Signal compatibility as a modulatory factor for audiovisual multisensory integration

Cesare Valerio Parise

St. Cross College, University of Oxford

Dissertation submitted for the degree of Doctor of Philosophy at the University of Oxford

Declaration

I declare that this work has not been submitted previously as an exercise for a degree at this or any other university, and that it is entirely my own work.

_____

Cesare Valerio Parise

Acknowledgments

I would like to express my deepest gratefulness to all the people who made this possible.

First of all, I would like to express my heartfelt gratitude to Charles Spence and Marc Ernst, outstanding scientists and superb mentors. Working with them has always been an honor for me, I could not be prouder of my academic roots.

I would also like to thank my formers scientific mentors: Stefano Vezzani, Klaus Oberauer, Daniele Zavagno, Emanuela Bricolo, Gabriel Baud-Bovy, and Francesco Pavani, for inspiring me with their passion for science; A better guidance would not have been possible.

A special thanks goes to my examiners, Paul Azzopardi, Georg Meyer and Oliver Braddick, for their fairness and for the enjoyable discussions that we had during both viva voce exams. Surely my thesis is a better piece of work after their contribution. I am also deeply grateful to Miles Hewstone, for having done so much to make this possible.

A big thanks goes to Fagiano, Fiore, Sara, Scu, Siluro, Stiv, Vale (let Glen Benton be with you!), André, David, Davide, Elena, Francesca, Giorgiana, Gustavo, Jess, Katya, Loes, Marieke, Matteo, Mario, Mark, Max D.L., Max H., Paniz, Sara, Vanessa, Verena, Vespa, and Will; Friends, colleagues, drinking fellows and cheerful companions in the twisted path of my DPhil.

I also owe a huge debt of gratitude to my amazing family, and in particular to my grandmothers Maria and Angela (you rock!), to my aunts and uncles, and especially to Rina, Lia, Elio and Tiziano, and to my four legged brother Igor.

I would like to express my deepest gratitude to Irene, for always being so close to me in spite of being so far away. You are the best thing that happened to my life. We'll be together soon!

Finally, I would like to dedicate this work to my parents, Graziella and Giovanni, for their endless and unconditional support. Without them I would have never managed to get my DPhil. Thanks.

Ringraziamenti

Vorrei esprimere la mia più profonda gratitudine alle persone che hanno reso possibile tutto questo.

Anzitutto, vorrei dare il mio più caloroso ringraziamento a Charles Spence e Marc Ernst, illustri scienziati e straordinari mentori. Lavorare con loro è sempre stato un onore per me, non potrei essere più fiero delle mie radici accademiche.

Desidero inoltre ringraziare i miei precedenti mentori scientifici: Gabriel Baud-Bovy, Emanuela Bricolo, Klaus Oberauer, Francesco Pavani, Stefano Vezzani, Daniele Zavagno per avermi ispirato con la loro passione per le scienze. Una guida migliore non sarebbe stata possibile.

Un ringraziamento speciale va ai miei esaminatori, Paul Azzopardi, Georg Meyer ed Oliver Braddick, per la loro equità e per i piacevoli scambi di idee avuti durante gli esami orali. Grazie al loro contributo la mia tesi è sicuramente un lavoro migliore. Sono anche profondamente grato a Miles Hewstone, per aver fatto così tanto per rendere tutto questo possibile.

Un grosso grazie va a Fagiano, Fiore, Sara, Scu, Siluro, Stiv, Vale (che Glen Benton sia con voi!), André, David, Davide, Elena, Francesca, Giorgiana, Gustavo, Jess, Katya, Loes, Marieke, Matteo, Mario, Mark, Max D.L., Max H., Paniz, Sara, Vanessa, Verena, Vespa, and Will; Amici, colleghi ed allegri compagni di bevute, nel tortuoso sentiero del mio dottorato.

Un enorme ringraziamento va inoltre alla mia splendida famiglia ed in particolare alle mie nonne Maria e Angela (siete grandi!), ai miei zii ed alle mie zie, in particolare Rina, Lia, Elio e Tiziano e ad Igor, il mio fratello a quattro zampe.

Vorrei esprimere la mia più profonda gratitudine ad Irene, per essere sempre così vicina nonostante sia così lontana. Sei la cosa migliore che mi sia capitata. Presto saremo insieme!

Vorrei infine dedicare questo lavoro ai miei genitori, Graziella e Giovanni, per il loro infinito e incondizionato supporto. Senza di loro non ce l'avrei mai fatta a conseguire questo dottorato. Grazie.

Contents

Publications resulting from the present work

Parise, C.V., & Spence, C. (2008) Synesthetic associations modulate the temporal ventriloquism effect. *Neuroscience Letters.* 442, 257-261

Parise, C.V., & Spence, C. (2009) 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE* 4(5):e5664

Parise, C.V., Spence, C., & Ernst, M.O. (2012) Multisensory integration: When correlation implies causation. *Current Biology. 22, 46-49*

Parise, C.V., & Spence, C. (under review) Audiovisual crossmodal correspondences. J. Simner & E. Hubbard (Eds.), *Oxford handbook of synaesthesia.* Oxford, UK: Oxford University Press

Parise, C.V., & Spence, C. (under review) Audiovisual crossmodal correspondences and sound symbolism: An IAT study. *Experimental Brain Research.*

Presentations resulting from the present work

Parise, C.V., & Spence, C. (2008) *Synaesthetic links in audiovisual temporal integration* – Poster presented at the International Multisensory Research Forum, 16-19 July, Hamburg (Germany)

Parise, C.V., & Spence, C. (2009) *Synesthetic constraints on multisensory integration* – Poster presented at the Symposium for MEG inauguration at CIMeC, 27 May, Trento (Italy)

Parise, C.V., & Spence, C. (2009) *'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes.* – Oral presentation at the International Multisensory Research Forum, 30 June - 2 July, New York (USA)

Spence, C., Navarra, J., Vatakis, A., Hartcher-O'Brien, J., & Parise, C.V. (2009) *The multisensory perception of synchrony.* - Oral presentation at the European Conference on Visual Perception, 24-28 August, Regensburg (Germany)

Parise, C.V., & Spence, C. (2009) *'Quando chi si somiglia si piglia': corrispondenze sinestetiche modulano l'integrazione audiovisiva in soggetti non-sinesteti ['When Birds of a Feather Flock Together': Synesthetic Correspondences Modulate Audiovisual Integration in Non-Synesthetes]* - Oral presentation at the annual meeting Congresso Nazionale della Sezione Sperimentale dell'Associazione Italiana di

Psicologia [National Congress of the Experimental Section of the Italian Association of Psychology], 24-26 September, Chieti (Italy)

Parise, C.V., Harrar, V., Spence, C., & Ernst, M.O. (2011) *Multisensory integration: When correlation implies causation* – Poster presented at the European Conference on Visual Perception, 28 August – 1 September, Toulouse (France)

Parise, C.V., Harrar, V., Spence, C., & Ernst, M.O. (2011) *Multisensory integration: When correlation implies causation* – Poster presented at the Bernstein symposium on multisensory perception and action – 28-29 September, Tübingen (Germany)

Parise, C.V., Harrar, V., Spence, C., & Ernst, M.O. (2011) *Multisensory integration: When correlation implies causation* – Oral presentation at the International Multisensory Research Forum, 17-20 October, Fukuoka (Japan)

Spence, C., Parise, C.V., & Deroy, O. (2011) *Crossmodal correspondences* – Oral presentation at the International Multisensory Research Forum, 17-20 October, Fukuoka (Japan)

Other publications

Spence, C., & Parise, C.V. (2010) Prior entry: A review. *Consciousness & Cognition* 19, 364-79

Spence, C., Parise, C.V., & Chen, Y-C. (2011) The colavita visual dominance effect. In M. Wallace & M. Murray (Eds.), *The neural basis of multisensory processes*. London, UK: Taylor and Francis Group

Parise, C., & Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research. 214*(3), 373-380

Parise, C.V., & Spence, C. (2012) Assessing the associations between brand packaging and brand attributes using an indirect performance measure. *Food Quality and Preference.* 24, 17-23

Other presentations

Spence, C., Parise, C.V., & Chen, Y. (2009) *Explaining the Colavita visual dominance effect* - Oral presentation at the International Multisensory Research Forum, 30 June - 2 July, New York (USA)

Spence, C., Navarra, J., Vatakis, A., Hartcher-O'Brien, J., & Parise, C.V. (2009) *The multisensory perception of synchrony.* - Oral presentation at the European Conference on Visual Perception, 24-28 August, Regensburg (Germany)

Parise, C.V., Stewart, N., Foecker, J., Ngo, M., Browning, M., Roeder, B., Spence, C., Rogers R.D. (2010) *Emotional information enhances audiovisual speech integration* – Poster presented at the International Multisensory Research Forum, 16-19 July, Liverpool (UK)

Parise, C.V., Di Luca, M., & Ernst, M. (2010) *Multiple criteria for multisensory stimuli* – Poster presented at the International Multisensory Research Forum, 16-19 July, Liverpool (UK)

Abstract

The physical properties of the distal stimuli activating our senses are often correlated in nature; it would therefore be advantageous to exploit such correlations to better process sensory information. Stimulus correlations can be contingent and readily available to the senses (like the temporal correlation between mouth movements and vocal sounds in speech), or can be the results of the statistical co-occurrence of certain stimulus properties that can be learnt over time (like the relation between the frequency of acoustic resonance and the size of the resonator). Over the last century, a large body of research on multisensory processing has demonstrated the existence of compatibility effects between individual features of stimuli from different sensory modalities. Such compatibility effects, termed crossmodal correspondences, possibly reflect the internalization of the natural correlation between stimulus properties.

The present dissertation assesses the effects of crossmodal correspondences on multisensory processing and reports a series of experiments demonstrating that crossmodal correspondences influence the processing rate of sensory information, distort perceptual experiences and

lead to stronger multisensory integration. Moreover, a final experiment investigating the effects of contingent signals' correlation on multisensory processing demonstrates the key role of temporal correlation in inferring whether two signals have a common physical cause or not (i.e., the correspondence problem). A Bayesian framework is proposed to interpret the present results whereby stimulus correlations, represented on the prior distribution of expected crossmodal co-occurrence, operate as cues to solve the correspondence problem.

Extended abstract

In this thesis, the effects of crossmodal correspondences and temporal correlation on the processing of multisensory signals are discussed, and their role in solving the correspondence problem are empirically demonstrated with a series of experiments. After a brief introduction to the topic of multisensory processing and crossmodal correspondence in Chapter 1, in Chapter 2, the literature on crossmodal correspondences is critically reviewed. Based on a large body or research on crossmodal correspondences accumulated over more than a century, an inventory of the defining features of crossmodal correspondence is provided. Next, a taxonomy of crossmodal correspondence is developed. Finally the literature on the effects of audiovisual correspondence on human information processing is reviewed.

In Chapter 3, evidence for the effect of crossmodal correspondences on human behavior is presented. A number of well-known examples of crossmodal correspondence, including the Mil-Mal effect, the Takete-Maluma effect, and the correspondence between auditory pitch and visual size are investigated using a modified version of the Implicit Association Test (IAT).

Moreover, evidence is provided for two new crossmodal correspondences, namely the association between pitch and size of angles, and between the waveform of auditory signals and the roundedness of visual shapes.

In Chapter 4, evidence is presented that crossmodal correspondences operate on a perceptual level, and systematically distort perceptual experiences. Human observers sometimes find it easier to judge the temporal order in which two visual stimuli have been presented if one sound is presented before the first visual stimulus and a second sound is presented after the second visual stimulus. This phenomenon has been term temporal ventriloquism. A manipulation of the crossmodal congruency between the visual and the auditory stimuli revealed a systematic modulation of the magnitude of this perceptual effect: Temporal sensitivity was higher for pairs of congruent auditory and visual stimuli than for incongruent pairs of stimuli. These results therefore provide the first empirical evidence that crossmodal correspondences operate on a perceptual level, and systematically distort perceptual experiences.

In Chapter 5, a series of experiments showing that crossmodal correspondences modulate multisensory integration are described. Observers

were presented with pairs of asynchronous or spatially discrepant visual and auditory stimuli that were either crossmodally congruent or incongruent, and had to report the relative temporal order of presentation or the relative spatial locations of the two stimuli. Sensitivity to spatial and temporal offsets between auditory and visual stimuli was lower for pairs of congruent as compared to incongruent audiovisual stimuli. Recent studies of multisensory integration have demonstrated that reduced sensitivity to perceptual estimates regarding intersensory conflicts constitutes the marker of a stronger coupling between unisensory signals. These results therefore indicate a stronger coupling of congruent vs. incongruent stimuli and provide the first psychophysical evidence that crossmodal correspondences promote multisensory integration.

In Chapter 6, an experiment investigating the role of the similarity of the temporal structure of visual and auditory signals for multisensory integration is presented. Inferring which signals have a common underlying cause, and hence should be integrated, (i.e., solving the correspondence problem), is a primary challenge for a perceptual system dealing with multiple sensory inputs. Here the role of correlation between the temporal structures of auditory and visual signals in causal inference is explored.

Specifically, it is tested whether correlated signals are inferred to originate from the same event and hence integrated optimally. In a pointing task with visual, auditory, and combined audiovisual targets, the improvement in precision for combined relative to unimodal targets was statistically optimal only when the audiovisual signals were correlated. These results therefore demonstrate that humans use the similarity in the temporal structure of multiple sensory signals to solve the crossmodal correspondence problem, hence inferring causation from correlation.

1. Introduction

Humans and other animals are equipped with multiple sensory channels with which to perceive the environment that surrounds them. Multiple sources of information concerning the state of the external world and our body can provide us with richer, more robust, and more precise information, ultimately allowing for more adaptive behavior. Think of a dog barking nervously behind a picket fence. Visual and auditory cues both provide valuable information about where the dog is located when it barks. The spatial and temporal properties of the environment (and of the events taking place in it) can often be redundantly sensed, though with different levels of precision, via multiple sensory channels, and hence are typically considered as being amodal stimulus properties (see Bahrick, Lickliter, &

Flom, 2004; Green & Angelaki, 2010). Multiple senses, however, also provide complementary (i.e., non-redundant) information. The colour of the barking dog in this example can only be perceived visually, while the pitch of the dog's growl can only be sensed auditorily. Being the perceptual correlates of different physical properties, colour and pitch can therefore be considered as complementary, or modal, cues (see Green & Angelaki, 2010; Bahrick, et al., 2004).

Over the course of the last decade, there has been a resurgence of research interest in studying the mechanisms by which our brains integrate redundant cues from multiple sensory channels (e.g., see Meyer & Noppeney, 2011; Murray & Wallace, 2011; Parise & Spence, under review; Spence, 2011). When jointly estimating the same physical property, different senses normally provide highly correlated sensory cues. A growing body of empirical research now demonstrates that the integration of redundant cues about the same physical properties allows humans and other animals to generate more precise and robust combined sensory estimates (Ernst & Bülthoff, 2004; Trommershäuser, Landy, & Körding, 2011). On

the other hand, having multiple complementary cues tuned to different properties of the environment also provides non-redundant information about the environment, and accounts for the richness of our sensory experiences. It should be noted that sometimes also complementary stimulus properties are correlated in nature: the size of an object is indeed correlated to its weight (e.g., see Cole, 2008; Ellis, & Lederman, 1993; Flanagan, & Beltzner 2000; Flanagan, Bittner, & Johansson, 2008; Johansson, & Westling, 1988)!

Dealing with multiple sensory inputs, however, presents the organism with an important challenge: inferring which signals have a common source and which not. This is the so-called *correspondence problem,* or causal inference (Ernst & Bülthoff, 2004; Shams & Beierholm, 2010; Welch & Warren, 1980). Correctly inferring whether signals have a common physical origin, and hence whether or not they provide redundant information, is indeed a precondition for multisensory integration: Integrating unrelated sensory signals would in fact lead to erroneous interpretations of the world, with potentially harmful consequences. An extensive literature has shown

the importance of 'bottom-up' and 'top-down' factors for multisensory integration in terms of solving the correspondence problem (see Spence, 2007, for a review). With regard to bottom-up factors, researchers have mainly focused on temporal or spatial coincidence, hence showing, for example, that integration breaks down when large temporal delays (Bresciani et al., 2005; Slutsky & Recanzone, 2001) and large spatial offsets are introduced between multiple unisensory signals (Bermant & Welch, 1976; Jackson, 1953; Körding et al., 2007; Meyer, Wuerger, Röhrbein, & Zetzsche, 2005). Moreover, it has been proposed that other finer grained bottom-up features might play a role in helping to solve the correspondence problem. One such feature that has often been postulated is the temporal correlation between multiple unisensory signals (Radeau & Bertelson, 1987; Spence, 2007): if two signals have the same temporal structure, it is quite likely that they also have the same physical cause. Returning to the example of the barking dog, if the temporal structure of the barks that we hear and the movement of the dog (especially of the dog's mouth) that we see are not correlated, then it is quite likely that there's another dog in

addition to the one that we see, and that other dog might actually be the one that is barking.

On top of contingent stimulus properties, higher-order factors are also known to modulate multisensory integration. This is the case for semantic congruency (Chen & Spence, 2010; Doehrmann & Naumer, 2008; Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004), whereby the causal inference is guided by previous knowledge about which signals normally go together (e.g., cats do not bark). Some of the studies on crossmodal perception that have dealt with this issue have used complex (i.e., not blobs, Gabors, beeps, flashes, etc.) and/or ecological stimuli, often involving audiovisual speech (Campanella & Belin, 2007; Vatakis, Ghazanfar, & Spence, 2008; Vatakis & Spence, 2007a), or naturalistic sounds (Chen & Spence, 2010, 2011). A common finding is that semantically congruent crossmodal stimuli are more likely integrated than semantically mismatching stimuli.

However, there are other examples of more subtle compatibility effects involving the individual physical features of elementary stimuli. This is the case of size (perceived either visually and/or haptically) and the pitch of

auditory stimuli. It has been demonstrated that larger objects are normally associated by human and dog observers with lower pitched sounds, and vice versa (Faragó et al., 2010; Gallace & Spence, 2006, Taylor, Reby & McComb, 2010). As mentioned already, complementary sensory cues are often correlated, and this is surely the case of pitch and size, two properties that normally correlate in nature due to the laws of acoustic resonance. Therefore, the existence of compatibility effects between elementary features of multisensory signals, termed crossmodal correspondences, might reflect an adaptive response of human perceptual system that learns the natural statistical correlations that exist between multisensory stimulus features. Let us now return to our original example, if the nervous dog barking behind the fence happens to be a small Chihuahua, it is very unlikely that his (or her) growls will be deep (i.e., low-pitched; see Faragó et al., 2010, Taylor et al., 2010)!

2. On crossmodal correspondences

2.1. Introduction

For more than a century now, scientists have acknowledged the existence of seemingly arbitrary compatibility effects between crossmodal complementary cues in the normal population (Külpe, 1893; Spence, 2011; Stumpf, 1883). As mentioned in Chapter 1, most people, for example, readily associate larger objects with low-pitched sounds, while considering it less natural to match them with shrill sounds (Gallace & Spence, 2006; P. Walker & Smith, 1984, 1985; R. Walker, 1987). Likewise, most observers find it more natural to match loud sounds to bright rather than dim lights

(Marks, 1987a). Such crossmodal correspondences, possibly reflecting the learnt associations between the features of naturally-occurring multisensory stimuli, constitute a pervasive aspect of multisensory perception.

In the present chapter, the role of crossmodal correspondences between complementary cues in multisensory perception is analyzed. The discussion will focus on those correspondences that exist between auditory and visual stimuli, since this is the modality pairing that has attracted the majority of research interest to date. First, a general introduction to the topic of crossmodal correspondences is provided. Next, their role in the processing of crossmodal signals is scrutinized.

2.2. A definition of crossmodal correspondences

Crossmodal correspondences can be defined as congruency effects between stimuli presented in different sensory modalities that result from the 'expected' (i.e. a priori) mapping between those sensory cues. In the simplest case of redundant cues, when two or more sensory channels are simultaneously sampling the same physical property, each modality should

provide virtually the same estimate. That is, it is sufficient to sample a physical property (e.g. spatial location) of a distal stimulus via one sensory modality (e.g. audition) in order to know what to expect if the same physical property was to be sampled with another sensory modality (e.g., vision; Alais & Burr, 2004; Wuerger, Meyer, Hofbauer, Zetzsche, & Schill, 2010). Conversely, when multiple senses provide complementary information, the mapping between multiple sensory estimates is highly uncertain (i.e., low- or no correlation), and it is often impossible, given a particular sensory estimate in one modality, to infer a complementary property in another. Nevertheless, even in the case of complementary cues, the mapping is sometimes not completely uncertain (or arbitrary), and it is still possible to observe *relative* compatibility effects. This is the case for auditory pitch and elevation: Although in nature there is no deterministic mapping between auditory pitch and vertical position, it is likely, given two sounds having different pitches and two visual stimuli with different elevations, that the sound with the higher-pitch would likely be mapped to the higher elevation and the lower sound to the lower elevation (Bernstein & Edelstein, 1971).

In the remainder of this chapter, the focus will be on those crossmodal correspondences between complementary polar features of stimuli presented in different sensory modalities that are shared by most (if not all) individuals. Four criteria constitute the core features of the crossmodal correspondences under discussion: *Complementarity*, *polarity*, *universality*, and *relativity*[1]. As mentioned earlier, crossmodal correspondences often involve complementary (i.e., non-redundant) sensory cues such as pitch and elevation, pitch and size, lightness and loudness, etc… Such features may themselves be either modal (i.e., modality specific) or amodal (i.e., modality independent).

Polar dimensions, often referred to as prothetic dimensions (see S.S. Stevens & J.C. Stevens, 1975), are those sensory dimensions along which stimuli can be experienced as one being 'more than' or 'less than' another[2].

[1] While, at first glance, the terms universality and relativity might appear contradictory, it is important to note that they refer to different things. Universality speaks to the commonality of the phenomenon across the population. By contrast, relativity refers to how congruency effect operates between the stimuli being compared.

[2] Examples of polar dimensions include: Loudness, lightness, saturation, pitch, elevation, duration, size, mass, temperature...

