Piesse, C. H. (1891). *Piesse's art of perfumery* (5th ed.). London: Piesse and Lubin.

Plato. (1961). In E. Hamilton & H. Cairns (Eds.), *The collected dialogues*. New York: Pantheon.

Poffenberger, A. T., & Barrows, B. E. (1924). The feeling value of lines. *The Journal of Applied Psychology, 8*, 187–205.

Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology, 13*, 278–285.

Radeau, M., & Bertelson, P. (1987). Auditory–visual interaction and the timing of inputs. Thomas (1941) revisited. *Psychological Research, 49*, 17–22.

Rader, C. M., & Tellegen, A. (1987). An investigation of synaesthesia. *Journal of Personality and Social Psychology, 52*, 981–987.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia—A window into perception, thought and language. *Journal of Consciousness Studies, 8*, 3–34.

Ramachandran, V. S., & Hubbard, E. M. (2003, May). Hearing colors, tasting shapes. *Scientific American, 288*, 43–49.

Ramachandran, V. S., & Oberman, L. M. (2006, May 12). Broken mirrors. *Scientific American, 295*, 62–69.

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology, 89*, 1078–1093.

Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: A strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society B, 273*, 2159–2168.

Roffler, S. K., & Butler, R. A. (1968). Factors that influence the localization of sound in the vertical plane. *The Journal of the Acoustical Society of America, 43*, 1255–1259.

Rogers, S. K., & Ross, A. S. (1975). A cross-cultural test of the Maluma–Takete phenomenon. *Perception, 4*, 105–106.

Root, R. T., & Ross, S. (1965). Further validation of subjective scales for loudness and brightness by means of cross-modality matching. *The American Journal of Psychology, 78*, 285–289.

Rouw, R., & Scholte, H. S. (2007). Increased structural connectivity in grapheme–color synesthesia. *Nature Neuroscience, 10*, 792–797.

Rudmin, F., & Cappelli, M. (1983). Tone–taste synesthesia: A replication. *Perceptual and Motor Skills, 56*, 118.

Sagiv, N., & Ward, J. (2006). Cross-modal interactions: Lessons from synesthesia. *Progress in Brain Research, 155*, 263–275.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12*, 225–239.

Seo, H.-S., Arshamian, A., Schemmer, K., Scheer, I., Sander, T., Ritter, G., et al. (2010). Cross-modal integration between odors and abstract symbols. *Neuroscience Letters, 478*, 175–178.

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences, 14*, 425–432.

Shams, L., Ma, W. J., & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *NeuroReport, 16*, 1923–1927.

Shepard, R. N. (1994). Perceptual–cognitive universals as reflections of the world. *Psychonomic Bulletin & Review, 1*, 2–28.

Shore, D. I., Barnes, M. E., & Spence, C. (2006). The temporal evolution of the crossmodal congruency effect. *Neuroscience Letters, 392*, 96–100.

Simner, J., Cuskley, C., & Kirby, S. (2010). What sound does that taste? Cross-modal mapping across gustation and audition. *Perception, 39*, 553–569.

Simner, J., & Ludwig, V. (2009). *What colour does that feel? Cross-modal correspondences from touch to colour*. Paper presented at the Third International Conference of Synaesthesia and Art, Granada, Spain, April.

Simner, J., Ward, J., Lanz, M., Jansari, A., Noonan, K., Glover, L., Oakley, D. A. (2005). Nonrandom associations of graphemes to colours in synaesthetic and non-synaesthetic populations. *Cognitive Neuropsychology, 22*, 1069–1085.

Simpson, R. H., Quinn, M., & Ausubel, D. P. (1956). Synaesthesia in children: Association of colors with pure tone frequencies. *The Journal of Genetic Psychology, 89*, 95–103.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *NeuroReport, 12*, 7–10.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *The Journal of the Acoustical Society of America, 117*, 305–318.

Smith, E. L., Grabowecky, M., & Suzuki, S. (2007). Auditory–visual crossmodal integration in perception of face gender. *Current Biology, 17*, 1680–1685.

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology, 24*, 99–142.

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science & Technology, 28*, 61–70.

Spence, C. (2010). The color of wine—Part 1. *The World of Fine Wine, 28*, 122–129.

Spence, C., & Driver, J. (Eds.). (2004). *Crossmodal space and crossmodal attention*. Oxford: Oxford University Press.

Spence, C., & Gallace, A. (in press). *Tasting shapes and words. Food Quality and Preference*.

Spence, C., Levitan, C. A., Shankar, M. U., & Zampini, M. (2010). Does food color influence taste and flavor perception in humans? *Chemosensory Perception, 3*, 68–84.

Spence, C., & Ngo, M. (in press). Does attention or multisensory integration explain the crossmodal facilitation of masked visual target identification? In B. E. Stein et al. (Eds.), *The new handbook of multisensory processing*. Cambridge, MA: MIT Press.

Spence, C., Ngo, M. K., Lee, J.-H., & Tan, H. (2010). Solving the correspondence problem in haptic/multisensory interface design. In M. H. Zadeh (Ed.), *Advances in haptics* (pp. 47–74). InTech. Available at www.sciyo.com/articles/show/title/solving-the-correspondence-problem-in-haptic-multisensory-interface-design

Spence, C., & Zampini, M. (2006). Auditory contributions to multisensory product perception. *Acta Acustica united with Acustica, 92*, 1009–1025.