In the case of crossmodal correspondences between complementary polar dimensions, a pole of the first dimension is compatible with one pole of the second dimension and incompatible with the other. This is, for example, the case of the above-mentioned crossmodal correspondence between pitch and elevation (Bernstein & Edelstein, 1971; Cabrera & Morimoto, 2007; Rusconi, Kwan, Giordano, Umilta, & Butterworth, 2006; Stumpf, 1883).

Far from being idiosyncratic, many congruency effects are universal with most, if not all, individuals sharing the same patterns of crossmodal correspondence (Spence, 2011). A number of researchers have conducted cross-cultural studies in order to investigate the universality of crossmodal correspondences between the sound of spoken words and meaningless visual stimuli, a phenomenon known as sound symbolism, and often found just minor variations between different cultures (Davis, 1961; Gebels, 1969; Hinton, Nichols, & Ohala, 1994; Osgood, 1960; Rogers & Ross, 1968; Taylor & Taylor, 1962). Moreover, it has recently been demonstrated that Westerners can readily interpret non-Western metaphors used to describe the elevation of auditory pitch (Eitan & Timmers, 2010). Nevertheless, not

all crossmodal correspondences satisfy the universality criterion: Although most cultures match the pitch of a tone to the elevation of a visual stimulus, this was not true (that is, there was no mapping) for certain Native American populations studied by Walker (R. Walker, 1987; see also Antovic, 2009). Moreover, Marks (1974) has demonstrated that even within a given population, while the majority of participants matched increasing loudness to increasing lightness, a small proportion consistently matched increasing loudness to increasing darkness.

The fourth defining criterion of the correspondences under discussion in the present chapter, relativity, speaks to the comparative nature of crossmodal correspondences. Indeed, compatibility effects have only been reported for two pairs of crossmodal stimuli when the two stimuli presented in each modality vary noticeably on a given dimension along which one stimulus is 'more' (or 'less') than the other[3]. This is the case for the polar crossmodal correspondence between lightness and loudness: Generally-

[3] Though see Smith, Grabowecky, & Suzuki (2007) for tentative evidence of absolute correspondence in the case of gender judgments. However it should be noted that gender is a categorical rather than a continuous dimension.

speaking, louder auditory stimuli crossmodally correspond to lighter visual stimuli, whereas quieter sounds are associated with darker visual stimuli. What is lighter, and what is darker, however, depends on the context: Irrespective of the absolute lightness level, the lighter light will always be associated with the louder stimulus (Marks, 1989). Moreover, congruency effects between complementary multisensory cues have only been reported in those studies in which congruent and incongruent pairs of stimuli have been randomly interleaved within each block of trials, but not when congruent and incongruent stimuli are presented in separate blocks of experimental trials (e.g., Bernstein & Edelstein, 1971; Gallace & Spence, 2006).

Having restricted the focus of the present discussion, the next step is to draw a taxonomy accounting for the communalities and differences that exist between the various types of crossmodal correspondences that have been documented to date. Based on their putative underlying cause, three types of crossmodal correspondence between complementary polar cues are hereby identified.

2.3. A taxonomy of polar crossmodal correspondences

Structural correspondences. The suggestion is that these arise from the features of the nervous system. This idea dates back to S. S. Stevens (1957), who pointed to the fact that stimulus intensity is coded in the firing rate of neurons: The more intense the stimulus, the higher the rate of neuronal firing. Critically, this appears to hold true for every sensory modality, and hence the crossmodal correspondence between stimulus intensity in different modalities might simply reflect a common response of the brain to stimulus intensity. Such a structural hypothesis for crossmodal correspondences advocates the principle of neural economy (see Anderson, 2010), whereby the brain adopts similar mechanisms to process a number of different features from different sensory modalities, which, as a consequence, might end-up being associated. Of course, being a by-product of the anatomo-functional features of the human nervous system, such structurally-based dimensional interactions are likely to be universal.

Another form of structural correspondence can be postulated on the basis of Walsh's (2003) suggestion that there exists a multi-purpose system

coding for magnitude in the inferior parietal cortex in humans. It has been argued that this system encodes the magnitude of spatial, temporal, and qualitative features of sensory inputs in a common metric irrespective of their modality of input. According to this view, crossmodal correspondences might reflect the outcome of such a system encoding multiple estimates of magnitude within a common pool of neurons. There are, however, also other means by which the structural features of our nervous system might underlie crossmodal correspondences. Indeed, multiple sensory dimensions might be associated as a consequence of being processed in neighboring (see Ramachandran & Hubbard, 2001) or interconnected brain areas (see Rouw & Scholte, 2007). In this case, reciprocal connections and interactions between brain areas processing different sensory attributes might give rise to crossmodal correspondences. While this hypothesis provides an account for the existence of interactions between the processing of given sets of sensory attributes, it should be noted that it does not make specific predictions with regard to what will count as congruent or incongruent. In other words, dimensional interactions resulting from the processing of sensory information in adjacent/interconnected brain areas might easily

give rise to what is known as Garner interference (Garner, 1974), but not necessarily to Stroop-like interference (Stroop, 1935)[4].

Statistical correspondences. The physical properties of the distal stimuli activating our senses are often correlated in nature. As a result of repeated exposure, sensory systems learn the statistical regularities of the environment and the potential correlations between sensory cues (Adams, Graf, & Ernst, 2004; Stocker & Simoncelli, 2006; Weiss, Simoncelli, & Adelson, 2002). This information can then be used in order to decide which stimuli normally go together, and should therefore be integrated, and which to keep segregated (Ernst, 2007). A paradigmatic example of such statistical regularity is the relation between the size of physical bodies and their resonance frequencies: *Ceteris paribus*, the resonance frequency is inversely proportional to the size of the resonator. That is, keeping density, tension, temperature, etc. constant, large bodies resonate at lower

[4] Both Garner interference and Stroop interference are behavioural effects resulting from the interaction between task-relevant and task-irrelevant stimulus dimensions. Garner interference relates to the costs associated to the *variation* of an irrelevant dimension on the response to the relevant dimension. Stroop interference relates to the costs associate to the *incongruence* between the relevant and the irrelevant stimulus dimensions.

frequencies than smaller bodies. It would be adaptive for the brain to learn this relation between pitch (the perceptual correlate of a sound's frequency) and size and then to exploit this knowledge in order to better process/predict audiovisual information. In the chapters that follow I will provide evidence to substantiate this claim.

The literature on multisensory perception contains numerous other cases of compatibility effects between crossmodal stimulus features that might plausibly involve similar learnt stimulus correlations (Bernstein & Edelstein, 1971; Gallace & Spence, 2006; Marks, 1987a, 1987b, 1989, 2004; Martino & Marks, 2000; R. D. Melara & O'Brien, 1987; R. D. Melara & O'Brien, 1990). Reflecting the properties of the environment (and hence also the laws of physics), most statistical correspondences are again likely to be universal. Nevertheless, it is also possible for researchers to artificially introduce correlations between crossmodal features and train observers until their perceptual systems learn such correlations (Baier, Kleinschmidt, & Müller, 2006; Ernst, 2007; Zangenehpour & Zatorre, 2010). Experimentally-induced crossmodal correspondences of this kind would certainly also

qualify as statistical correspondences, though obviously they fail to meet the universality criterion. It is worth noting that the natural statistics of the environment might provide the basis for a number of the associations between crossmodal dimensions that have been documented in the literature to date. This is, for example, the case for the association between color and taste/flavour (e.g., see Levitan, Zampini, Li, & Spence, 2008; Spence, Levitan, Shankar, & Zampini, 2010).

Semantically (or linguistically) mediated correspondences. This kind of correspondence originates from the use of a common lexicon to describe multiple perceptual properties (Gallace & Spence, 2006; Long, 1977; Martino & Marks, 1999; R.D. Melara, 1989; P. Walker & Smith, 1984). So, for example, the first correspondence between complementary cues to have been reported in the literature was that between pitch and elevation (Eitan & Timmers, 2010; Stumpf, 1883). It turns out that many languages use the same words – 'high' and 'low' – to describe the pitch of sounds and the elevation of objects. Some crossmodal correspondences might, therefore, result from the common linguistic labels used to describe various perceptual

dimensions that eventually come to be associated. Of course, the very existence of such widespread linguistically-mediated correspondences begs the question of why so many languages 'just so happen' to use the same words to describe distinct perceptual properties from different senses. It might even be argued that linguistically-mediated correspondences build-on structural or statistical correspondences that, in the long run, might shape the languages' lexicons.

Irrespective of lexical labels, however, semantic information can underlie crossmodal correspondences because associated items show similar semantic profiles. For example, Bozzi and Flores D'Arcais (1967) studied the correspondences between invented words and simple shapes. They found that compatible word-figure pairs also shared similar semantic associations (see also Osgood, Suci, & Tannenbaum, 1957). That said, it is ambiguous whether a common semantic profile or lexical label was the cause of the crossmodal correspondences documented in this study, or whether instead crossmodal correspondences were the underlying cause for similar semantic profiles and common linguistic coding.

It is important to note that the taxonomy of crossmodal correspondences outlined here is by no means unequivocally defined nor should it necessarily be taken to be exhaustive. In fact, a given crossmodal correspondence can often be classified in terms of more than one category. It could even be argued that most (perhaps all) crossmodal correspondences originate in the statistical regularities of the environment: In a relatively short time-scale, statistical regularities might shape the lexicon so that features that are commonly associated in the world come to be described using the same words (this might even be thought of as an example of lexical, rather than neural, economy; cf. Anderson, 2010). Then, in a longer time-scale, the brain might come, through evolution, to develop similar strategies to process associated features (hence internalizing them as common neural coding principles, or co-locating them in adjacent, or well-connected brain regions), which, as a consequence, end-up giving rise to structural correspondences. However, in order to further substantiate any claims about the grounding of crossmodal correspondences on the natural statistics of the environment, researchers will, in the future, need to directly measure the physical properties of the environment and the natural

correlations that exist between the multiple physical dimensions that humans are sensitive to.

2.4. A review of the effects of crossmodal correspondences

Research on the topic of crossmodal correspondences has a long history in the field of experimental psychology. Linguistic and semantic correspondences were investigated first, and the history of research on this topic is intimately connected with that of sound symbolism. The first mention of a natural association between vocal stimuli and their meaning dates back to Plato's Cratylus, where the eminent philosopher discusses the arbitrariness of language and the symbolic properties of vocal sounds. The idea that a simple vocal sound might convey some meaning, namely the notion of sound symbolism, saw a surge of interest in the 20th century, primarily due to the work of Edgar Sapir (1929) and Wolfgang Köhler (1929, 1947).

In an attempt to investigate the symbolic values of vowels, Sapir (1929) gave participants two meaningless words, Mil and Mal, and told

them that they both meant 'table', but that one referred to a small table while the other word referred to a large table. When Sapir asked his participants to match those words to the sizes of the tables in the most natural way, he found that the majority associated the word Mil with the small table while associating Mal with the large table. Sapir interpreted this result by claiming *'that vocalic and consonantal contrasts tended with many, indeed with most, individuals to have a definite symbolic feeling-significance that seemed to have little relation to the associative values of actual words'* (p. 228). In the same year, Köhler (1929) documented what is now possibly the most famous example of sound symbolism. He gave participants two made-up words, Takete and Baluma, and two abstract shapes, a spiky and a rounded one, and told them that those two words were the names of the two shapes, but which was which? Most participants agreed that Baluma was the rounded figure, while Takete was the spiky one[5].

[5] In the 1947 version of the book 'Gestalt Psychology', Köhler changed 'Baluma' into 'Maluma', to avoid confusion with the word 'balloon', clearly denoting a round object

In the wake of Sapir (1929) and Kohler's (1929) early research, a number of scientists went on to further investigate sound symbolism, providing additional support for the notion of there being a non-arbitrary link between vocal sounds and meaning (Nuckolls, 2003; Parise & Pavani, 2011). Notably, cross-cultural studies found little difference in the magnitude or type of sound symbolic effects between different cultures (Davis, 1961; though see Diffloth, 1994), thus pointing to the existence of a universal mapping between meaning and at least certain speech sounds.

Another line of research on crossmodal correspondences developed in parallel with that on sound symbolism and focused on the associations between elementary features of non-linguistic crossmodal stimuli. One of the first such crossmodal correspondences to capture the attention of psychophysicists was the association between auditory loudness and visual brightness (Cohen, 1934; Hornbostel, 1938; Külpe, 1893), whereby both adults and children readily matched loud sounds with light patches and soft sounds to dark patches (Bond & Stevens, 1969; Root & Ross, 1965; Stevens & Marks, 1965).

The major scientific contribution of these pioneering studies was the discovery of a variety of crossmodal correspondences that were present in the normal (i.e., non-synesthetic) population. In the following years, researchers started to investigate the function of crossmodal correspondences and the focus of research interest rapidly shifted to studying the effects of such crossmodal phenomena on human perception and information processing (see Marks, 2004). Typically, researchers studied the interactions between corresponding crossmodal dimensions on the *speed* of information processing. Often the participants in these studies would have to classify a target dimension of stimuli presented in a given modality as rapidly as possible, while trying to ignore irrelevant stimuli presented in another sensory modality. As mentioned in Chapter 1, if the dimensions in the two modalities are not processed independently (i.e., if the dimensions are not orthogonal), reaction times (RTs) to the target stimuli should be modulated by variations of the stimuli in the irrelevant modality (Garner, 1974). Moreover, if certain dimensions of the target and irrelevant stimuli happen to correspond crossmodally, participants should respond more

rapidly when the stimuli in the two modalities are congruent as compared to when they are incongruent (Bernstein & Edelstein, 1971).

Congruency effects have been documented between many of the crossmodal correspondences highlighted in previous studies (see Spence, 2011, for a review) but, notably, only when compatible and incompatible stimulus pairs alternated in a random fashion on a trial-by-trial basis (Bernstein & Edelstein, 1971; Gallace & Spence, 2006). The effects of crossmodal correspondences on RTs led certain researchers to propose an interpretation of such dimensional interactions in terms of rate of information accrual (Martino & Marks, 1999), that is in term of the speed of accumulation of information necessary to initiate a response. The idea here is that in order to select a response, the human information processing system accumulates information simultaneously in the two channels, and only when the evidence reaches a certain threshold is a response initiated. If the two channels provide congruent information, the evidence accumulated in one modality interacts with the information provided by the other sensory modality and hence information accrual is speeded-up. Conversely,

when the information provided by the two channels is incompatible, the irrelevant modality slows down information accrual. More recently, congruency effects elicited by crossmodal correspondences have been interpreted in terms of a failure of selective attention (see Marks, 2004), the claim being that even if participants are instructed to attend exclusively to a single modality, they simply cannot ignore whatever stimulus happens to be presented in the other (irrelevant) modality.

Most RT studies of crossmodal correspondences have investigated the effects of stimulus congruency in tasks in which the participants always had to pay attention to the stimuli presented in a single modality, while the stimuli presented in the other modality provided congruent or incongruent task-irrelevant information. This paradigm, however, seems to tackle an observer's ability to ignore irrelevant information rather than investigating how multiple sources of sensory information jointly contribute to a participant's responses. In order to investigate whether stimulus congruency speeds-up human information processing, Miller (1991) tested whether the crossmodal correspondence between auditory pitch and elevation (specified

visually) could give rise to the redundant-targets effect (RTE). Miller's participants had to respond as quickly as possible to target stimuli. It is well known that the simultaneous presentation of multiple targets can speed-up observers' responses as compared to when single targets are presented due simply to statistical facilitation (Raab, 1962). According to the race model, when multiple stimuli are processed independently and the response is initiated by the faster process, as a result of probability summation it is more likely to provide faster responses, as compared to conditions with only a single target (Raab, 1962). Conversely, if multiple targets jointly co-activate a response, participants should exhibit an additional facilitation and respond faster than is predicted by the race model, an effect termed RTE (Miller, 1982).

In a go/no-go RT experiment, Miller (1991) instructed his participants to respond as rapidly as possible to visual stimuli presented above or below fixation, and to high- and low-pitched auditory stimuli, while refraining from responding in the presence of visual stimuli presented at fixation or to auditory stimuli having an intermediate pitch. Unimodal or bimodal stimuli

were presented pseudo-randomly on each trial. In line with the results of previous studies, participants responded more rapidly to congruent pitch-elevation combinations as compared to incongruent combinations on the bimodal trials. Furthermore, RTs to congruent bimodal stimuli systematically violated the race-model, thus providing evidence that congruent information about pitch and elevation is jointly processed, and concurrently activates observers' responses, hence leading to a significant RTE (though see Otto & Mamassian, 2010). Conversely, the race model was not violated when the audiovisual stimulus pairs were incongruent.

While such a paradigm allows for the detection of crossmodal congruency effects on human information processing, it is important to note that these results fail to provide any information about the processing level at which such effects take place. Indeed, dimensional interactions can arise on a *perceptual* level, thus leading to systematic distortions of perceptual experiences or at a subsequent stage of information processing (e.g., at the stage of *response selection*).

2.5. Concluding remarks

A century of research on the topic of crossmodal correspondences has provided robust empirical evidence on the effects of compatibility, or lack thereof, between elementary features of multisensory signals. Crossmodal correspondences are phenomenologically experienced as crossmodal similarities, and congruent crossmodal signals are processed faster and more accurately than incongruent stimuli. However, so far, no studies tried to investigate at which level of information processing (e.g., sensory vs decisional level) crossmodal correspondences operate, nor their function in multisensory integration. The following chapters provide evidence that crossmodal correspondences indeed operate on a perceptual level (Chapter 4 and 5) and that they promote multisensory integration by informing the system that congruent signals are more likely to originate from a common physical source (Chapter 5).

3. The implicit association test

3.1. Introduction

As mentioned in the Chapter 2, crossmodal correspondences can be defined as congruency effects between stimuli presented (either physically present, or merely imagined) in different sensory modalities that result from the 'expected' (i.e., a priori) mapping between those sensory cues (Parise & Spence, under review). This definition, however, begs the question of how to define congruency operationally. Over the years, researchers have utilized a number of experimental techniques in order to measure such crossmodal congruency effects. From the early studies, in which observers were explicitly required to match pairs of visual and auditory stimuli (see Davis,

1961; Köhler, 1929; Sapir, 1929), cognitive scientists have switched to more sophisticated paradigms allowing for repeated measures (rather than just a single measure as in many early studies) within a single observer, and often not relying on introspection. The most common technique over the last two or three decades has relied on the modulation of reaction times (RTs) in speeded classification tasks in which participants have to respond to stimuli on a target sensory modality, while trying to ignore task-irrelevant stimuli presented in a different sensory modality (see Marks, 2004; Spence, 2007, for a review). Despite the fact that the stimuli presented in one sensory modality are completely task-irrelevant, participants' RTs are often faster for certain combinations of (relevant and irrelevant) stimuli, and slower for others. Based on this Stroop-like crossmodal interference on response latencies, stimulus combinations leading to faster RTs (and more correct responses) are considered to be compatible, while those leading to longer RTs (and more incorrect responses) are considered as being incompatible.

Other techniques, instead, rely on explicit measures of similarity. This is the case of the crossmodal matching task (Stevens & Marks, 1965), where

participants have to adjust the magnitude of a stimulus along a given sensory dimension (e.g., loudness) until it phenomenologically 'matches' the magnitude another sensory dimension in a different modality (e.g., brightness). A more constrained variant of this technique was later proposed by Marks (1989). He had participants select which of two stimuli (whose properties were parametrically manipulated on a trial by trial basis) in a given modality better matched a target stimulus in a different sensory modality. In addition to these techniques, many others approaches have also been used in the study of crossmodal correspondences over the years. They include the use of the semantic differential technique (Bozzi & Flores D'Arcais, 1967; Osgood, 1960; Osgood, et al., 1957; see also Hevner, 1935; Poffenberger & Barrows, 1924), preferential looking (P Walker et al., 2010), and cueing (whereby the crossmodal congruence or incongruence of a cue preceding the target stimulus modulated RT to the target stimulus, Melara & O'Brien, 1990; see also Klein, Brennan, & Gilani, 1987) to name but a few.

In spite of providing important insights into the underlying nature of crossmodal correspondences, most of these techniques suffer from various methodological limitations that potentially compromise the interpretation of many of the empirical results. Explicit measures of associations, such as the crossmodal matching task in its various forms, or the semantic differential technique, rely on observers' introspection. Therefore the results critically depend on observers' ability (and/or willingness) to report on their introspections. Such limitations have been overcome by indirect techniques based on RTs, such as the speeded classification task. Nevertheless, these tasks exhibit a number of further limitations. First, while the speeded classification paradigm provides evidence that compatibility between, say, auditory and visual stimuli interacts in terms of the processing of visual information, it does not address the reciprocal effects on audition within the same experiment. For this reason, using the speeded classification task, it is harder and more time consuming to measure any potential asymmetries between the magnitude of the effect of one modality over another and vice-versa. Moreover, given that two stimuli in different modalities have often been presented at the same time in each trial, any stimulus-dependent

modulation of response latencies might reflect some form of failure of selective attention (Gallace & Spence, 2006; R. D. Melara & O'Brien, 1987), with participants being unable to fully focus their attention on the target stimuli and ignore the distracting stimuli. In addition to the various limitations of the traditional techniques, such methodological fragmentation inevitably leads to further difficulties when it comes to trying to compare the results of different studies.

In order to overcome such issues and problems, the compatibility between crossmodal stimuli was measured using a variant of the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998). The IAT is currently one of the most popular tools with which to study the association between different items. In the simplified version of the task used here, participants respond as rapidly as possible to a series of stimuli, taken from a set of four stimuli (i.e., two auditory stimuli, *mil* and *mal* and two visual stimuli, a small and a large circle). They use just two response keys, with two stimuli (i.e., one auditory and one visual) being assigned to the same response key in a given block of trials. Previously, it has been demonstrated

that participants' performance improves when the set of stimuli assigned to a given response key are also associated with each other (the compatible conditions), as compared to conditions in which a set of unrelated (or incompatible) stimuli are assigned to the same response key (the incompatible conditions; Greenwald, et al., 1998). In the present study, the assignment of the four stimuli to each response key was experimentally manipulated from block to block during the course of the experiment, so that half of the blocks were assumed to be compatible and the other half incompatible. Discrepancies in RT between different stimulus-response key assignments are taken to provide evidence of the existence of a compatibility effect: faster RTs indicate associations between the stimuli assigned to the same response key, while slower RTs indicate weaker associations. One important advantage associated with using such a technique (over, say, the speeded classification task) is that given that on each trial only visual or auditory stimuli are individually tested, the IAT allows to rule out any potential account for the effects of crossmodal correspondences in terms of selective attention. Moreover, it is based on a standard technique that has proven to be very sensitive to associations

between stimuli from a variety of categories, and it is flexible enough to be adapted to crossmodal settings (Crisinel & Spence, 2009, 2010; Dematte, Sanabria, & Spence, 2006, 2007). Finally, given that observers have to respond to both visual and auditory stimuli, the modified version of the IAT used here provides evidence of crossmodal compatibility effects on both visual and auditory responses within a single experiment.

In the experiments reported in this chapter, the IAT is used to replicate some well-known examples of crossmodal correspondence (including classic examples from the literature on sound symbolism) never studied using the IAT before, namely takete-maluma, and mil-mal, and the association between auditory pitch and the size of visual objects. Moreover, the existence of two additional postulated crossmodal correspondences is also investigated – namely that between auditory pitch and the size of visual angles, and that between the waveform of auditory stimuli and the spikiness-roundedness of visual stimuli. Given that in the present study the same method is used in all of the five experiments, the description of the experiments is combined into a single methods section.

3.2. Methods

Participants

Overall, fifty participants (twenty-six females) took part in the present study (ten different participants for each of the five experiments). Their mean age was twenty-three years (range 18-35 years), and all of the participants reported normal or corrected-to-normal vision and audition. The gender and age of the participants were roughly matched across experiments. Each session lasted for approximately 35 minutes and participants received a £5 (UK Sterling) voucher in return for taking part in the study. The experimental procedure was approved by the Ethics Committee of the Department of Experimental Psychology, University of Oxford.

Apparatus and materials

The presentation of the stimuli and the collection of the responses were controlled by a personal computer running the Psychtoolbox v.2.54

(Brainard, 1997; Pelli, 1997). Each participant was seated in front of a 21' CRT computer monitor with a resolution of 1280x1024 pixels (75Hz refresh rate) flanked by a pair of loudspeakers. Participants responded to the target stimuli by pressing a key of a computer keyboard. The experiment was conducted in a dark and quiet room.

Stimuli

Two visual stimuli and two auditory stimuli were used in each experiment. Details of the stimuli used in each experiment are reported in Table 3.1.

Experiment 3.1: The visual stimuli consisted of two light grey circles, one subtending 5 cm and the other subtending 2 cm (5.2° vs. 2.1° of visual angle, respectively), presented at the centre of the screen against a white background. The auditory stimuli consisted of the words 'mil' and 'mal' pronounced by a female voice.

Experiment 3.2: The visual stimuli consisted of two shapes, one spiky, the other curved (Köhler, 1947, see Table 3.1), respectively subtending

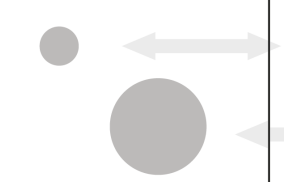
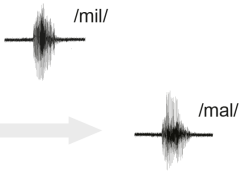
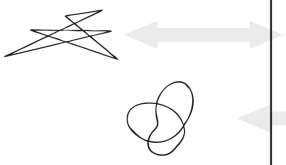
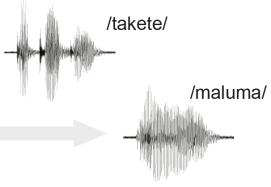
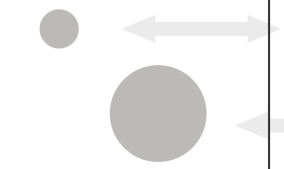
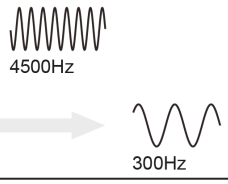
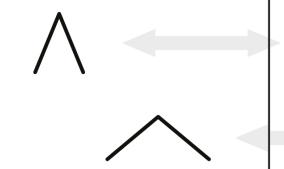
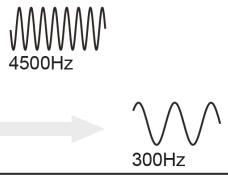
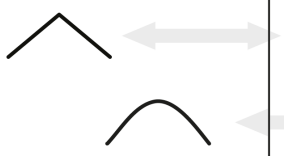
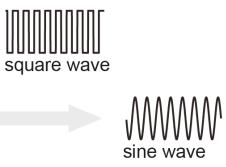
Exp	Visual stimuli	Auditory stimuli	IAT Results
3.1			Congruency F(1,9)=23.84 p<.001 Modality F(1,9)=33.42 p<.001 Compatibility X Modality F(1,9)<1 n.s.
3.2			Congruency F(1,9)=22.08 p=.001 Modality F(1,9)=38.26 p<.001 Compatibility X Modality F(1,9)=2.45 p=.15
3.3			Congruency F(1,9)=11.07 p=.009 Modality F(1,9)=12.92 p=.006 Compatibility X Modality F(1,9)<1 n.s.
3.4			Congruency F(1,9)=16.54 p=.003 Modality F(1,9)=13.42 p<.006 Compatibility X Modality F(1,9)<1 n.s.
3.5			Congruency F(1,9)=5.71 p=.041 Modality F(1,9)=21.45 p=.001 Compatibility X Modality F(1,9)=2.45 p=.15

Table 3.1. Stimuli and results of the statistical analysis of the experiments reported in the present chapter (see main text for further details regarding the stimuli and the analysis). The light gray double-headed arrows connect compatible audiovisual pairs of stimuli.

6.24x3.12°, and 4.16x4.68° of visual angle, and presented at the centre of the screen against a white background. The auditory stimuli consisted of the words ‘takete’ and ‘maluma’ pronounced by a female voice.

Experiment 3.3: The visual stimuli consisted of two light grey circles, one subtending 5 cm and the other subtending 2 cm (5.2° vs. 2.1° of visual angle, respectively), presented at the centre of the screen against a white background. The auditory stimuli consisted of two pure tones, a high and a low pitched one (4500Hz and 300Hz respectively). The perceived intensities (loudness) of the 300ms tones were individually matched for each participant with a brief preliminary psychophysical experiment based on the QUEST procedure (Watson & Pelli, 1983).

Experiment 3.4: The visual stimuli consisted of the two angles (i.e., arrowheads), one acute and the other obtuse (42° and 126°, respectively) subtended by two segments, each segment subtending 4.3° of visual angle. The auditory stimuli were the same as those used in Experiment 3.3.

Experiment 3.5: The visual stimuli consisted of an angle and a curve, both subtending a visual angle of 6.8°x2.9°, presented at the centre of the screen against a white background. The auditory stimuli consisted of two tones with a fundamental frequency of 440Hz and varying in waveform, with one being sinusoidal and the other being square. The perceived intensities (loudness) of the two tones were individually matched for each participant with a brief preliminary psychophysical experiment based on the QUEST procedure (Watson & Pelli, 1983).

Procedure

The participants were instructed to maintain their fixation on the centre of the screen and to respond to the stimuli as rapidly and accurately as possible, by pressing one of two keys on a computer keyboard. Two patches, representing an arrow pointing either to the left or to the right, marked the relevant response keys. Each trial began with the presentation of a red fixation point from the centre of the screen for a randomized interval of 500-600ms. After the removal of the fixation point, there followed a random interstimulus interval of 300-400 ms. Next, the target

stimulus, either visual or auditory, was presented. The visual stimulus remained on the screen for 300ms before being removed. The auditory stimuli, also lasting for 300ms (or approximately 300ms in Experiments 3.1 and 3.2), were repeated only once on each trial. Feedback in the form of a red cross was provided after each incorrect response and remained on the screen for 500ms.

At the beginning of each block of trials, the participants received new instructions about the mapping between the stimuli and the appropriate response for the upcoming block of experimental trials. On each block of trials, two of the four stimuli, one figure and one word, were assigned to either the left of the right key and the remaining stimuli to the other response key. The instructions remained visible on the screen until the participants initiated the new block of trials by pressing the space bar. The mapping of the stimuli onto the response keys was manipulated during the experiment thus generating four different stimulus pairings, of which two were hypothesized to be compatible (e.g., in Experiment 3.1 the small circle and the word ‘mil’ associated with the same key; Sapir, 1929) while the

remaining two were judged as being incompatible (e.g., in Experiment 3.1, the large circle and the word ‘mil’ associated with the same key). Note that a block of trials was considered as being ‘compatible’ when the two stimuli associated with a given response key were hypothesized to be associated with one another. Conversely, a block of trials was considered as being incompatible when the hypothetically associated stimuli were mapped onto different response keys. Each of the four pairings was repeated six times for a total of 24 randomly alternating blocks. Each block consisted of 20 trials (each stimulus repeated four times) for a total of 480 trials. Block order was randomized across participants. Participants were allowed to take a break at the end of each block. RTs and the accuracy of participants’ responses were collected.

Instruction about the mapping between the stimuli and the relevant response keys consisted of a schematic representation of the two response keys with the corresponding visual stimuli displayed next to them. The participants were required to press two keys ‘1’ and ‘2’ on the keyboard to listen to the stimulus associated respectively with the left and right

response keys. There were no time limits to learn the new stimulus-response mapping, and participants were encouraged to listen to the auditory stimuli as much as they wanted, until they were sure that they had learnt the new assignment.

3.3. Results

The first four trials of each block, in which the participants were presumably still learning the new stimulus-response mapping were not included in the data analysis. In order to normalize the RT distributions, the RT data were log-transformed, and responses that fell three standard deviations above or below the individual log-means were excluded from further analyses. The RTs from those trials in which participants responded correctly were submitted to a repeated-measures analysis of variance (ANOVA) with the within-participants factors of stimulus-response compatibility (compatible versus incompatible) and stimulus type (words versus pictures). The results of the analysis are reported in Table 3.1 (see Figures 3.1-3.5).

Overall, a significant crossmodal congruency effect was observed in *all* five experiments, indicating that all of the crossmodal correspondences investigated here significantly modulated the latency of participants' behavioural responses. Moreover, in all five experiments, there was also a significant effect of stimulus modality, showing that participants responded more rapidly to visual than to auditory stimuli overall (see also Evans & Treisman, 2010). Interestingly, there was no interaction between compatibility and stimulus modality in any of the experiments, indicating that the effect of compatibility influenced participants' performance regardless of the stimulus type.

In order to study the build-up of the compatibility effect and thus determine at which stage of information-processing it was taking place, I ran a bin analysis of RTs (see De Jong, Liang, & Lauber, 1994), by dividing the RT data into 5 bins, from fastest to slowest. This procedure was performed separately for each participant, modality and stimulus-response compatibility. This analysis revealed that participants' RTs were slower in incompatible than in compatible trials irrespective of the bin. To further

highlight this difference, the effect size (d-score, Cohen, 1977, 1988) of compatibility for each bin was calculated by diving the RT difference between incompatible and compatible trials by the overall standard deviation of that bin (calculated by pooling together compatible and incompatible responses for each bin). Overall, the effect size was higher in the central bins. However, for all five of the experiments reported here, the d-scores were positive even in the first bin, thus indicating that the stimulus compatibility modulated response latencies even when RTs are very fast. Such a result argues for an early onset of the crossmodal compatibility effect.

In order to compare the size of the congruency effects for the visual and auditory targets across the five experiments reported in the present chapter, the overall d-score for visual and auditory responses for each participant and experiment were calculated. An ANOVA on the d-scores, with stimulus modality as a within-participants factor and experiment as a between-participants factor, revealed no main effect of experiment ($F(4,45)<1$, ns), no main effect of modality ($F(1,4)=1.053$, $p=.31$), and no interaction ($F(4,45)=1.075$, $p=.38$, see Figure 3.6).

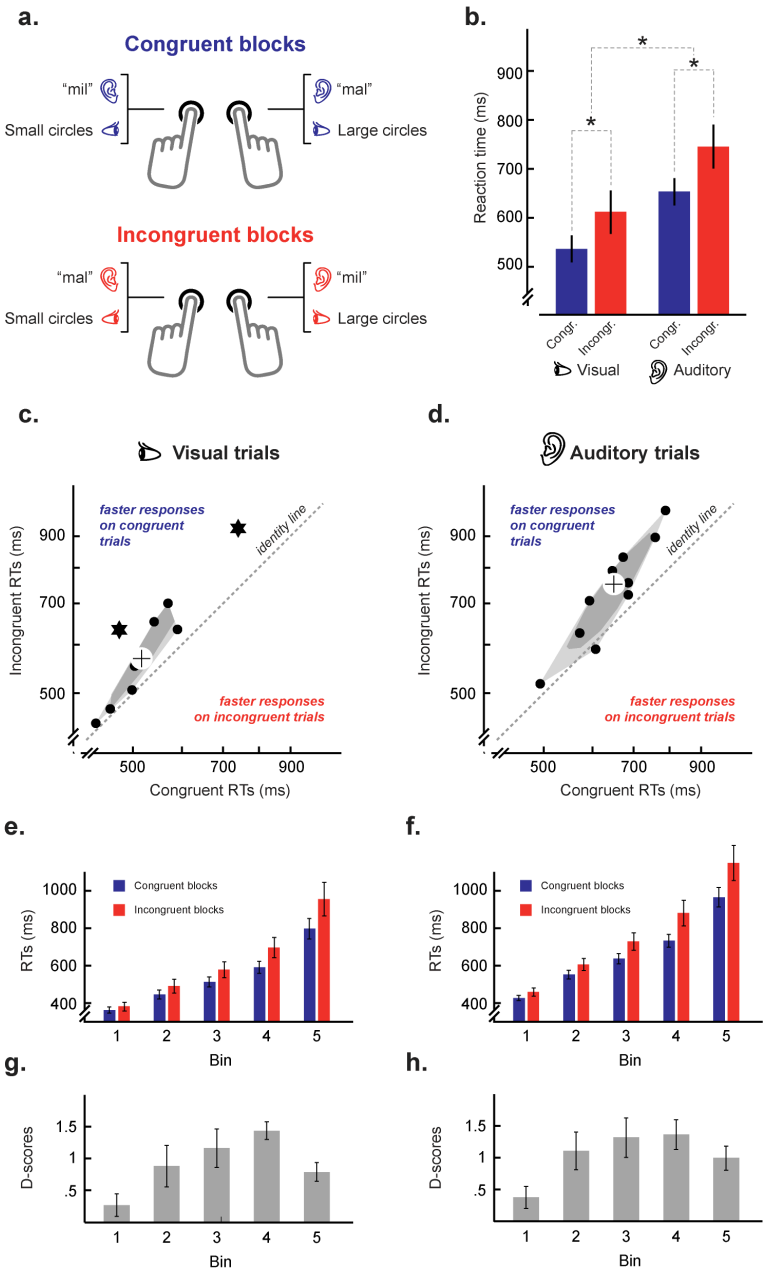


Figure 3.1. The mil-mal effect modulates observers' RTs in Experiment 3.1. (a) Examples of stimulus-response key assignment (top: congruent; bottom: incongruent). (b) Mean RTs for congruent and incongruent trials on visual and auditory trials. Error bars represent the standard error of the mean across participants and the asterisks indicate statistical difference ($p<.05$). (c) Scatter and bagplot of participants' mean visual RTs on congruent vs. incongruent trials. (d) Scatter and bagplot of participants' mean auditory RTs on congruent vs. incongruent trials. The cross at the centre of the bagplot represent the centre of mass of the bivariate distribution of empirical data (i.e., the halfspace depth), the dark gray area (i.e., the bag) includes the 50% of the data with the largest depth, the light gray polygon contains all the non-outliers data points, and the stars represent the outliers (Rousseeuw, Ruts, & Tukey, 1999). (e-f) Mean RTs of congruent (blue) and incongruent (red) visual (e) and auditory (f) trials for each bin. Mean effect size of the RT difference between incongruent and congruent RTs for each bin on visual (g) and auditory (h) trials. In all four panels, error bars represent the standard error of the mean.

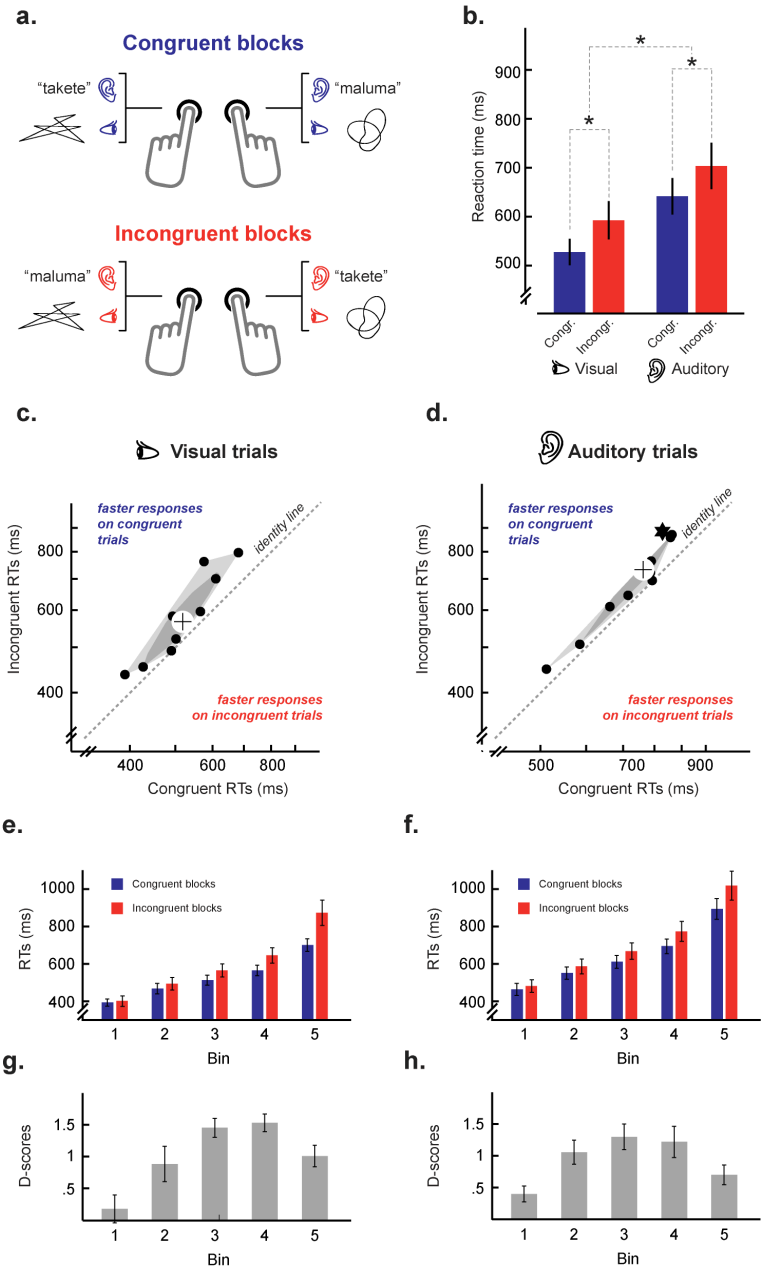


Figure 3.2. The takete-maluma effect modulates observers' RTs in Experiment 3.2. (a) Examples of stimulus-response key assignment (top: congruent; bottom: incongruent). (b) Mean RTs for congruent and incongruent trials on visual and auditory trials. Error bars represent the standard error of the mean across participants and the asterisks indicate statistical difference ($p<.05$). (c) Scatter and bagplot of participants' mean visual RTs on congruent vs. incongruent trials. (d) Scatter and bagplot of participants' mean auditory RTs on congruent vs. incongruent trials. (e-f) Mean RTs of congruent (blue) and incongruent (red) visual (e) and auditory (f) trials for each bin. Mean effect size of the RT difference between incongruent and congruent RTs for each bin on visual (g) and auditory (h) trials. In all four panels, error bars represent the standard error of the mean.

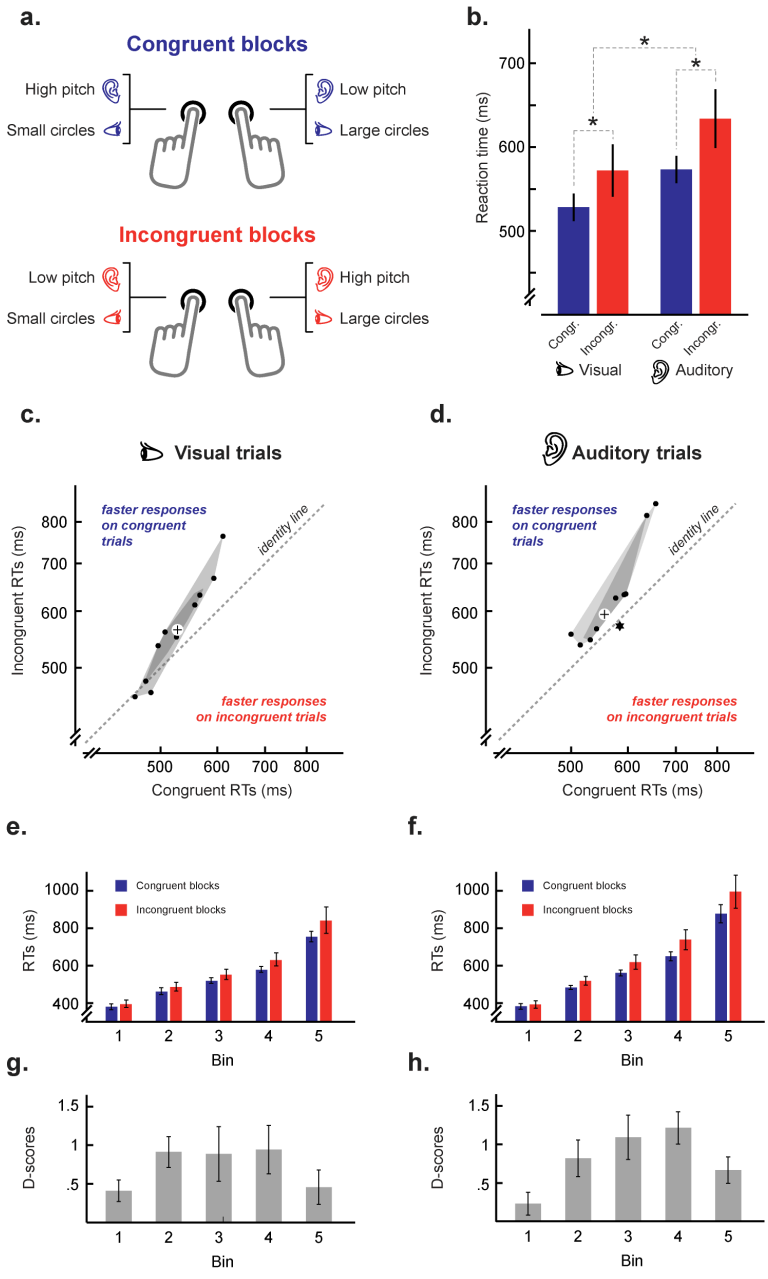


Figure 3.3. Pitch-size compatibility modulates observers' RTs in Experiment 3.3. (a) Examples of stimulus-response key assignment (top: congruent; bottom: incongruent). (b) Mean RTs for congruent and incongruent trials on visual and auditory trials. Error bars represent the standard error of the mean across participants and the asterisks indicate statistical difference ($p<.05$). (c) Scatter and bagplot of participants' mean visual RTs on congruent vs. incongruent trials. (d) Scatter and bagplot of participants' mean auditory RTs on congruent vs. incongruent trials. (e-f) Mean RTs of congruent (blue) and incongruent (red) visual (e) and auditory (f) trials for each bin. Mean effect size of the RT difference between incongruent and congruent RTs for each bin on visual (g) and auditory (h) trials. In all four panels, error bars represent the standard error of the mean.