Stevens, J. C., & Marks, L. E. (1965). Cross-modality matching of brightness and loudness. *Proceedings of the National Academy of Sciences, 54*, 407–411.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*, 153–181.

Stevens, S. S. (1971). Issues in psychophysical measurement. *Psychological Review, 78*, 426–450

Stoffregen, T. A., & Bardy, B. G. (2001). On specification and the senses. *The Behavioral and Brain Sciences, 24*, 195–261.

Stroop, J. R. (1935). Studies of interference in serial-verbal reaction. *Journal of Experimental Psychology, 18*, 643–662.

Stumpf, K. (1883). *Tonpsychologie I [Psychology of the tone]*. Leipzig: Hirzel.

Taylor, I. K. (1963). Phonetic symbolism re-examined. *Psychological Bulletin, 60*, 200–209.

Taylor, I. K., & Taylor, M. M. (1962). Phonetic symbolism in four unrelated languages. *Canadian Journal of Psychology, 16*, 344–356.

Teramoto, W., Hidaka, S., & Sugita, Y. (2010). Sounds move a static visual object. *PLoS ONE, 5*, e12255.

Thomas, G. J. (1941). Experimental study of the influence of vision on sound localization. *Journal of Experimental Psychology, 28*, 163–177.

Uznadze, D. (1924). Ein experimenteller Beitrag zum Problem der psychologischen Grundlagen der Namengebung [An experimental contribution to the problem of the psychological bases of naming]. *Psychologische Forschung, 5*, 25–43.

van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron, 43*, 271–282.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory–visual speech perception. *Neuropsychologia, 45*, 598–607.

Vatakis, A., Ghazanfar, A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity assumption": Is speech special? *Journal of Vision, 8*(9), 14:1–11.

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics, 69*, 744–756.

Vatakis, A., & Spence, C. (2008). Evaluating the influence of the "unity assumption" on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica, 127*, 12–23.

von Békésy, G. (1959). Similarities between hearing and skin sensations. *Psychological Review, 66*, 1–22.

Von Hornbostel, E. M. (1931). Über Geruchshelligkeit [On odour/smell brightness]. *Pflügers Archiv für Gesamte Physiologie, 227*, 517–538.

Von Hornbostel, E. M. (1927/1950). The unity of the senses. In W. D. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 210-216). London: Routledge and Kegan Paul

von Kriegstein, K., & Giraud, A.-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology, 4*, e326.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17*, 367–376.

Vroomen, J., & Keetels, M. (2006). The spatial constraint in intersensory pairing: No role in temporal ventriloquism. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 1063–1071.

Wagner, K., & Dobkins, K. (2009). Shape–color synaesthesia in the first year of life: A normal stage of visual development? [Abstract]. *Journal of Vision, 9*(8), 699a.

Wagner, S., Winner, E., Cicchetti, D., & Gardner, H. (1981). "Metaphorical" mapping in human infants. *Child Development, 52*, 728–731.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., Johnson, S.P. (2010). Preverbal infants' sensitivity to synesthetic cross-modality correspondences. *Psychological Science, 21*, 21–25.

Walker, P., Francis, B. J., & Walker, L. (in press). The brightness–weight illusion: Darker objects look heavier but feel lighter. *Experimental Psychology*.

Walker, P., & Smith, S. (1984). Stroop interference based on the synaesthetic qualities of auditory pitch. *Perception, 13*, 75–81.

Walker, P., & Smith, S. (1985). Stroop interference based on the multimodal correlates of haptic size and auditory pitch. *Perception, 14*, 729–736.

Walker, R. (1987). The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. *Perception & Psychophysics, 42*, 491–502.

Walker, S., Bruce, V., & O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics, 57,* 1124–1133.

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research, 158,* 252–258.

Walsh, V. (2003). A theory of magnitude: Common cortical metrices of time, space and quality. *Trends in Cognitive Sciences, 7,* 483–488.

Ward, J., Huckstep, B., & Tsakanikos, E. (2006). Sound–colour synaesthesia: To what extent does it use cross-modal mechanisms common to us all? *Cortex, 42,* 264–280.

Welch, R. B. (1972). The effect of experienced limb identity upon adaptation to simulated displacement of the visual field. *Perception & Psychophysics, 12,* 453–456.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 3,* 638–667.

Wertheimer, M. (1958). The relation between the sound of a word and its meaning. *The American Journal of Psychology, 71,* 412–415.

Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language, 93,* 10–19.

Wicker, F. W. (1968). Mapping the intersensory regions of perceptual space. *The American Journal of Psychology, 81,* 178–188.

Widmann, A., Kujala, T., Tervaniemi, M., Kujala, A., & Schröger, E. (2004). From symbols to sounds: Visual symbolic information activates sound representations. *Psychophysiology, 41,* 709–715.

Yau, J. M., Olenczak, J. B., Dammann, J. F., & Bensmaia, S. J. (2009). Temporal frequency channels are linked across audition and touch. *Current Biology, 19,* 561–566.

Zangenehpour, S., & Zatorre, R. J. (2010). Cross-modal recruitment of primary visual cortex following brief exposure to bimodal audiovisual stimuli. *Neuropsychologia, 48,* 591–600.