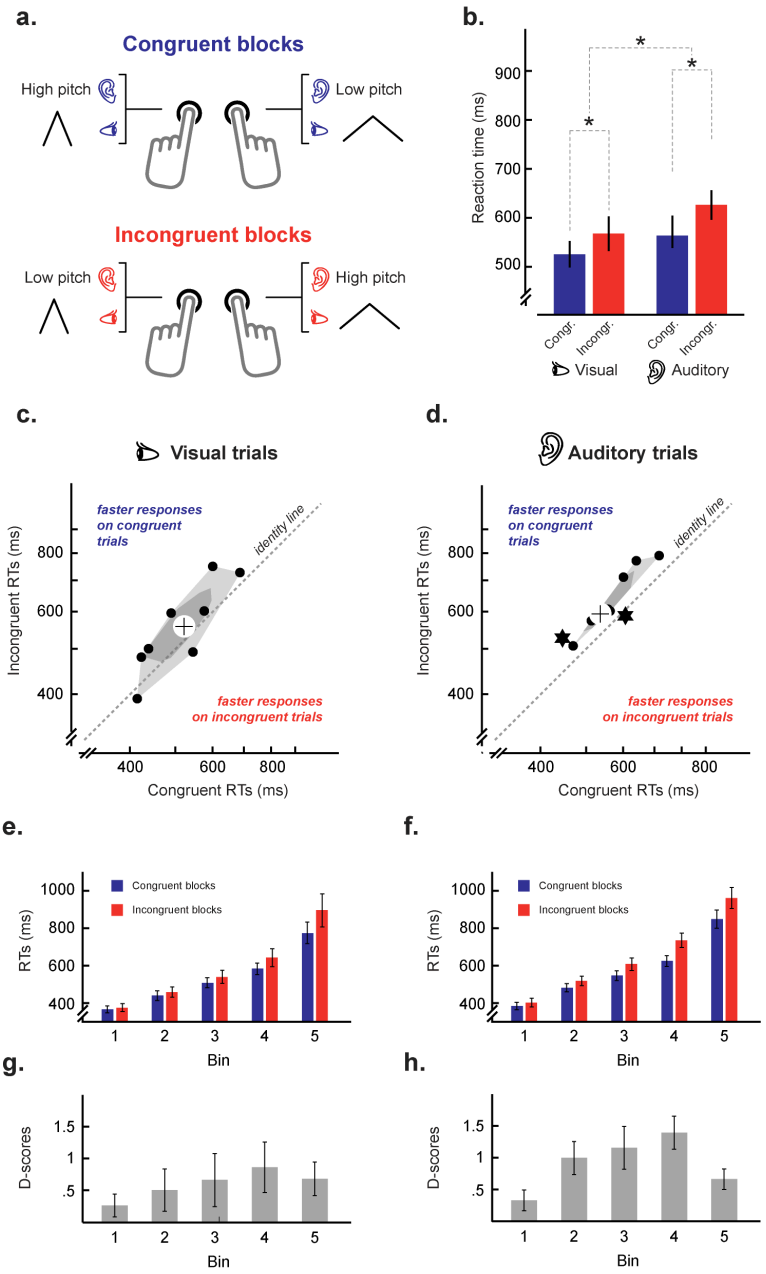


Figure 3.4. Pitch-angle compatibility modulates observers' RTs in Experiment 3.4. (a) Examples of stimulus-response key assignment (top: congruent; bottom: incongruent). (b) Mean RTs for congruent and incongruent trials on visual and auditory trials. Error bars represent the standard error of the mean across participants and the asterisks indicate statistical difference ($p<.05$). (c) Scatter and bagplot of participants' mean visual RTs on congruent vs. incongruent trials. (d) Scatter and bagplot of participants' mean auditory RTs on congruent vs. incongruent trials. (e-f) Mean RTs of congruent (blue) and incongruent (red) visual (e) and auditory (f) trials for each bin. Mean effect size of the RT difference between incongruent and congruent RTs for each bin on visual (g) and auditory (h) trials. In all four panels, error bars represent the standard error of the mean.

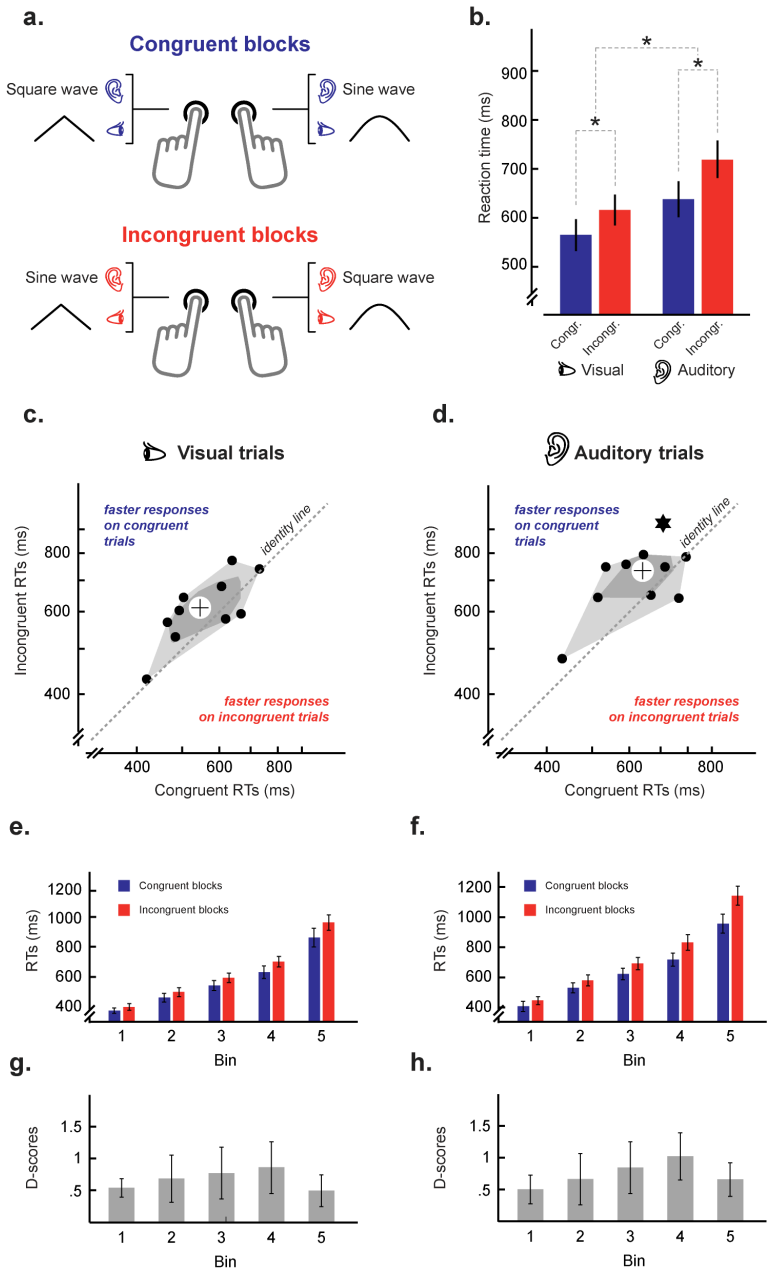


Figure 3.5. Waveform -roundedness compatibility modulates observers' RTs in Experiment 3.5. (a) Examples of stimulus-response key assignment (top: congruent; bottom: incongruent). (b) Mean RTs for congruent and incongruent trials on visual and auditory trials. Error bars represent the standard error of the mean across participants and the asterisks indicate statistical difference ($p<.05$). (c) Scatter and bagplot of participants' mean visual RTs on congruent vs. incongruent trials. (d) Scatter and bagplot of participants' mean auditory RTs on congruent vs. incongruent trials. (e-f) Mean RTs of congruent (blue) and incongruent (red) visual (e) and auditory (f) trials for each bin. Mean effect size of the RT difference between incongruent and congruent RTs for each bin on visual (g) and auditory (h) trials. In all four panels, error bars represent the standard error of the mean.

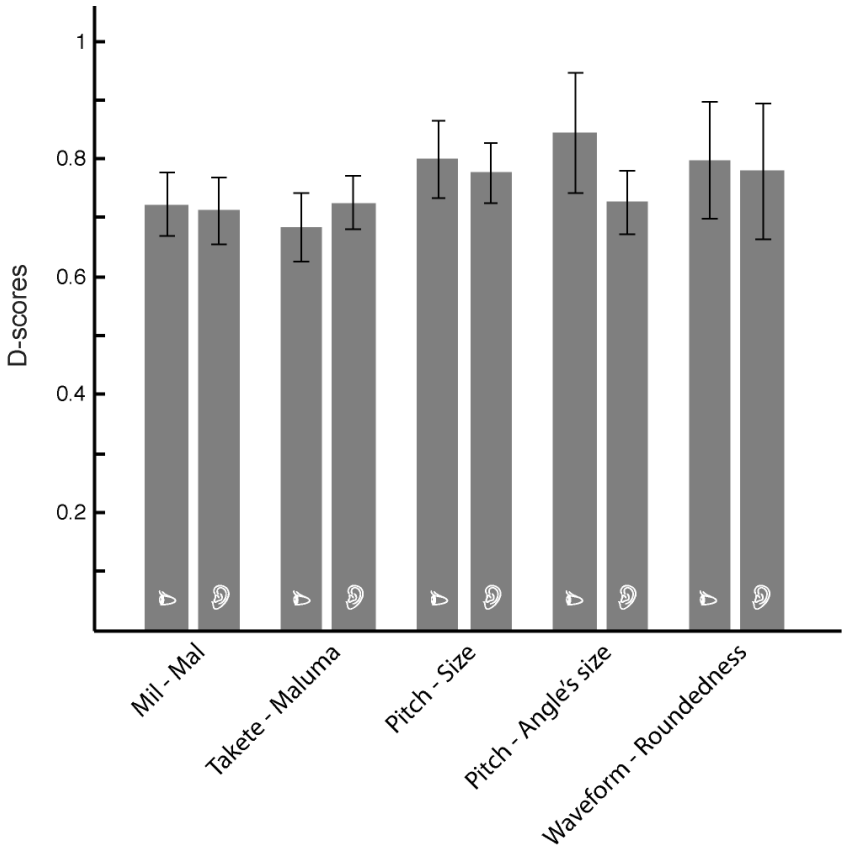


Figure 3.6 Comparison of the effect size (d-score) for vision and audition between the five experiments. Note that all of the crossmodal correspondences had a very similar effect size. Error bars represent the standard error of the mean.

3.4. Discussion

The results of the five experiments reported in the present study demonstrate the existence of several crossmodal associations between

auditory and visual stimuli. In particular, several of the traditional results from the literature on sound symbolism (i.e., takete/maluma, and mil/mal) have been replicated together with a finding from the literature on crossmodal correspondences (i.e., the association between auditory pitch and visual size). Moreover, the existence of two entirely new crossmodal correspondences have been demonstrated, namely the association between auditory pitch and the size of visual angles, and between the waveform of auditory stimuli and the roundness of visual shapes.

This is the first study specifically to investigate the famous takete/maluma and mil/mal effects using an *indirect* performance measure with auditory (rather than written) verbal stimuli. These results therefore demonstrate that such effects cannot simply be attributed to some kind of similarity between the shape of the visual stimuli and the visual appearance of the written words (see Westbury, 2005). This claim is further supported by the bin analysis, where the early effects of compatibility make it unlikely that participants had time to imagine the shape of the words. Rather, the

results reported here demonstrate that the similarity involves, at least in the early stage, the physical features of the visual and the auditory stimuli.

Interestingly, all of the crossmodal correspondences studied in this chapter had effect sizes that were similar in magnitude (see Figure 3.6). This result suggests that crossmodal correspondences involving elementary stimulus features, such as pitch and size, and those involving more complex stimuli, such as nonsense words and line drawings, are equally effective in inducing crossmodal compatibility effects. Moreover, these results demonstrate that all of the crossmodal correspondences tested here affected participants' responses to visual and auditory stimuli similarly. This also suggests that crossmodal compatibility effects are not modality-specific or modality-dependent (see Evans & Treisman, 2010). The fact that the effect sizes of crossmodal correspondences are remarkably similar between different experiments and modalities is consistent with there being a unique modality -independent mechanism coding for crossmodal correspondences.

These results move beyond simply replicating previous findings showing that the processing of unimodal sensory stimuli are faster under

conditions of crossmodal compatibility, but do so with a single technique that has been specifically designed to measure associations between stimuli. As mentioned in the Introduction, the modified IAT utilized here has several advantages over other traditional techniques. First, the IAT provides an indirect (i.e., non-explicit) measure of the association between stimuli presented in different sensory modalities, therefore the results demonstrate that all of the crossmodal correspondences investigated here are automatically encoded by participants, irrespective of any voluntary effort aimed at recognizing the compatibility or incompatibility of the stimulus pairs. This conclusion is further supported by the bin analysis, demonstrating that congruency effects modulated RTs even in the fastest responses, supposedly less influenced by top-down cognitive control.

Second, given that in the IAT only one stimulus is presented on each trial and that both modalities were equally relevant to the task, unlike previous findings, the present results cannot be interpreted in terms of costs and/or benefits associated to the simultaneous presentation of certain combinations of stimuli, nor in terms of a failure of selective attention (see

Marks, 2004; Spence, 2011), nor in terms of costs and benefits of multisensory integration (see Chapters 4 and 5). In classic interference tasks, two stimuli are always presented at the same time with participants being instructed to respond to only one of them. It is therefore unclear how much of the reported effects are due to the presence of an irrelevant stimulus and how much to the effect of stimulus compatibility per se. Given that these confounds are not present in the modified version of the IAT used here, the present results qualify as more genuine effects of stimulus compatibility. Moreover, given that both modalities are task relevant, the IAT allows one to measure how crossmodal compatibility affects the processing of both visual and auditory stimuli.

Additionally, by ensuring that only a single (unimodal) stimulus is presented at any given time, the IAT overcomes every issue concerning potential spatiotemporal inconsistencies in the combined presentation of audiovisual signals. When auditory and visual stimuli are jointly presented, any offset in their relative position, such as when the visual stimuli are presented on the screen while the auditory stimuli are played through

headphones, might alter multisensory processing and hence interfere with the crossmodal congruency effects that are observed (e.g., see Soto-Faraco, Lyons, Gazzaniga, Spence, & Kingstone, 2002). Similar problems also occur in the temporal domain, where asynchronies between auditory and visual stimuli might occur due to physical and neural delays. Both physical delays (e.g., due to timing inaccuracies in the experimental setup) and neural delays, (e.g., due to the auditory system being generally faster than the visual system, Spence & Squire, 2003) can underlie potential asymmetries in the effect of compatibility, whereby stimuli in a given modality can alter the processing on a second modality (Chen & Spence, 2010), but not vice-versa (see Evans & Treisman, 2010).

Previous studies claimed that sound symbolic effects are robust based on the consistency of the responses provided by a large number of participants (see Robson, 2011). In this regard, the IAT allows one to assess the strength of crossmodal correspondences and sound symbolic association in a more subtle way than traditional techniques. Being based on a large number of responses from a single observer, the IAT also allows one to

measure the strength of crossmodal correspondences within individual participants, hence providing a measure of individual differences. Moreover, being an indirect technique, the measure of crossmodal correspondence provided by the IAT should not be biased by consensual thinking. That is by participants trying to respond as they believe that others would (Koriat, 2008).

All of the crossmodal correspondences previously reported in the literature, and investigated in the present study have been successfully replicated with a modified version of the IAT. Together with the fact that the IAT provides a standard method for measuring (implicit and explicit) associations between a wide range of items, these results suggests that the IAT should be more extensively used for measuring correspondence between crossmodal and unimodal sensory signals, and might be a key technique for discovering novel correspondences. Moreover, not relying on explicit responses, the IAT might be suitable to investigate crossmodal correspondences and sound symbolism in special populations, such as autistic individuals (which according to previous research do not show

direct evidence of sound symbolism, Oberman & Ramachandran, 2008; V. S. Ramachandran & Oberman, 2007) or even in primates or other animals (e.g., see Cowey & Weiskrantz, 1975; Parker & Easton, 2004; Weiskrantz & Cowey, 1975; Parker & Easton, 2004).

4. The temporal ventriloquist

4.1. Introduction

It is now well-documented that people are typically unable to direct their behaviour on the basis of the information provided by a single sensory channel without also potentially being influenced, often without their awareness, by whatever stimuli may be being presented to the other modalities at the same time. A classic example of this phenomenon is the spatial ventriloquism effect (see Radeau, 1994, for an exhaustive review), where the source of a sound is mislocalized toward the position of a concurrent and task-irrelevant visual stimulus (e.g., Alais & Burr, 2004;

Bertelson & Aschersleben, 1998; Caclin, Soto-Faraco, Kingstone, & Spence, 2002).

More recently, a similar phenomenon has been demonstrated in the temporal domain, whereby the perceived time of occurrence of a visual stimulus can be biased by the presentation of an irrelevant, and slightly asynchronous auditory stimulus (Scheier, Nijhawan, & Shimojo, 1999); though see Fendrich & Corballis, 2001). This phenomenon has been labelled the temporal ventriloquism effect (see Morein-Zamir, Soto-Faraco, & Kingstone, 2003). The claim is that the sensitivity of a participant's judgments concerning the temporal order in which a pair of visual stimuli were presented is enhanced (i.e., the just noticeable difference, JND, is reduced) when two auditory stimuli are presented, one shortly before the first visual stimulus and the other shortly after the second visual stimulus. Researchers have interpreted this phenomenon in terms of the auditory capture of vision, with the second visual stimulus shifted temporally toward the time of occurrence of the second auditory stimulus, thus expanding the

perceived temporal gap between the two visual events (Morein-Zamir, et al., 2003)[6].

A number of subsequent studies have gone on to examine the spatial and temporal constraints on the temporal ventriloquism effect (see Aschersleben & Bertelson, 2003; Bertelson & Aschersleben, 2003; Jaekl & Harris, 2007; Keetels, Stekelenburg, & Vroomen, 2007; Morein-Zamir, et al., 2003; Vroomen & Keetels, 2006). For example, Morein-Zamir and her colleagues found that auditory stimuli could shift the perceived time of occurrence of visual stimuli that had been presented 200ms earlier. More recently, using a somewhat different paradigm, Jaekl et al. (2007) observed a temporal cross-capture of audiovisual stimuli only when the separation between the auditory and visual stimuli did not exceed 125ms. Interestingly, Vroomen and Keetels (2006) have shown that the relative spatial position from which the auditory and visual stimuli are presented

[6] The temporal ventriloquism paradigm, as proposed by Morein-Zamir et al. (2003), only allows one to measure the auditory capture of vision, but it should be noted that other studies (Ascherleben & Bertelson, 2003; Fendrich & Corballis, 2001) have also reported that visual stimuli may give rise to a modest capture of audition as well, thus demonstrating that the perceived time of occurrence of asynchronously-presented auditory and visual stimuli are both shifted toward each other.

does not seem to modulate the size of this particular multisensory (temporal) effect.

To date, however, no one has investigated whether the qualitative aspects of the stimuli might influence the strength of this temporal ventriloquism effect. That is, in all of the studies that have been published thus far, just one type of stimulus was presented in each sensory modality (though see Bushara, Grafman, & Hallett, 2001). As a consequence, we still do not know whether the temporal ventriloquism effect would be modulated by other multisensory integration phenomena triggered by the intrinsic (or relative) features of the stimuli presented.

The present chapter investigates the role of crossmodal correspondences in audiovisual temporal ventriloquist. In particular, the working hypothesis is that auditory stimuli that are crossmodally congruent (with visual stimuli) might give rise to a stronger temporal ventriloquism effect than sounds that are incongruent with their respective visual stimuli. If demonstrated, such finding would provide the first evidence that crossmodal correspondences not only alter the speed of information

processing (as demonstrated in Chapter 3), but also operate on a perceptual level and modulate observers' sensitivity when making unspeeded perceptual judgments. In other words, this experiment aim at testing whether crossmodal correspondences alter temporal perception rather than just modulate reaction times as previously demonstrated (see Chapter 2 and 3).

This prediction was tested capitalizing on the crossmodal correspondence between pitch and size (see Chapter 3, Experiment 3.3). Participants had to judge the temporal order in which two asynchronous visual stimuli (one delivered to either side of a computer monitor) were presented, while ignoring two irrelevant auditory stimuli, one presented slightly ahead and the other slightly behind the two relevant visual stimuli. Given that temporal ventriloquism has been interpreted in terms of the auditory capture of vision in the temporal domain (Morein-Zamir, et al., 2003), it was hypothesized that the strength of any such crossmodal capture might be modulated by the crossmodal congruency between the stimuli in the two modalities, with stronger capture taking place between congruent

stimuli as compared to incongruent stimulus pairs. According to this hypothesis, it was expected to find that participants' temporal sensitivity in a temporal order judgment (TOJ) task might be higher on congruent trials (where the first sound was congruent with the first visual stimulus and the second visual stimulus was congruent with the second sound) than on incongruent trials (where the first sound was congruent with the second visual stimulus and first visual stimulus was associated with the second sound), as measured by the relative difference in the JNDs. The participants were instructed to make a 'Which came second?' visual TOJ given that previous research suggests that it is only the second auditory stimulus that appears to play a role in the temporal capture of vision (Morein-Zamir, et al., 2003, Experiments 2 and 3; cf. Shore, Spence, & Klein, 2001).

4.2. Methods

Nine paid volunteers (4 male and 5 female) with a mean age of 24 years (range 18-38 years) took part in this study in return of a £5 (UK

Sterling) gift voucher or course credit. This study was conducted in accordance to the Declaration of Helsinki, and had ethical approval from the Department of Experimental Psychology at the University of Oxford.

The participants sat in front of a 21' CRT screen (75Hz refresh rate) flanked by a pair of loudspeaker cones and responded to the stimuli by pressing one of two buttons on a computer keyboard. A personal computer running Matlab v.7.2 with Psychtoolbox v.2.54 (Brainard, 1997; Pelli, 1997) was used to control the presentation of the stimuli and the recording of a participant's responses. The visual stimuli consisted of two light grey circles. Their diameter subtend 5 cm and 2 cm (5.2° vs. 2.1° of visual angle, respectively). The circles were placed 5 cm (5.2°) to the left or right of a central red fixation point against a white background. The auditory stimuli consisted of two sine wave tones (frequency of 300 and 4500Hz) presented for 5 ms each[7] (note that the stimuli used in this study are identical to

[7] In order to test whether participants could readily perceive the pitch of the auditory stimuli, a control study on pitch discrimination was conducted. Twelve participants (4 males, 8 females, mean age 23 years) were asked to rate the pitch of the two 5 ms auditory stimuli with a frequency of 300Hz and 4500Hz by drawing a mark on a 10 cm line representing a scale going from 'low pitch' (left end) to 'high pitch' (right end). The two auditory stimuli were presented in succession with an inter-stimulus interval of

those used by (Gallace & Spence, 2006), in their speeded classification study).

The stimuli were presented approximately 55 cm from the participant's head. Each trial began with the presentation of the central fixation point. The first auditory stimulus was presented after a random interval of between 520 and 910 ms. The first visual stimulus was presented to the left or right of the fixation point after a further 150 ms, and remained on the screen until the end of the trial. The second visual stimulus was subsequently presented on the other side of fixation (after the SOA) and also remained visible until the end of the trial. The onset of the second visual stimulus was followed, 150 ms later, by the presentation of the second auditory stimulus. The participants had to indicate whether the

---

500ms. The participants listened to the stimuli 3 times before rating the pitch. The order of presentation was constant for each participant but was balanced across participants. Each participant made a single rating and the value of zero was assigned to the left end (corresponding to 'low pitch') while the value of ten was assigned to the other end (corresponding to 'high pitch'). The average rating by participants for the 300 Hz stimulus was 2.08 (SE 0.26) whereas the average score for the 4500 Hz stimulus was 7.49 (SE 0.37). A two-tailed repeated measures T-test revealed that participants consistently rated the 300 Hz stimulus as being lower pitched than the 4500 Hz stimulus (two tailed T-stat=-10.372, df=11, $p<.001$). These results therefore demonstrate that participants could discriminate between the frequency of the two tones used in the present study.

second visual stimulus had been presented on the left or right by pressing the left or right arrow key of a computer keyboard while trying to ignore the task-irrelevant auditory stimuli (see Figures 4.1A).

Since previous research (using a speeded classification task) has shown there to be a correspondence between higher-pitched auditory tones (H) and smaller visual images (S) and between lower-pitched tones (L) and larger images (B; see Chapter 3, Experiment 3.3), two types of congruent trial (L-B-S-H and H-S-B-L) and two types of incongruent trial (L-S-B-H and H-B-S-L) were presented in this study (i.e., Experiment 4.1).

The SOA between the two visual stimuli in each condition was varied using the method of constant stimuli (Hegelmaier, 1852). SOAs of ±117, ±78, ±39, ±26, ±13, and 0 ms were used. Negative values indicate that the smaller of the two visual stimuli was presented second while positive values indicate that the larger of the two visual stimuli was presented second. The interval between the first auditory stimulus and the first visual stimulus, and between the second visual stimulus and the second auditory stimulus was 150 ms. In the 0 ms SOA condition, the two visual stimuli appeared

simultaneously while the auditory stimuli preceded and trailed their presentation by 150 ms; in half of the trials the first auditory stimulus was high pitched and the second low pitched, while in the other half the order was inverted. The congruent and incongruent trials were presented equiprobably in a random order in each participant's experimental session, which consisted of 480 trials overall.

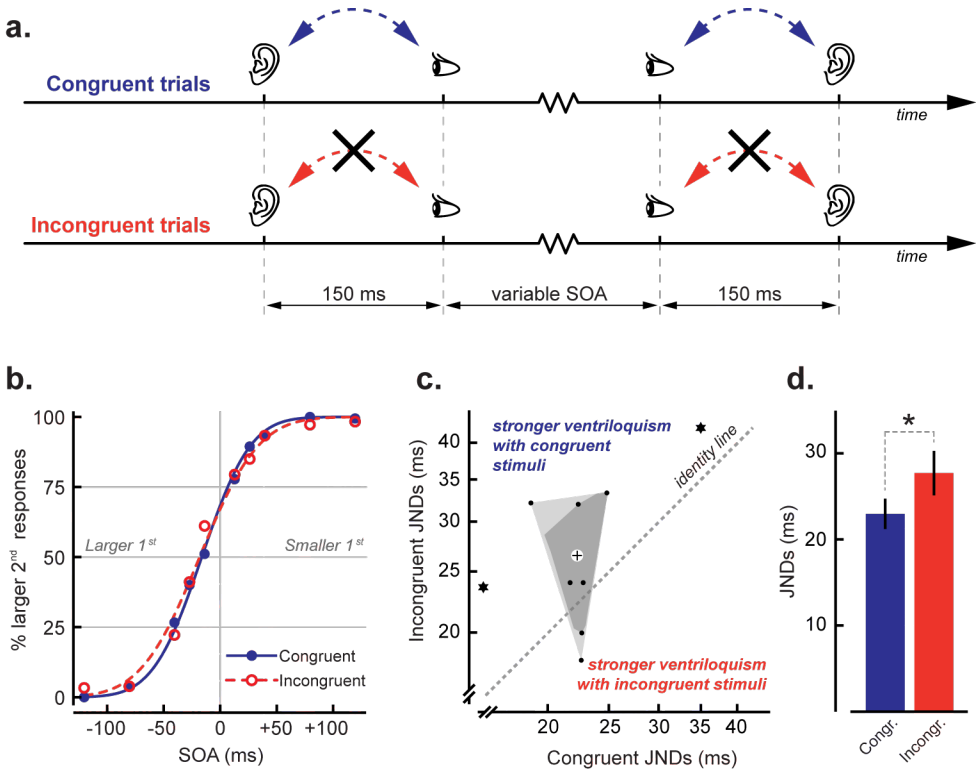


Figure 4.1 Pitch-size congruency modulates the temporal ventriloquist effect. (A) Schematic illustration of the events presented on each trial. Arrows represents the compatibility between visual and auditory stimuli. The SOA between the two visual stimuli was experimentally varied using the method of constant stimuli. (B) Psychometric function describing participants' temporal

order judgment (TOJ) performance on congruent (blue) and incongruent (red) trials. Filled and empty circles represent the proportion of 'larger second' responses for each SOA tested averaged over all participants. (C) Scatter and bagplot of participants' JNDs on congruent vs. incongruent trials. (D) Mean JNDs on congruent and incongruent trials. Error bars indicate the standard error of the mean, asterisk indicate statistical difference ($p<.05$)

4.3. Results

Psychometric functions were calculated for each participant and condition by fitting a cumulative Gaussian function to the percentage of 'larger circle second' responses. Given that congruency is defined in term of the relative order of presentation of the visual stimuli, the 0 ms SOA trials, in which the two visual stimuli were presented simultaneously, cannot be coded as either congruent or incongruent, and hence were excluded from the data analysis. The JNDs were then calculated from each psychometric function by subtracting the value of asynchrony where 75% of 'larger second' responses were made from the asynchrony where participants made 25% 'larger second' responses, and then dividing the result by two. The point of subjective simultaneity (PSS), which indicates the asynchrony at which the participants were maximally uncertain concerning the temporal

order of the stimuli, was also calculated from each function as the asynchrony where 50% 'larger second' responses were made.

A two-tailed paired-samples T-test conducted on the JND data revealed that the JND was significantly smaller for the congruent trials (M = 21 ms) than for the incongruent trials (25 ms; T =-2.367, df=8, p=.045). No such difference was observed when a similar comparison was conducted on the PSS data (-13 vs. -14 ms, respectively; two tailed T- =0.881, df=8, p=.404). Therefore, as predicted, the crossmodal congruency between the auditory and the visual stimuli modulated the sensitivity of participants' TOJ responses while having no effect on the subjective perception of synchrony (see Figure 4.1B). An analysis of the overall (congruent and incongruent) PSS value (-17 ms) conducted with a single sample T-test shows that it differed significantly from zero, indicating a bias toward 'larger second' responses (T=-2.983, df=8, p=.018).

4.4. Discussion

The critical result to emerge from Experiment 4.1 was the significant effect of crossmodal congruency on participants' sensitivity (as measured by the change in their JNDs). The sensitivity of participants' ability to resolve the temporal order in which the two visual stimuli had been presented was significantly higher on the congruent trials than on the incongruent trials. This result therefore provides the first empirical evidence demonstrating that the crossmodal temporal capture of vision by audition can be modulated by the crossmodal congruency between the stimuli, with more pronounced temporal ventriloquism taking place when the auditory and visual stimuli were congruent than when they were incongruent (as reflected by the enhanced auditory capture of vision leading to more sensitive visual TOJs by the participants in the present study). Presumably the auditory stimuli on the congruent trials exerted a stronger attraction on the temporally adjacent visual stimuli, and hence resulted in a larger temporal shift of the visual stimuli toward the time of onset of the congruent auditory stimuli. On a perceptual level, this result implies an illusory

expansion of the delay between the two visual stimuli, thus making it easier for participants to reliably judge their correct temporal order of occurrence, and therefore resulting in enhanced temporal sensitivity on the TOJ task (as highlighted by the lower JNDs in the congruent trials than in the incongruent trials).

The results of the Experiment 4.1 provide the first empirical evidence that the strength of the attraction between (even relatively simple) auditory and visual stimuli in the temporal domain is not fixed. The qualitative features of the stimuli that are presented within each sensory modality (such as the pitch and size of the stimuli in the present study) and the relation between the stimuli presented in each modality (i.e., the crossmodal associations between pitch and size), can modulate the auditory capture of vision, as measured by the difference in the JND between the congruent and incongruent conditions in Experiment 4.1.

The fact that the crossmodal congruency between the auditory and visual stimuli affected the JND but not the PSS rules out any interpretation of the current results in term of response bias (see Spence et

al., 2001), suggesting instead that the modulation of the precision of participants' TOJs observed in this study reflects a genuine perceptual effect (cf. Gallace & Spence, 2006; Long, 1977). Moreover, the present results also provide additional evidence for the claim that crossmodal associations exist between the size of visual stimuli and the pitch of auditory stimuli (see Chapter 3). It is important to note that the crossmodal congruency between the auditory and visual stimuli was completely irrelevant to the participant's task in Experiment 4.1.

In a recent study, Keetels and Vroomen (2011) replicated the present experiment and manipulated the SOA between the visual and the auditory stimuli. They found that when the SOA was set at 150ms (as in the present study), temporal sensitivity was overall lower as compared to a control condition with 0 ms audiovisual SOA. Although in the 150ms SOA condition they replicated the results presented here (i.e., increased sensitivity with congruent stimuli), the fact that there was an overall decrease in performance relative to a baseline led the authors to conclude that crossmodal correspondences do not modulate the temporal

ventriloquist effect. Their argument was based mainly on the original definition of the temporal ventriloquist effect, which is described as an increase in temporal sensitivity with respect to a baseline (Morein-Zamir, et al., 2003). However, in spite of mere lexical arguments, the key point is that crossmodal correspondences do modulate temporal sensitivity, be it proper temporal ventriloquism or not.

According to Keetels and Vroomen (2011), the claim that crossmodal correspondences do not modulate the temporal ventriloquist effect was substantiated by two facts: First that there is no effect of crossmodal correspondences on temporal sensitivity at 75ms SOA between the auditory and visual stimuli, where temporal ventriloquist should be stronger (Morein-Zamir, et al., 2003), and second by the fact that training had no effects on temporal sensitivity. With respect to their first argument, it is unfortunate that Keetels and Vroomen *did not* corroborate their claim with statistical analysis: Indeed, a visual inspection of their plots clearly suggests a significant effect in the expected direction. Moreover, the very small

number of trials used by the authors to fit the psychometric curves makes it hard to draw any meaningful conclusions from their data.

With respect to the second point, it is not obvious why a very brief training on what counts as congruent and what not should change temporal sensitivity. In their training session, congruent and incongruent trials were equiprobably presented, and participants had to judge on the stimulus (in)congruency. It might easily be argued that a standard Pavlovian training session, so successful in inducing cue recruitment and perceptual learning (Di Luca, Ernst, & Backus, 2010; Ernst, 2007; Haijiang, Saunders, Stone, & Backus, 2006), might have led to different results.

Moreover, Keetels and Vroomen (2011), went on saying that the present results might be due to a response bias (e.g., when the first sound is low participants are more likely to say that the large visual stimulus occurred first). However, if this is the case, then a significant modulation of JND due to crossmodal correspondences should occur at any audiovisual SOAs. With this respect, it should be noted that crossmodal correspondences had no effect in the 0ms audiovisual SOA condition, thus

ruling out the response bias account. Nevertheless, in order to silence any potential criticism concerning the use of temporal ventriloquist paradigms to investigate the effects of crossmodal correspondences on temporal perception, in Chapter 5 a new paradigm was devised that overcomes all of the issues highlighted here.

The different speed at which sound and light propagate through air, as well as the different neural processing latencies associated with visual and auditory pathways (King & Palmer, 1985; Spence & Squire, 2003), introduce asynchronies in the time of arrival of visual and auditory information originating from a common event. The ability of our perceptual systems to compensate (at least partially) for such asynchronies is fundamental for multisensory integration and is a necessary condition for the construction of a coherent representation of the external world (see Spence & Squire, 2003). In the audiovisual domain, the outcome of such compensation is the temporal auditory capture of vision, a phenomenon that has been shown to depend on both temporal constraints (Morein-Zamir, et al., 2003) and, under certain conditions, the 'unity assumption'

(Vatakis, et al., 2008; Vatakis & Spence, 2007a). The results of the present study go beyond previous research by demonstrating that some properties of the individual features of audiovisual stimuli, namely crossmodal correspondences, also seem to modulate the temporal ventriloquism effect. Kanai and his colleagues have demonstrated that the crossmodal binding of visual and auditory stimuli is a key factor in the reduction of the perceived asynchrony between auditory and visual signals (Kanai, Sheth, Verstraten, & Shimojo, 2007). In light of this claim, the present results are consistent with the suggestion that more pronounced crossmodal binding takes place between congruent audiovisual stimuli as compared to incongruent stimuli. It is currently an open question as to just how many other phenomena in the field of multisensory perception research might also be modulated by the degree of crossmodal correspondence between the various unimodal stimuli that the experimenters may have happened to incorporate in their study.

5. Spatiotemporal offsets

5.1. Introduction

In the previous chapters, through a review of the literature and a series of experiments, the effects of crossmodal correspondences on sensory processing have been extensively documented. A growing number of studies has now demonstrated that crossmodal correspondences can modulate both the speed of sensory processing (see Chapter 2 and 3) and temporal sensitivity (see Chapter 4). However, there is to date no psychophysical evidence documenting the role of crossmodal congruency for multisensory integration.

Here we investigate the role of crossmodal correspondences on the integration of pairs of temporally (Experiment 5.1 and 5.2) or spatially (Experiment 5.3) conflicting auditory and visual stimuli. When spatiotemporally conflicting stimuli from different modalities are integrated, small offsets are often compensated for, giving rise to the ventriloquist effect, whereby the conflicting stimuli are perceptually 'pulled' together toward a single spatiotemporal onset (Alais & Burr, 2004; Bertelson & Aschersleben, 1998; Morein-Zamir, et al., 2003; Slutsky & Recanzone, 2001). Participants therefore tend to perceive combinations of spatiotemporally conflicting stimuli as unitary multisensory events and become less sensitive to any crossmodal conflicts that may be present (Welch & Warren, 1980). Multisensory integration, in fact, has the cost of hampering the brain's access to the individual sensory components feeding into the integrated percept, thus reducing the reliability of estimates of potential crossmodal conflicts (Ernst, 2005; Hillis, Ernst, Banks, & Landy, 2002). Reliability is defined here as the inverse of the squared discrimination threshold, the just noticeable difference (JND), that is the minimal difference along a given

dimension between a test and a standard stimulus that an observer can detect at a specified level above chance.

Within such a framework, if crossmodal correspondences are exploited by the perceptual system to integrate stimuli from different modalities, the strength of coupling should be higher for congruent combinations of stimuli as compared to incongruent combinations. Therefore, when presented with congruent audiovisual stimuli that are either asynchronous or spatially discrepant, participants' estimates requiring access to such conflicts, such as judgments regarding the relative temporal order or the relative spatial location of the stimuli, should be less reliable (i.e., higher discrimination thresholds for spatiotemporal conflicts) as compared to conditions in which the conflicting stimuli are incongruent.

A similar effect has recently been reported in the temporal domain with audiovisual speech stimuli (human voices and moving lips) presented asynchronously that were either matched (i.e., voices and moving lips belonging to the same person) or mismatched (i.e., voices and moving lips belonging to a different person). When both modalities provide congruent

information, more pronounced multisensory integration takes place, leading to a 'unity effect', which is evidenced behaviorally by an increase of the discrimination thresholds for audiovisual temporal asynchronies (Vatakis, et al., 2008; Vatakis & Spence, 2007a; see also Petrini et al., 2009). Interestingly, it has been shown that the phenomenon disappears when participants are presented with realistic non-speech stimuli, thus suggesting that the 'unity effect' might be specific to speech (Vatakis, et al., 2008; Vatakis & Spence, 2007b; Vroomen & Stekelenburg, 2010; though see Petrini et al., 2009).

An increase of the discrimination thresholds for spatial and temporal conflict when audiovisual stimuli are congruent would provide the first psychophysical evidence that crossmodal correspondences promotes multisensory integration, thus qualifying them as a novel, additional cue to multisensory integration. Moreover, such a result would constitute the first empirical demonstration that the 'unity effect' is not a prerogative of speech stimuli and that it can also occur in the spatial domain. In keeping with previous predictions, participants' estimates regarding both spatial

and temporal conflicts were less reliable with congruent audiovisual stimuli than with incongruent stimuli, thus supporting the claim that crossmodal correspondences promote multisensory integration.

5.2. Experiment 5.1: Temporal conflict – pitch-size

Materials and methods

Twelve participants, with normal vision and audition, made unspeeded audiovisual temporal order judgments (TOJs) regarding which stimulus (i.e., visual or auditory) had been presented second (Shore, et al., 2001).

Visual stimuli consisted of light grey circles presented for 26ms at the centre of a CRT screen against a white background, and subtending 2.1° (small stimulus) or 5.2° (large stimulus) of visual angle at a viewing distance of 55cm. The auditory stimuli consisted of 26ms pure tones, with 5ms linear ramps at on- and off-set and delivered via headphones against background white noise. The frequency of the tones was 300Hz (low pitched) or 4500Hz (high pitched). High and low pitched tones in this and the following experiments were made equally loud for each participant

through an adaptive psychophysical procedure (QUEST, Watson & Pelli, 1983).

A visual and an auditory stimulus were presented on each trial with a variable stimulus onset asynchrony (SOA; ±467, ±333, ±267, ±200, ±133, ±76 and 0 ms, negative values indicate that visual stimulus trailed the auditory stimulus, positive values indicate that visual stimulus led). Each SOA was presented 10 times (20 for the 0 ms SOA) in each condition (i.e., in both the congruent and conditions). The auditory and visual stimuli presented on each trial were equiprobably either congruent along the above-mentioned pitch-size dimension (i.e., a higher-pitched tone was paired with a smaller visual stimulus or a lower-pitched tone was paired with a larger visual stimulus) or else incongruent (i.e., a higher-pitched tone was paired with a larger visual stimulus and a lower-pitched tone was paired with a smaller visual stimulus, see Figure 5.1A). In order to maximize the alternation of congruent and incongruent trials, no more than 2 trials from the same condition were presented in a row. The participants had to perform an unspeeded discrimination task in which they had to indicate the

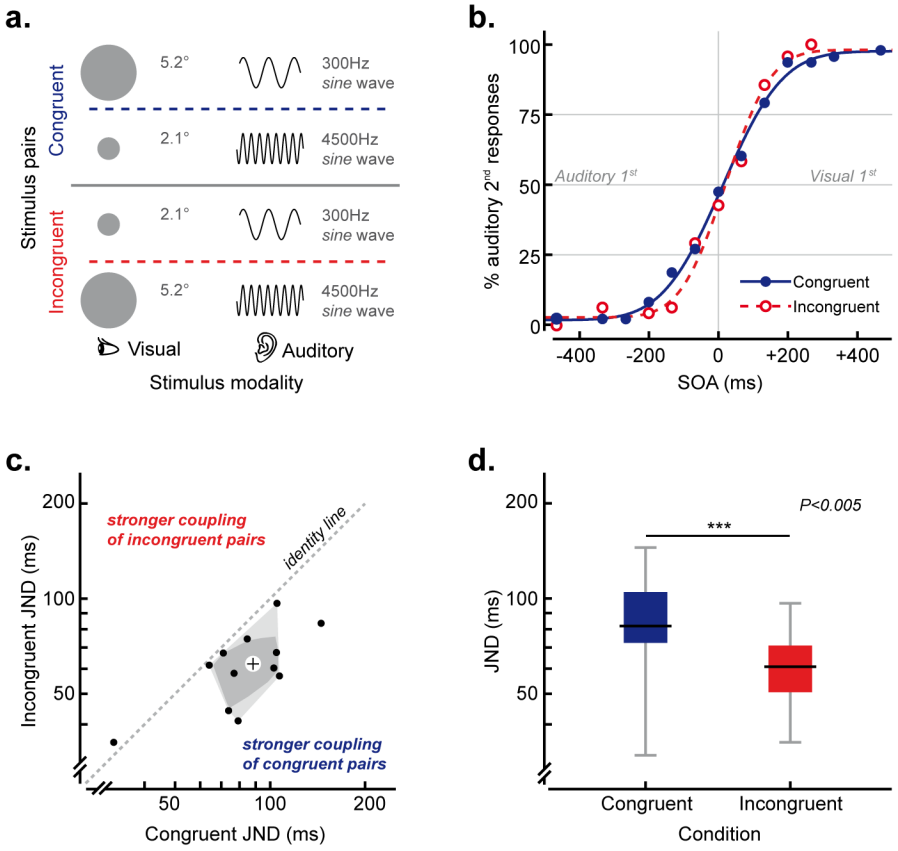


Figure 5.1. Stimuli and results of Experiment 5.1. a. Pairs of auditory and visual stimuli presented in congruent (top) and incongruent trials (bottom). b. Psychometric functions describing performance on congruent (continuous line) and incongruent (dashed line) conditions. Filled and empty circles represent the proportion of 'auditory second' responses for each SOA tested averaged over all participants. c. Scatter and bagplot (Rousseeuw, et al., 1999) of participants' sensitivity (JNDs) on congruent vs. incongruent trials (log-log coordinates). Points below the identity line indicate a stronger coupling of congruent stimuli. The cross at the centre of the bag represents the depth median. d. Sensitivity of participants' responses (JNDs) on congruent and incongruent trials in log scale. The central lines in the boxes represent the median JND, the boxes indicate the first and third quartiles, and the whiskers, the range of the data.

modality of the second stimulus presented on each trial by pressing one of two response keys.

Results

Separate psychometric functions for congruent and incongruent trials were calculated for each participant by fitting the ratios of 'auditory second' responses plotted against SOAs with a cumulative Gaussian distribution (Wichmann & Hill, 2001) (see Fig. 5.1B). The just noticeable differences (JNDs), providing a measure of the reliability (i.e., the discrimination threshold) of participants' TOJs, were calculated for both congruent and incongruent conditions by subtracting the SOA at which participants made 75% 'auditory second' responses from the SOA at which they made 25% 'auditory second' responses and halving the result (see Figure 5.1B-D). Crossmodal congruency had a significant influence on the reliability of participants' estimates (Wilcoxon Signed Rank Test Z=-2.903, p=.004), with smaller JNDs (indicating increased reliability) reported for incongruent trials (median=61ms, interquartile range (IQR)=51-71ms) than for congruent trials (median=82ms, IQR=72-104ms). This result provides

support for the claim that enhanced multisensory integration takes place for congruent as compared to incongruent audiovisual stimulus pairs. Eleven out of the 12 participants tested exhibited less reliable TOJ estimates for congruent as compared to incongruent stimulus pairs (Sign Test, p=.006). Although the PSE data (denoting the point of maximum uncertainty in participants' judgments) do not provide relevant information regarding the strength of coupling (e.g., see Ernst, 2005, 2007) nor the 'unity effect' (e.g., see Vatakis, et al., 2008; Vatakis & Spence, 2007b), statistical outcomes on the effect of crossmodal correspondences on the PSE are reported for completeness: Z=-0.549, p=.583.

5.3. Experiment 5.2: Temporal conflict – pitch/waveform-shape

Materials and methods

The generalizability of the results of Experiment 5.1 was tested in a second experiment (Experiment 5.2) by manipulating the crossmodal correspondence between the auditory features of pitch and waveform and

the visual features of curvilinearity and the magnitude of the angles of regular shapes (see Marks, 1987a; O'Boyle & Tarte, 1980; see Figure 5.2A). The visual stimuli consisted of black 7-pointed stars presented for 26ms against a white background and subtending 5.2° of visual angle. One star was curvilinear and had a ratio of inscribed to circumscribed circles of 0.65, whereas the other star was angular and had a ratio of inscribed to circumscribed circles of 0.55. The auditory stimuli, delivered via headphones against background white noise, consisted of 26ms tones with 5ms linear ramps at on- and off-set. One auditory stimulus consisted of a high pitched (1760Hz), square waved tone, whereas the other had lower frequency (440Hz) and sinusoidal wave.

The experimental procedure was the same as Experiment 5.1, with the exception that the compatible stimulus combination here consisted of the presentation of the pointed star together with the higher pitched tone and the curvilinear star with the lower pitched tone. Conversely the incompatible stimulus pairs consisted of the pointed star coupled with the lower pitched tone and the curvilinear star with the higher tone.

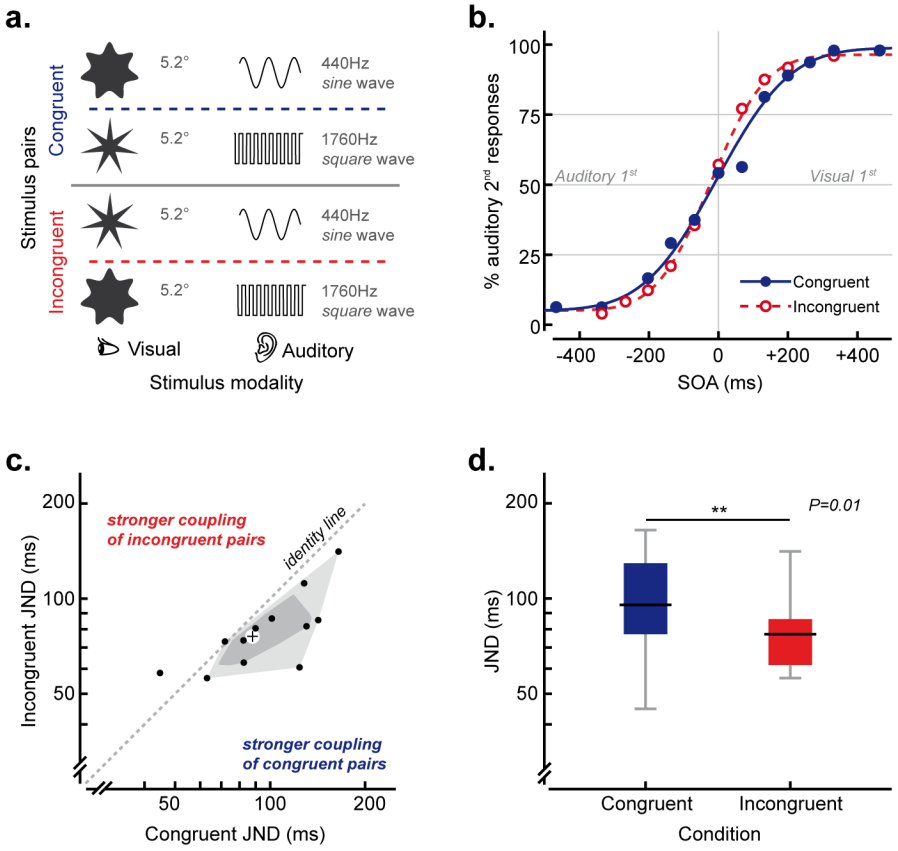


Figure 5.2. Stimuli and results of Experiment 5.2. **a.** Pairs of auditory and visual stimuli. **b.** Psychometric functions describing performance on congruent (continuous line) and incongruent (dashed line) conditions. **c.** Bagplot (Rousseeuw, et al., 1999) of participants' sensitivity (JNDs) on congruent vs. incongruent trials. **d.** Participants' sensitivity (JNDs), on congruent and incongruent trials.

Results

JNDs (calculated with the procedure described in Experiment 5.1) were again significantly higher on the congruent trials (median=95ms, IQR=77-129ms) than on the incongruent trials (median=77ms, IQR=61-

86ms, Wilcoxon-Test Z=-2.589, p=.010), with 10 out of 12 of the participants tested exhibiting higher discrimination thresholds in the congruent as compared to the incongruent condition (Sign Test, p=.039, see Figure 5.2B-D). No significant effect of condition was found in the PSE data (Z=.893, p=.343).

5.4. Experiment 5.3: Spatial conflict

Materials and methods

Twelve participants, with normal vision and audition, made unspeeded judgments as to whether an auditory stimulus was presented to either the left or the right of a visual stimulus. The visual stimuli consisted of white Gaussian blobs projected for 200ms against a black background on a fine fabric screen (width: 107.7cm; height: 80.8cm) (see Figure 5.3). The standard deviation of the Gaussian luminance profile of the blobs subtended 0.26° (small stimulus) or 2.3° (large stimulus) of visual angle at a viewing distance of 110.5cm (a chinrest was used to control the head position). The auditory stimuli consisted of 200ms pure tones with 5ms linear ramps at on-

and off-set; the frequency of the tones was 300Hz (low pitched) or 4500Hz (high pitched, see Figure 5.4A). In order to provide richer spectral cues for auditory localization, the tones were convolved with white noise (King & Oldfield, 1997) and their intensity was modulated with a sinusoidal profile with a frequency of 50Hz. The auditory stimuli were delivered from one of four loudspeaker placed behind the fabric screen (placed 5.2cm and 15.6cm to the left and the right of the midline of the screen) and their intensity was randomly jittered from trial to trial (between ±1% of the standard intensity) in order to avoid participants using any potential slight differences in the intensities of the sounds delivered by the 4 loudspeaker as auxiliary cues for sound source localization. Furthermore, in order to mask the noise emitted by the relay that gated the signal to the relevant speaker, white noise was delivered by an additional pair of loudspeaker placed behind the screen throughout the whole experimental session.

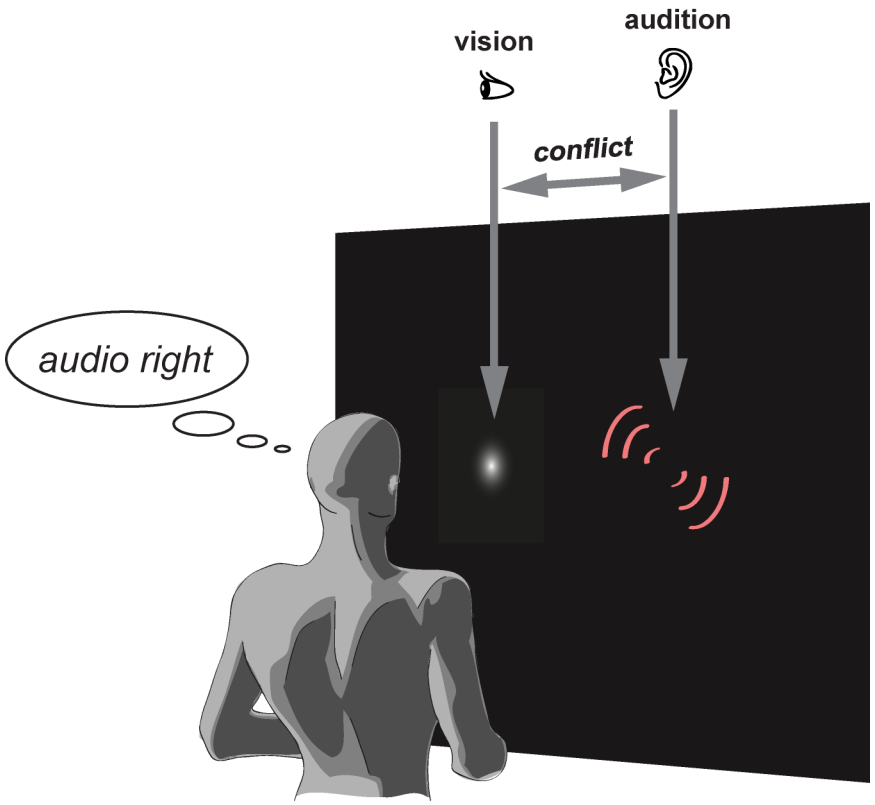


Figure 5.3. Schematic illustration of the participant and experimental apparatus for Experiment 5.3. Visual stimuli (blobs) were projected on a white sound-transparent screen placed in front of an array of 8 loudspeakers. This apparatus allowed for the introduction of physical conflicts between the spatial location of the visual stimuli and the auditory ones

A train of 3 synchronous audiovisual events, with an interstimulus interval randomized between 150ms and 300ms, was presented on each trial with the source of the auditory stimulus randomly located to the left or the right of visual stimulus with the magnitude of the azimuthal displacement determined using an adaptive psychophysical procedure. At the beginning

of the experiment a psychometric function was fitted over a small set of hypothetical data points. In particular, it was assumed that participants correctly responded 'left' or 'right' in 4 trials in which the auditory stimulus was placed 9.7° and 4.9° to the left or the right of the visual stimulus and fitted a cumulative Gaussian curve over these four points. Then, after each response, the curve was fitted again with the newly-collected data and the auditory stimulus that was presented on the next trial was randomly placed to the left or the right of the visual stimulus with a displacement normally randomized around 1 JND (s.d. 1 JND). This procedure was selected after preliminary results which indicated high variability in participants' ability to localize sounds, making it hard to select in advance an efficient placement of the stimuli (as required by the method of limits (Dixon & Mood, 1948) and the method of constant stimuli (Watson & Fitzhugh, 1990)). Moreover, this procedure maximises the information provided by each response by placing the stimuli around the regions that are more relevant to calculate the JND. In order to train participant to localize sounds, before running the experiment, they were required to perform a quick task (96 trials) where a sound was emitted by

one of 8 loudspeakers placed behind the screen (4 to the left and 4 to the right of the vertical midline) and they had to determine whether it was coming from the left or the right of the screen's midline (visual feedback was provided after incorrect responses in the training block).

The auditory and visual stimuli presented on each trial were equiprobably either congruent along the pitch-size dimension (i.e., a higher-pitched tone was paired with a smaller visual stimulus or a lower-pitched tone was paired with a larger visual stimulus) or incongruent (i.e., a higher-pitched tone was paired with a larger visual stimulus and a lower-pitched tone was paired with a smaller visual stimulus, see Figure 5.4A). Two hundred and eighty trials were presented on each session (140 congruent and 140 incongruent). Participants performed an unspeeded discrimination task in which they had to press either the left or the right key of a computer mouse in order to indicate whether the auditory stimulus was coming from the left or the right of the visual stimulus.

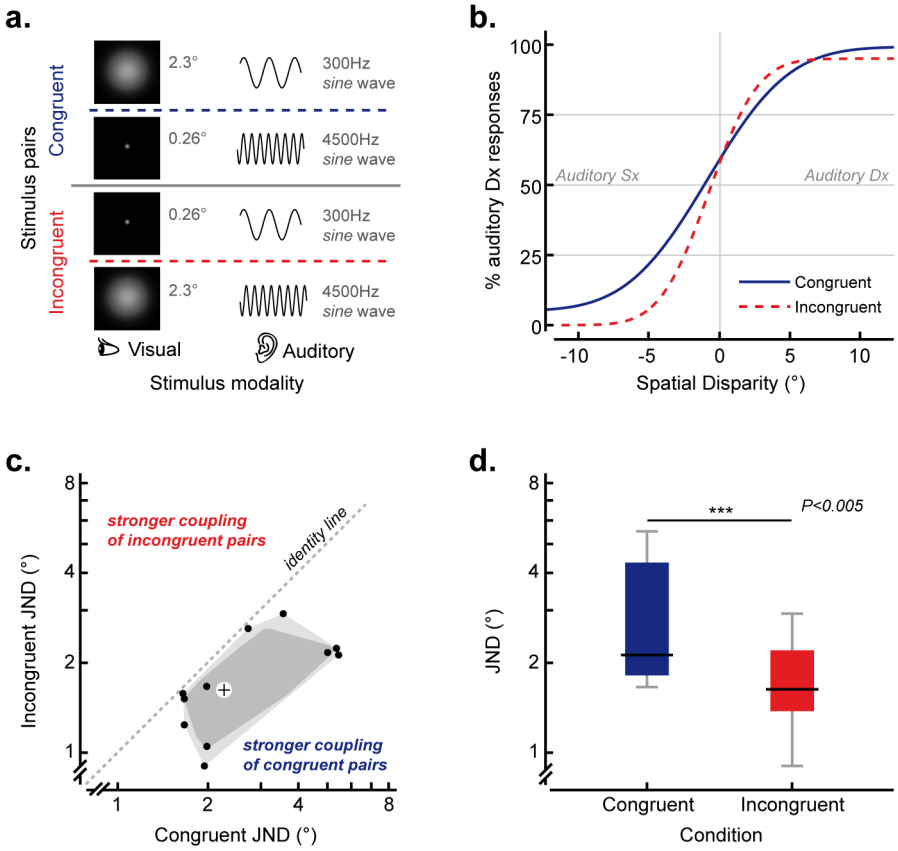


Figure 5.4. Stimuli and results of Experiment 5.3. **a.** Pairs of auditory and visual stimuli. **b.** Psychometric functions describing performance on congruent (continuous line) and incongruent (dashed line) conditions. **c.** Bagplot (Rousseeuw, et al., 1999) of participants' sensitivity (JNDs) on congruent vs. incongruent trials. **d.** Participants' sensitivity (JNDs), on congruent and incongruent trials.

Results

Separate psychometric functions were calculated for congruent and incongruent trials for each participant by fitting the ratios of 'auditory right' responses plotted against spatial displacement (measured in degrees

of visual angle, with negative values indicating that the auditory stimulus was placed to the left of the visual one) with a cumulative Gaussian distribution (Wichmann & Hill, 2001; see Figure 5.4B). Crossmodal congruency significantly influenced the reliability of participants' estimates (Wilcoxon Signed Rank Test Z=-3.059, p=.002), with smaller discrimination thresholds reported for incongruent trials (mean=1.7°, IQR=0.9°) than for congruent trials (median=2.2°, IQR=2.6°), thus providing support for the claim that enhanced multisensory integration takes place for congruent as compared to incongruent pairs of audiovisual stimuli. All of the participants exhibited lower discrimination thresholds in response to spatial conflicts between congruent as compared to incongruent stimulus pairs (Sign Test, p<.001, see Figure 5.4C-D). Interestingly, congruency also had a significant effect on the PSE data in this experiment: Z=-2.432, p=.015.

5.5. Discussion

The results of the three experiments reported in the present Chapter demonstrate that crossmodal correspondences affect multisensory integration, as assessed by their effect on the reliability of participants' audiovisual TOJs and spatial localization judgments. In particular, estimates requiring access to temporal (Experiments 5.1 & 5.2) and spatial (Experiment 5.3) conflicts between congruent auditory and visual stimuli were found to be less reliable (i.e., higher discrimination thresholds) than those requiring access to conflicts between incongruent stimuli. A reduced reliability of the estimates requiring access to intersensory conflicts reflects the cost of multisensory integration and is the marker of a stronger coupling between the unisensory signals (Ernst, 2005, 2007; Hillis, et al., 2002). These results therefore, go beyond the results of Chapter 4, and indicate a stronger coupling of congruent stimuli as compared to incongruent stimuli and provide the first psychophysical evidence that crossmodal correspondences can actually promote multisensory integration. However, it should be noted that the crossmodal correspondences studied

here (as well as in many other studies, see Bernstein & Edelstein, 1971; Gallace & Spence, 2006; Marks, 1987a, 1989; Martino & Marks, 2000; R. D. Melara & O'Brien, 1987; Parise & Spence, 2008; P. Walker & Smith, 1984, 1985) are likely relative rather than absolute, depending on the particular range of stimuli used (see Chapter 2). What is called a 'big' circle, in fact, would most likely behave like a small circle if we happened to pair it with an even larger circle and the same argument would apply, mutatis mutandis, to any other potential stimulus features that happen to be considered (see Marks, 1989, on this issue).

Considering that the unimodal signals used in Experiments 5.1-5.3 were identical in both congruent and incongruent conditions (i.e., same signal reliability in both conditions), the difference in the strength of coupling reported here should be attributed to the knowledge that the participants' perceptual systems had internalized about which stimuli 'belong together' (or, rather, which normally co-occur) and should therefore be integrated. According to Bayesian integration models, such prior knowledge about stimulus mapping, the coupling prior, determines the

strength of the coupling between the stimuli proportionally to its reliability (with reliability defined as the inverse of the squared variance of the coupling prior distribution), that is, the more the system is certain that two stimuli belong together (i.e., the smaller the variance of the coupling prior), the stronger such stimuli will be coupled (Ernst, 2005, 2007). The effect of crossmodal correspondences in multisensory integration could, therefore, be interpreted in terms of differences in the variance of the coupling prior (i.e., smaller variance for congruent stimulus pairs than for incongruent pairs), that is to say that crossmodal correspondences determine the strength of coupling by modulating the variance of the coupling prior distribution (see Chapter 7 for a more detailed discussion).

It should, however, be noted that the present results might also be accounted for by the possibility that crossmodal correspondences modulate the tuning of multisensory spatio-temporal filters (see Burr, Silva, Cicchini, Banks, & Morrone, 2009). The early stages of sensory processing have, in fact, traditionally been modelled in terms of spatial and temporal filters operating upon the incoming sensory information (e.g. see de Lange Dzn,

1954; J. Robson, 1966). Their role, in a crossmodal setting, would be critical to determining the perceived temporal simultaneity and spatial coincidence of multisensory signals (Burr, et al., 2009). Crossmodal correspondences might act on those filters by increasing their spatial and temporal constants under conditions of congruent crossmodal stimulation and by reducing such constants when the stimuli are incongruent. In keeping with the data reported here, a similar modulation of the tuning of the multisensory spatio-temporal filters could also determine larger windows of both subjective simultaneity and spatial coincidence for congruent as compared to incongruent pairs of audiovisual stimuli.

The results of Experiments 5.1 and 5.2 also extend the finding of previous research on the 'unity effect' by showing that an increase of the discrimination threshold for temporal asynchronies is not specific to matched audiovisual speech events: crossmodal correspondences can also trigger robust unity effects. Vatakis and her colleagues (Vatakis, et al., 2008; Vatakis & Spence, 2007b) have conducted a number of studies on the integration of asynchronous but ecologically-valid audiovisual stimuli and

consistently found that the 'unity effect' is restricted to speech stimuli, thus concluding that speech is 'special' inasmuch as the facilitatory effect on multisensory integration leading to the unity effect is specific to speech. The results of Experiments 5.1 and 5.2, therefore, not only extend the class of stimuli that are known to lead to a unity effect, but also suggest the hypothesis that crossmodal correspondences might also be 'special' (or rather that audiovisual speech stimuli may not be so special, or unique, after all; Vroomen & Stekelenburg, 2010). In addition, the results of Experiment 5.3, showing that participants' discrimination thresholds for the spatial separation between auditory and visual stimuli are increased when the stimuli are synesthetically congruent, constitutes the first experimental evidence that the unity effect also occurs in the spatial domain, and thus provides additional evidence for the claim that the unity effect results from more pronounced multisensory integration.

While research has tended to focus on the spatiotemporal constraints of multisensory integration over the past 25 years (Calvert, Spence, & Stein, 2004), the results reported here demonstrate that crossmodal

correspondences provides an additional constraint on such processes. It should be noted that in the experiments described so far, the focus was mainly on statistical crossmodal correspondences (see Chapter 2), that is those correspondences reflecting an internalization of the natural correlation between the properties of the sensory signals. Nevertheless, signals' correlations can also be contingent, and not relying on previous learning. This is the case, for example, of the correlation between the temporal structures of two or more simultaneous signals. What is the role of contingent signals' correlation on multisensory integration? In the next chapter, I will try to answer this important question by directly manipulating the correlation between auditory and visual signals and measuring their effects on simple spatial task aimed at measuring the optimality of multisensory integration.

6. Temporal correlation

6.1. Introduction

Multisensory signals originating from the same distal event are often similar in nature. Think of fireworks on New Year's Eve, an object falling and bouncing on the floor, or the footsteps of a person walking down the street. The temporal structures of the visual and auditory events are always almost overlapping, and we often effortlessly assume an underlying unity between our visual and auditory experiences. In fact, the similarity in temporal structure of unisensory signals provides a potentially powerful cue with which the brain can determine whether or not multiple sensory signals have a common origin.

One measure that is commonly used in signal processing to quantify the similarity between two signals as a function of their time lag is cross-correlation; the higher the cross-correlation between two signals at a given time lag, the higher their similarity. Notably, cross-correlation (hereafter simply referred to as 'correlation') is an important cue for humans when attempting to solve the correspondence problem in stereo vision (Tyler & Julesz, 1978). In the present study, the temporal correlation between auditory and visual signals was manipulated experimentally in order to assess the strength of multisensory integration (MSI): If correlation promotes MSI, sensory fusion should occur only when the auditory and visual stimuli are temporally correlated.

Previous research has demonstrated that when multisensory stimuli are optimally integrated, the resulting percept ($\hat{S}$) is a weighed average of the individual sensory estimates ($\hat{S}_i$) with weights (w_i) proportional to their precision (Alais & Burr, 2004; Ernst & Banks, 2002). If the noise of the individual sensory estimates (σ_i^2) is independent and Gaussian

distributed, the maximum-likelihood estimate (MLE) of a physical property is:

$$\hat{S} = \sum_i w_i \hat{S}_i \qquad \text{(eqn. 6.1)},$$

where

$$w_i = \frac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2} \qquad \text{(eqn. 6.2)},$$

and σ_i^2 is the variance of a sensory estimate $\hat{S}_i$ (see Figure 6.1). Notably, if unimodal sensory cues are integrated according to MLE, the resulting sensory estimate should also be more precise than either of the individual sensory estimates, and its variance is given by

$$\sigma_{ij}^2 = \frac{\sigma_i^2 \sigma_j^2}{\sigma_i^2 + \sigma_j^2} \qquad \text{(eqn. 6.3)}.$$

MLE therefore allows one to make clear predictions concerning the combined estimates, hence providing a powerful benchmark against which to test for the optimality of multisensory integration.

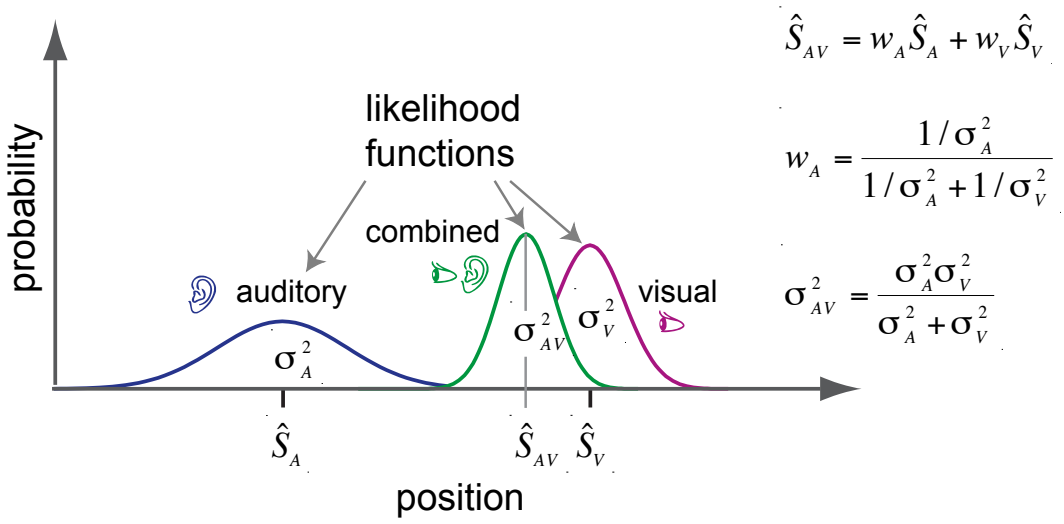


Figure 6.1. The Maximum Likelihood Estimation model.

To measure the effects of temporal correlation on the integration of auditory and visual spatial cues, observers were presented with streams of visual, auditory, and combined audiovisual stimuli that were sometimes correlated (Figure 6.2A). These stimuli, consisting of trains of white noise clicks and/or Gaussian blobs, were presented from a variety of different spatial locations within a large 2D display. Observers were instructed to point to the perceived location of the stimuli on the screen by moving a cursor controlled by a graphic tablet (Figure 6.2B). Note that in the bimodal trials, the position of the auditory and visual stimuli always coincided, and participants were explicitly informed of this fact. If MSI is

modulated by temporal correlation, one would expect observers to optimally integrate multisensory spatial cues when the stimuli are correlated, and otherwise to integrate the cues suboptimally. Given that participants were instructed to point to the perceived location of the stimuli on a 2D display, the present paradigm allowed us to simultaneously test for the effects of stimulus correlation on MSI in both the horizontal and vertical dimensions.

6.2. Methods

Nine naïve observers and C.P. took part in the experiment. All of the participants had normal or corrected-to-normal vision and audition. The study was conducted in accordance to the Declaration of Helsinki. The experiment consisted of two two-hour sessions conducted on consecutive days. Participants provided written informed consent and received 6 euros/hour in return for taking part in the study. The study was conducted in accordance with the Declaration of Helsinki and had ethical approval from the University of Tübingen.

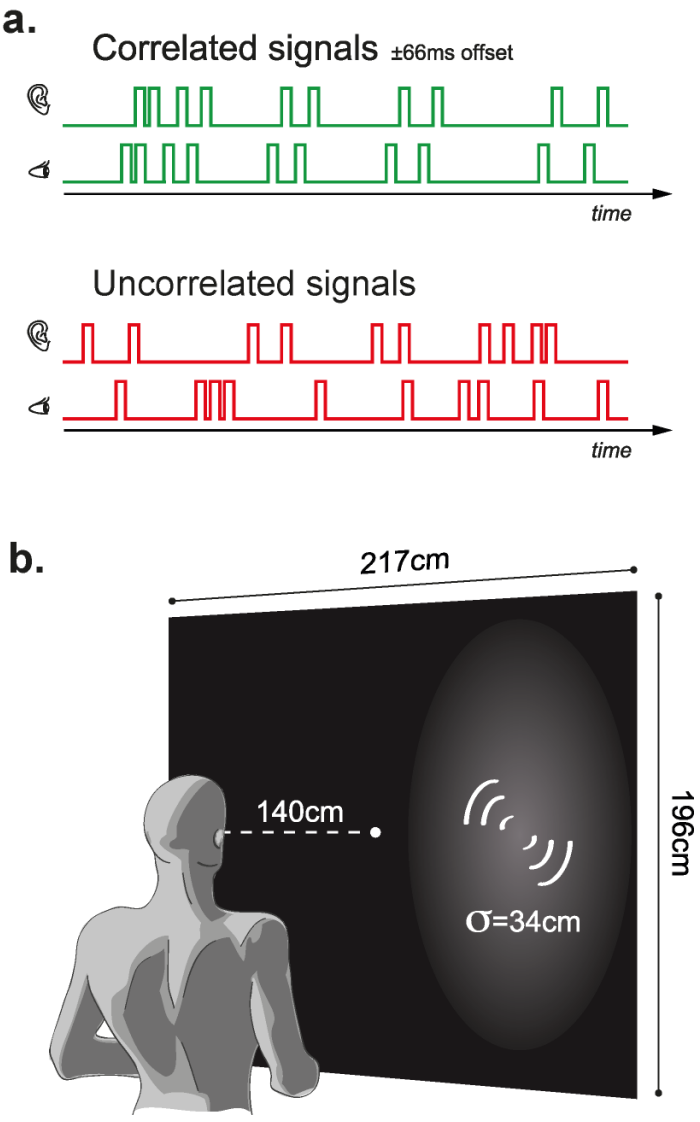


Figure 6.2 Stimuli and apparatus for Experiment 6.1. **a.** Examples of the intensity profiles of correlated (correlation 1, lag =±66ms) and uncorrelated audiovisual stimulus pairs. Auditory stimuli consisted of trains of 10 white noise bursts. The overall duration of visual and auditory stimuli was 2s. **b.** Schematic representation of the experimental apparatus.

The visual stimuli consisted of trains of flashes of large, low contrast (30%) Gaussian blobs (sigma=34cm) back-projected against a black background on a matte plexiglass screen (size 217x196cm). The auditory stimuli consisted of trains of white noise clicks delivered through earbuds. The participants were seated 140cm from the screen. The participant's head was constrained by a chinrest and a headrest with the head fixed by a temple-clamp. The participants were instructed to localize the visual and auditory stimuli on the screen by moving a red cursor controlled by a pen and a graphic tablet.

Each train of visual and auditory stimuli consisted of ten flashes or ten clicks, respectively, randomly scattered over a 2s temporal interval. Each click and blob lasted for 16ms. A new temporal structure (i.e., train of flashes and/or clicks) was generated for each trial. In the bimodal trials, where both visual and auditory stimuli were presented, the temporal structure was either identical in the two modalities (correlated trials) or random (uncorrelated trials). In the case of the correlated bimodal trials, a delay of ±66ms (pseudorandomly vision- or audition-lead, in order to avoid

temporal adaptation and recalibration, see Fujisaki, Shimojo, Kashino, & Nishida, 2004) was introduced between the visual and auditory stimuli. This value corresponded to the average temporal gap between the closest neighboring visual and auditory points in the uncorrelated condition, so as to equate the mean audiovisual delay present in the uncorrelated trials.

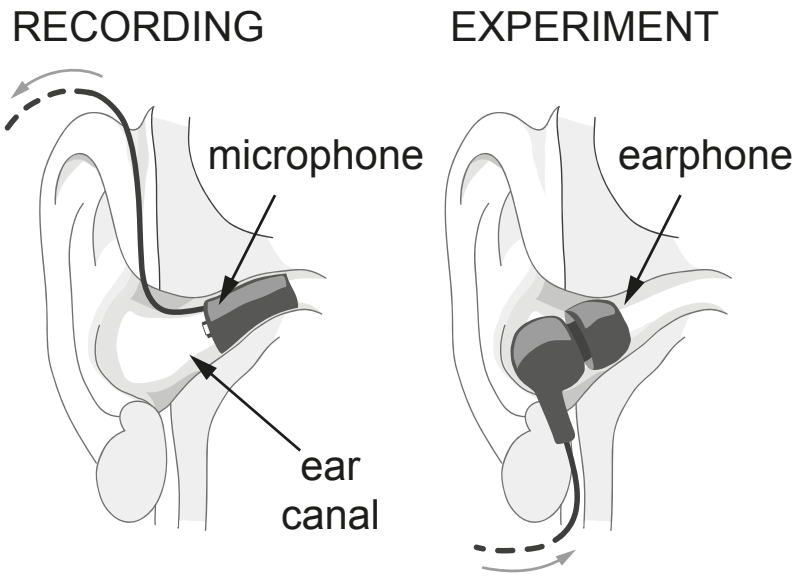


Figure 6.3. Recording and presentation of the auditory stimuli. In a preliminary phase, auditory stimuli were played from loudspeakers placed in various spatial locations and recorder from two microphones placed inside observers' ear canals (left panel). During the experiment, the stimuli recorded in the preliminary phase were played directly into observers' ear canals, thus preserving the spatial cues of the original signals recorded in the preliminary phase (right panel).

On each trial, the visual and auditory stimuli were presented pseudo-randomly from one of 25 spatial positions arranged along 5x5 two-dimensional grid (87.8 x 87.8 cm) aligned with observers' line of sight. The experimental trials consisted of the presentation of stimuli coming from one of the nine central positions in the grid. The external positions, closer to the edges of the screen, were included in order to broaden the stimulus space, thus increasing the positional uncertainty, and hence reduce any bias by participants to point straight ahead.

In order to create compelling spatialized sounds, in a preliminary session the raw auditory stimuli were played from a loudspeaker from each of the 25 spatial positions, and recorded the clicks with a pair of custom-built miniature microphones placed inside the left and the right ear canals of the blindfolded participants (see Figure 6.3). Tailored auditory stimuli were provided by performing this procedure individually for each participant. This ensured that the clicks were filtered by the individual's head-related transfer function (HRTF), hence providing rich and ecological cues for sound-localization. In order to have a large stimulus set and

minimize the effect of potential artifacts in the recording procedure, eight clicks were recorded from each spatial position for each participant. On each trial, a new train of clicks was generated by randomly sampling with replacement from the set of stimuli recorded from the relevant spatial position.

Before starting the main experiment, observers completed a sound localization training session to familiarize themselves with the task and the auditory stimuli. Each trial in this training session started with the presentation of 5 audiovisual stimuli, presented from 5 different spatial positions, followed by a 2s stimulus presentation interval containing 10 clicks played from a single spatial location. Participants pointed to the source of the auditory stimuli. After responding, a green dot appeared on the screen, marking the actual position from which the auditory stimuli had been presented. The participants were instructed to move the cursor to the position marked by the green dot and press a key; at that point, the auditory stimuli presented before were played again. At the same time, a series of blobs (sigma=6cm) were flashed in synchrony with the sounds.

Notably, while aiming for the green dot, a thin gray line, representing the pointing error, connected the perceived position with the moving cursor. The training session consisted of 150 trials and took place in the first experimental session right after the recording of the auditory stimuli.

In the main experiment, the participants were instructed to point to the perceived location of the visual and/or the auditory stimuli, guessing if necessary. In the case of the visual stimuli, they were instructed to point to the center of the blobs. When they were happy with the position of the cursor, participants had to press the spacebar of a computer keyboard to submit their response.

During the experiment, unimodal and bimodal trials were presented in separate blocks, preceded by an instruction screen informing the participants about the stimulus type. Auditory and visual trials alternated pseudorandomly in the unimodal blocks; correlated and uncorrelated audiovisual stimuli alternated pseudorandomly in the bimodal trials. Participants were not informed about the stimulus (un)correlation, but they were told that visual and auditory stimuli would always be presented from

the very same spatial position. Each block consisted of 25 trials, and unimodal and bimodal blocks alternated during the experiment[8].

6.3. Results

In a pointing task, two types of error can affect observers' responses: A systematic error, arising from sensorimotor biases, and a random error, reflecting the noise in the sensorimotor system. In the present analysis, these two error components were separated in order to estimate the average endpoint location in relation to the one stimulated (bias) and the precision of observers' pointing responses to this average endpoint location: For each participant and stimulus position, the average endpoint location was calculated as the median pointing movements (in both the vertical and the

[8] In order to engage participants with the experiment, the whole task was presented as a shooting video-game: A bullet-hole graphic effect (spatially aligned with the pointing response), and the sound of a gunshot accompanied each response, closely followed by the sound of a loading gun. Note that the sound effects were not spatialized (mono), and participants perceived those sounds as coming from 'within their head'. To avoid interference of those effects with the experimental stimuli, a temporal interval randomized between 2s and 3s separated two consecutive trials. In order to further motivate the participants, they were told that they could get points as a function of their performance. At the end of each block, a fake high score list was presented in which participants on average ranked third out of ten.

horizontal dimension), and subtracted it from individual pointing responses. This difference (residual) constituted the random component.

For each observer, four datapoints were collected for each condition and spatial position, and the data from all participants was jointly analyzed. This procedure was required to average out from the data the noise introduced by small miscalibrations in the position of the participants' head, and other potential individual differences (e.g., different sensitivity of the two ears; Lewald, 1997; Lewald & Ehrenstein, 1998). Responses slower than 2 standard deviations from the (log transformed) average RT (6.5s) were excluded from statistical analysis (this resulted in the removal of less than 4% of the data).

The precision of participants' pointing was calculated as the variance of the random errors across all participants (residual). Given that the sampling distribution of the variance follows a Chi-squared probability distribution, statistical inferences on the random errors were based on the well-known variance's sampling distribution (Sokal & Rohlf, 1995, p. 155). In the unimodal conditions, the precision of participants' pointing responses

in the visual trials was higher than in the auditory trials in both the vertical ($\chi^2(325)=110.65$, $p<.001$), and horizontal dimensions ($\chi^2(325)=133.96$, $p<.001$; see Figure 6.4). As expected, in the bimodal condition, the precision of the correlated trials was higher than the uncorrelated trials in both the horizontal ($\chi^2(313)=366.20$, $p=.020$), and vertical dimension ($\chi^2(313)=361.54$, $p=.030$).

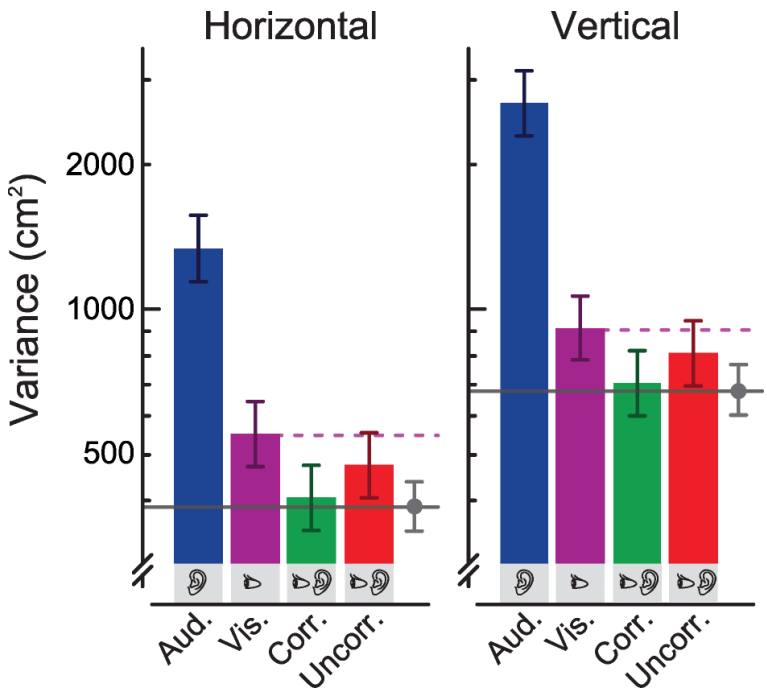


Figure 6.4 Variance of the pointing responses in the four conditions in the horizontal (left) and vertical dimension (right). Error bars indicate the 95% confidence intervals. The gray dot on the right of each panels indicate the MLE prediction from Eqn 3. The dashed line represents the variance of the most precise unimodal condition (i.e., vision).

To test for the optimality of MSI (the main purpose of the present analysis), the precision of the bimodal trials was compared to the MLE prediction of pointing precision computed from Eqn. 6.3. Interestingly, the variance of the correlated trials closely approached statistical optimality (i.e., the maximum expected benefit from MSI), with the empirical variance not differing from the MLE predictions in both the horizontal ($\chi^2(315)=326.85$, $p=.68$) and vertical dimensions ($\chi^2(315)=324.59$, $p=.66$). Conversely, uncorrelated trials were significantly less precise than predicted by MLE in both the horizontal ($\chi^2(313)=382.18$, $p=.0057$) and vertical dimensions ($\chi^2(313)=372.55$, $p=.0116$). Taken together, these results therefore demonstrate that MSI is optimal in the correlated but not in the uncorrelated condition.

To confirm statistically optimal integration over a winner-take-all strategy it is important to show that combined performance was better than either of the unimodal conditions. In keeping with the hypothesis, visual trials were significantly less precise than correlated bimodal trials in

both the horizontal ($\chi^2(325)=442.31$, $p<.001$) and vertical dimensions ($\chi^2(325)=422.78$, $p<.001$).

To test whether the uncorrelated condition was nevertheless still better than the best of the unimodal estimates, and hence to provide evidence of suboptimal MSI, the precision of the uncorrelated condition was compared to the visual-only condition. Visual-only pointing responses were overall less precise than uncorrelated bimodal pointing responses in both the horizontal ($\chi^2(325)=378.06$, $p=.0226$), and vertical dimensions, though in the vertical dimension this difference only just approached statistical significance ($\chi^2(325)=366.01$, $p=.058$).

It is important to note that the random errors in the vertical and horizontal dimensions were not correlated (Pearson's $\rho= -0.008$, $p =0.7762$), hence the error data from the vertical and the horizontal dimension provide two statistically independent measures of the effect of stimulus correlation on MSI. The fact that the same pattern of results were observed in both the vertical and horizontal pointing variance provides converging evidence that

MSI is optimal when audiovisual stimuli are correlated in time, and suboptimal otherwise.

To further test the robustness of the present results, the analyses were repeated applying a leave-one-out Jackknife technique, consisting of re-computing the variance of the pointing responses excluding one participant at a time (Efron, 1982). This procedure allows one to assess the importance of each observer in the measured effects, hence verifying whether or not the present results are due to an outlying participant biasing the overall estimates. Notably, the precision in the correlated condition was again always found to be higher than both the uncorrelated condition and the best unimodal condition (i.e., vision) in both the horizontal and the vertical dimension, hence substantiating the robustness of the results reported in this chapter.

In accordance with previous studies, unimodal pointing responses revealed a uniform compression of space in vision and a vertical compression and horizontal expansion of space in audition (Grüsser, 1983;

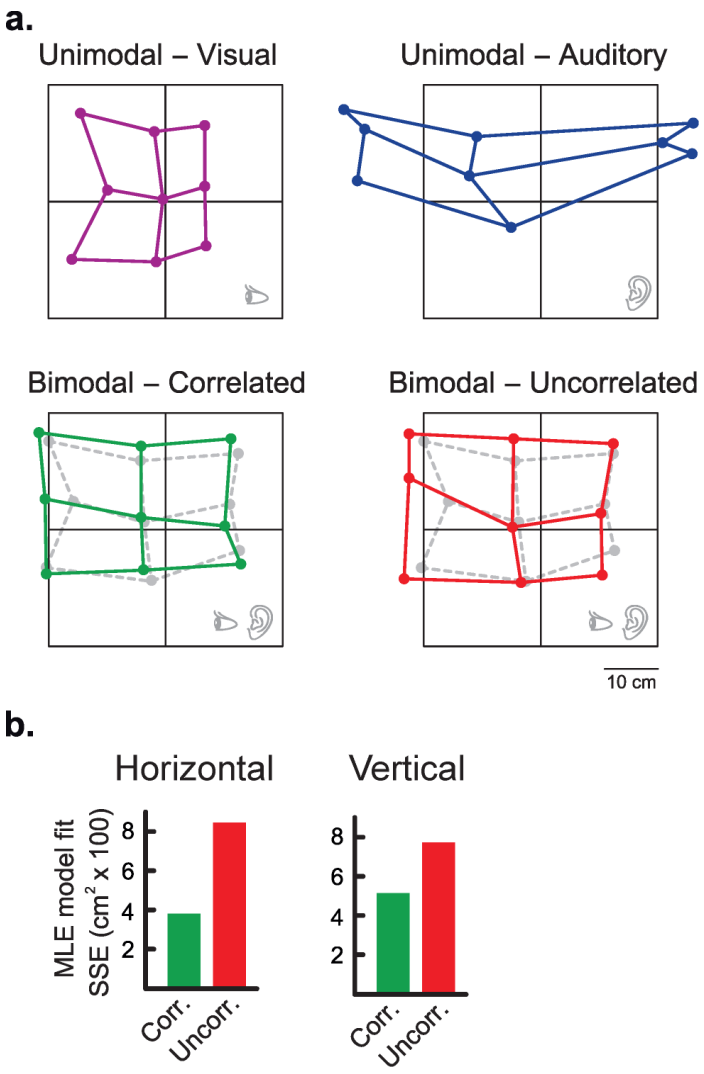


Figure 6.5. Average endpoints of pointing responses in the four conditions in Experiment 6.1. **a.** The filled points correspond to the average end points. The black thin grid represents the actual position of the stimuli. In the lower two panels (bimodal condition), the dashed lines and gray dots represent the MLE prediction from Eqn 6.1. **b.** Goodness of fit (sum of squared error, SSE) of MLE prediction (Eqn. 6.1) to the empirical average endpoint of pointing responses in the horizontal (left) and vertical dimension (right).

Lewald & Ehrenstein, 1998; see Figure 6.5). Such a result is not surprising, given that it is well known that inaccuracies and distortions in perception can often be induced by prior knowledge (Knill and Richards, 1996). It should be noted that these spatial distortions do not undermine the analysis of the precision of participants' responses, nor any conclusion regarding whether or not there is optimal integration.

On top of predictions regarding pointing precision, MLE allows also for a prediction of bimodal perceptual space. Taking into account the average endpoints and the precision of unimodal auditory and visual pointing movements for each stimulus position, Eqn. 6.1 provides a prediction of the bimodal endpoints. Overall, the bimodal perceptual space predicted by the MLE qualitatively agreed with the empirical data in both dimensions and in both conditions. A quantitative analysis, however, revealed that the perceptual space predicted by the MLE provided a better fit to the empirical data in the correlated than in the uncorrelated condition. Indeed, the sum of squared errors (SSE) between empirical data

and MLE prediction in both the horizontal and vertical dimension was lower in the correlated condition (horizontal SSE=369.5cm^2, vertical SSE=502.9cm^2) than in the uncorrelated condition (horizontal SSE=834.2cm^2, vertical SSE=759.8cm^2). Again, the overall squared errors between MLE predictions and the empirical data in the horizontal and vertical dimension was not correlated (Pearson's ρ= -0.065, p =0.7982), thus the results from both dimensions provide independent converging evidence that correlated data more closely followed the predictions of MLE than the uncorrelated data.

6.4. Discussion

The results of Experiment 6.1 demonstrate that human observers use the correlation between signals presented in different sensory modalities to integrate audiovisual signals. That is, when faced with temporally correlated auditory and visual signals, the sensory system infers a common underlying cause, and eventually integrates the two sources of information optimally into a coherent multisensory representation of a single

audiovisual event. In other words, when crossmodal signals are correlated, observers seems to expect such stimuli to refer to the same event; conversely, when crossmodal stimuli are uncorrelated, participants are more likely to consider those signals as being independent (that is, as belonging to different physical events).

It should be noted that these results cannot be explained in terms of the auditory and visual stimulus streams being more asynchronous in the uncorrelated condition than in the correlated condition. The correlated and uncorrelated conditions were equated in terms of the average audiovisual delays involved by introducing a ±66ms temporal offset (i.e., asynchrony) between the visual and auditory stimuli in the correlated condition (i.e., a cross-correlation with a maximum correlation=1 at a lag of 66ms either visual-leading or auditory-leading; see Methods). Therefore, the results of Experiment 6.1 demonstrate that it is the correlation between the temporal structures of the unisensory signals, rather than simply their (a)synchrony, that modulates audiovisual integration. It will be a challenge for future research to investigate the effect of different temporal offset between

correlated stimuli, and the possible interactions between correlation and delay.

Interestingly, the present results demonstrate that MSI in the spatial dimension is influenced by the correlation between the signals along the (orthogonal) time dimension. In other words, the temporal correlation between multiple sensory signals promotes spatial MSI by informing the system that two signals have a common physical cause. In statistical terms, correlation provides a measure of dependence between two variables: the higher the correlation, the lower the probability that those two variables are independent, but this by no mean implies causation. Sensory systems, however, have no direct access to the causal structure of the real world, hence causality must be inferred from the available sensory cues. Therefore, knowing that two signals are correlated (i.e., not statistically independent), makes it more likely that the organism will assume a common underlying cause.

In this sense, for the human sensory system, correlation really does imply causation.

7. General discussion

7.1. Summary of results

The series of experiments reported in the present thesis demonstrates the effects of crossmodal correspondences on the rate of information processing, on the perception of simultaneity and eventually their role in multisensory integration. After reviewing the literature on crossmodal correspondences in Chapter 2, in Chapter 3 a set of experiments were described highlighting the effects of crossmodal correspondences on choice RT using a simplified version of the Implicit Association Task (IAT). In line with previous studies, observers were faster and more accurate on compatible as compared to incompatible trials. Additionally, it was found

that audiovisual crossmodal correspondences equally affected responses to visual and to auditory stimuli. Notably, such effects occurred at a very early stage of information processing, hence reflecting an automatic detection of crossmodal correspondences. Moreover, different crossmodal correspondences lead to very similar effect sizes, thus supporting the notion that there might be a single mechanism underlying the effects of a large spectrum of audiovisual correspondences.

In Chaper 4, the perceptual effects of crossmodal correspondences were investigated. Exploiting the temporal ventriloquist effect, it has been demonstrated that crossmodal correspondences systematically distort the perceived timing of transient sensory inputs. This result highlights the effects of crossmodal correspondences on sensory processing and provides the first empirical evidence that crossmodal correspondences operate at a perceptual level.

The experiments reported in Chapter 5 demonstrate that crossmodal correspondences modulate sensitivity to both spatial and temporal intersensory conflicts. These results provide further support for the claim

that crossmodal correspondences operate on a perceptual level and further demonstrate that crossmodal correspondences modulate audiovisual integration.

Finally, the experiment reported in Chapter 6 demonstrates the role of the similarity between sensory inputs in the integration of audiovisual signals. The results, showing that optimal integration occurs only for correlated audiovisual signals, demonstrate that the brain exploits the correlation between sensory signals in order to infer whether or not stimuli perceived through different senses have a single physical cause, and hence whether they should be integrated or not.

Taken together, the results reported in this thesis demonstrate the important role that the correlation between signals (both contingent and learnt) plays in multisensory integration. Two kinds of correlations have been investigated here, one relates to the statistical correlations present in the environment and learnt through experience and interaction with the physical world, and the other relates to the contingent stimulus correlation defining the similarity between the signals (i.e., the cross-correlation). Both

of them similarly affect multisensory processing, by informing the system about whether or not multiple sensory signals have a common physical cause, and hence whether or not they should be integrated.

7.2. A Bayesian framework

Previous studies have demonstrated that spatial and temporal cues, along with semantic information are exploited by the brain in order to solve the correspondence problem (Bresciani, et al., 2005; Chen & Spence, 2010, 2011; Doehrmann & Naumer, 2008; Gepshtein, Burge, Ernst, & Banks, 2005; Körding, et al., 2007; Vatakis & Spence, 2007a). Crossmodal correspondences and temporal correlation might operate as additional cues to help solve the correspondence problem by biasing the brain toward integrating congruent stimuli and segregating incongruent ones. In other words, auditory and visual stimuli are more likely assumed to originate from a single event when their features are congruent and/or correlated. This should be especially valid for what it was earlier termed statistical correspondences (see Chapter 2), where congruency is defined as the

probability that two or more features co-occur in nature: If sensory systems exploit such statistical regularities when integrating multisensory signals, congruent stimuli should be more strongly integrated than incongruent ones (Ernst, 2007).

An influential Bayesian model of multisensory integration (Ernst, 2005; Ernst & Di Luca, 2011) provides an elegant framework in which to interpret the results presented in Chapter 5 and 6. According to this model, humans operate as optimal integrators, combining multiple sensory signals with prior knowledge in order to derive more precise combined estimates (Ernst & Bülthoff, 2004). Statistical correlations between multisensory signals can be modeled as Bayesian priors describing the expected (a priori) joint distribution of those signals (Ernst, 2007).

In order to model the effects of a Bayesian prior describing the statistical co-occurrence of two signals, the MLE model introduced in Chapter 6 should be extended to a two-dimensional audio-visual space (see Figure 7.1). The axes defining this two-dimensional space represent a property (e.g., spatial location or temporal occurrence) of visual and

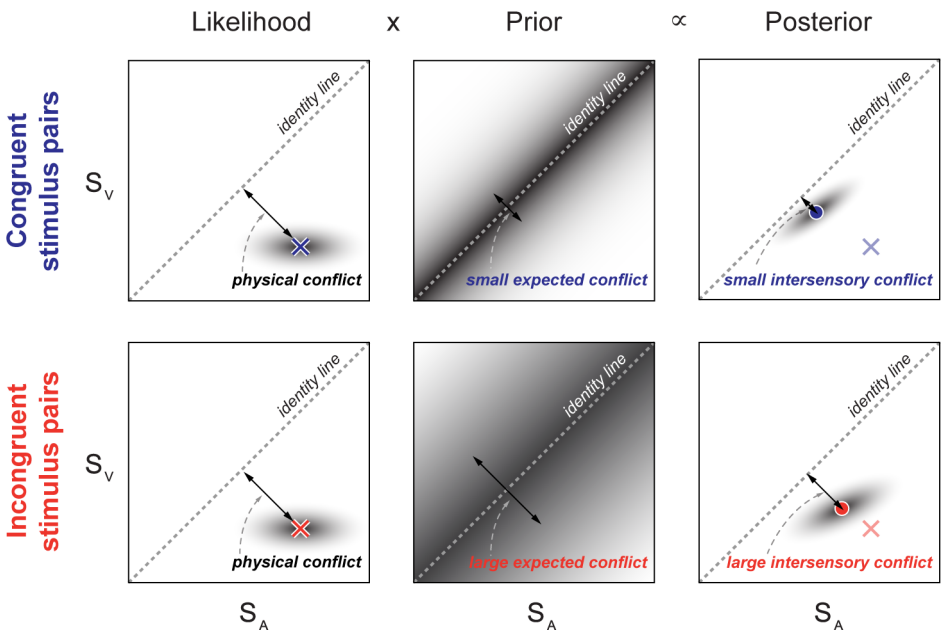


Figure 7.1 Crossmodal congruency promotes multisensory integration by acting on the coupling prior distribution. Multisensory integration results from the combination of sensory inputs (likelihood distribution) and prior knowledge (prior distribution) through Bayes' rule (i.e., likelihood*prior∝posterior). The likelihood distribution (left panels) represents the distribution of physical stimuli inducing a given sensory response. The axes represent a property S (e.g., spatial location) of visual and auditory stimuli. The prior (central panels) represents the expected joint distribution of S from the two sensory channels. A prior that is narrowly distributed along the identity line indicates that observes have a strong expectation that the two cues are identical and conflict free. In contrast, a flat prior indicates complete uncertainty about the mapping between the two cues (i.e., large and small crossmodal conflicts are equally likely). The integrated percept (the posterior, right panels) is the product of the prior and the likelihood distributions. A prior narrowly distributed along the identity line will bias the percept toward the identity line thus silencing crossmodal conflict (see upper panels). Conversely, a shallower prior would leave access to intersensory conflicts (see lower panels). (Note: all the distributions are assumed by the model to be bivariate Gaussian probability density functions). Adapted from (Ernst, 2005, Figure 11).

auditory signals. The prior distribution representing the mapping between the visual and auditory signals, the coupling prior, can be modeled as a bivariate Gaussian distribution with the axis aligned along the diagonals. Assuming that the probability of encountering the two signals is independent of the signals' mean value, the variance of the prior along the positive diagonal (i.e., the identity line) will tend to infinity. The mapping uncertainty between the signals is instead encoded in the variance of the prior along the negative diagonal: The more certain is the mapping, the lower the variance will be.

In the extreme case in which the mapping between sensory signals is completely unknown, the variance of the coupling prior along the negative diagonal will also tend to infinity (i.e., a flat prior). This is equivalent to say that the system expects the two signals to be independent. Conversely, if the mapping between the signals is known perfectly, the variance of the prior along the negative diagonal will be zero. In this case, the sensory signals are mandatorily fused, and the model corresponds to the standard

MLE model described in Chapter 6. In a plausible scenario, even when the mapping between two sensory signals is known, due to the noise arising at any stage of sensory processing (see Green & Swets, 1966), the variance of the prior will always be greater than zero.

According to Bayesian Decision Theory, the sensory inputs (the likelihood function) and prior knowledge (the prior distribution) are combined to produce a posterior distribution, which, in the case of multisensory cue combination, represents the final combined percept. The posterior is proportional to the product of the likelihood and the prior: The lower the variance of the prior, the more the posterior would be biased by the prior. Therefore if the prior happens to be flat (i.e., the mapping between the signals is unknown), the sensory signals will not be integrated, and the final percept is not biased by the prior toward the identity line. That is, any potential discrepancy between the individual sensory estimates should not be cancelled by the prior, and hence should still be accessible to the system. Conversely, if the variance of the prior approaches zero (delta function), the sensory signals are completely fused and biased toward the

prior. Critically, in this case any potential discrepancy between the individual signals should be completely cancelled out.

Therefore, the reduced sensitivity to spatial and temporal conflicts between congruent audiovisual stimuli, as described in Chapter 5, suggests that observers have encoded the natural mapping between the sensory signals (e.g. between pitch and size) and exploit this information (i.e., the coupling prior) in order to integrate audiovisual signals. That is, whenever they hear a high-pitched sound, observers should (implicitly) expect a small object to have produced it, and eventually integrate the two sources of information into a coherent multisensory representation of a single event involving a small object resonating at a high frequency. In other words, when the size of an object and the pitch of a sound are congruent, observers should expect such stimuli to be involved in the same event, hence their spatial location and temporal occurrence should coincide. Conversely, when crossmodal stimuli are incongruent, the sensory systems might consider those signals to be independent, with no need to postulate either the spatial

or temporal coincidence between the unisensory signals. A similar argument would apply in the case of temporal correlation.

In statistical terms, the expected probability of a large conflict (either spatial or temporal) between congruent stimuli should be lower than between incongruent stimuli (see Figure 7.1). Therefore, combining this prior probability (the coupling prior) with discrepant multisensory input (i.e. the likelihood function) according to Bayes' rule, should lead to different results depending on the congruency between the signals: When they are congruent, the conflict is largely cancelled, whereas when they are incongruent the conflict would still be accessible (see also Bresciani, Dammeier, & Ernst, 2006; Ernst, 2007). This prediction is largely supported by the results reported in Chapter 5.

A similar argument can be applied to the results of the experiment described in Chapter 6. The effects of temporal correlation on multisensory integration can also be interpreted in terms of mapping uncertainty. When two signals correlate, it is likely that they come from the same physical event, therefore the mapping between the spatial positions of the visual and

auditory signals should be highly certain (i.e., variance of the coupling prior should be low). Conversely, when the signals do not correlate, it is unlikely that they come from the same physical event, and hence their mapping should be more uncertain. Given that when the mapping between the signals is completely certain the model is equivalent to the standard MLE model, and that the data from the correlated condition were in close agreement with the prediction of MLE, it might be argued that the variance of the coupling prior for the correlated condition approached zero. Conversely, the suboptimal integration of the uncorrelated signals might be interpreted according to this framework in terms of higher mapping uncertainty.

7.3. Crossmodal correspondences and perceptual development

Previous studies of multisensory integration have demonstrated that infants are not optimal integrators (Gori, Del Viva, Sandini, & Burr, 2008; Nardini, Jones, Bedford, & Braddick, 2008; see also Bremner, Lewkowicz, &

Spence, in press). That is, they do not appear to combine multisensory information according to a maximum likelihood estimation strategy. This finding has been interpreted as a lack of knowledge in infants about the mapping between multisensory signals (Ernst, 2008; though see Meltzoff, 1993; Meltzoff & Borton, 1979). Considering that crossmodal correspondences may play a role in solving the correspondence problem (see Chapter 5), it is somewhat surprising to find that a number of studies have also demonstrated the existence of crossmodal correspondences in infants (Maurer, Pathman, & Mondloch, 2006; Mondloch & Maurer, 2004; P Walker, et al., 2010; see also Spector & Maurer, 2009, for a review). It should, however, be noted that those studies did not directly investigate multisensory cue integration, rather their main aim was to highlight the existence of crossmodal correspondences in infants; That is, although the paradigms used in those studies were well-suited to measure compatibility effects between crossmodal stimuli, they did not allow one to measure the effects of such compatibility on multisensory integration. Therefore, it might be argued that infants recognize crossmodal congruency without necessarily using this information in order to solve the correspondence

problem. This apparent inconsistency between the lack of optimal integration in infants and their ability to recognize crossmodal correspondences parallels the inconsistency between those studies that have shown that infants have quite developed multisensory abilities (Lewkowicz, 2000; Lewkowicz, Leo, & Simion, 2010) and those reporting that there is no optimal integration (Gori et al., 2008; Nardini et al., 2008). It will be a challenge for future research to investigate the developmental trajectory of crossmodal correspondences in terms of their role as cues that help a rapidly-developing infant to solve the correspondence problem. In this regard, if it is true that crossmodal correspondences are encoded in the coupling prior distribution, and that what infants are missing is the knowledge about the mapping between multisensory signals (which is also represented in the coupling prior), it can be hypothesized that the effects of crossmodal correspondence in multisensory integration should not appear before optimal integration.

7.4. Crossmodal correspondences and synesthesia

A final open question concerns the relation between crossmodal correspondences and synesthesia (see Table 6.1). As mentioned already, crossmodal correspondences have often been linked to full-blown synesthesia (see Martino & Marks, 2001; Maurer, 1997), and a number of authors have actually chosen to refer to what are here called 'crossmodal correspondences' as 'synesthetic correspondences' (e.g., Martino & Marks, 2000, 2001; Parise & Spence, 2008, 2009; Walker, et al., 2010). Although, at first sight, it might be tempting to connect the two phenomena, as both involve multisensory perception and congruency effects, there are a number of major differences between them that deserve further consideration.

Crossmodal Correspondences	**Synesthesia**
Universal	Rare
Shared correspondences	Idiosyncratic correspondences
No concurrent experiences	Concurrent sensory experiences
Relative compatibility effects	Absolute mapping between inducer and concurrent
Can be learnt following training	Cannot be acquired by training

Table 7.1 Differences beween crossmodal correspondences and synesthesia

First of all, to the best of my knowledge, the literature on crossmodal correspondences has never reported the presence of a concurrent experience of a congruent stimulus in an unstimulated modality as a result of the presentation of an inducing stimulus in another sensory modality. Nevertheless it should be noted that recent studies on synesthesthetes have highlighted the existence of a phenomenon that bears close resemblance to crossmodal correspondences. In particular, it has been demonstrated that even in those cases in which synesthesia is apparently unidirectional (i.e., when stimulation in one modality elicits a concurrent sensation in another modality but not the other way round), synesthesia sometime also occurs in the opposite direction (albeit below the level of awareness; see Cohen Kadosh, Cohen Kadosh, & Henik, 2007; Cohen Kadosh & Henik, 2007; Cohen Kadosh, Henik, & Walsh, 2007; Johnson, Jepma, & De Jong, 2007; Meier & Rothen, 2007). That said, it should be remembered that full-blown synesthetic experiences are mostly idiosyncratic (e.g., see Grossenbacher & Lovelace, 2001; Rouw & Scholte, 2010; though see Ward, Huckstep, & Tsakanikos, 2006), whereas crossmodal correspondences are regular (and sometimes universal).

Second, crossmodal correspondences between 'prothetic' (or polar) dimensions appear to be largely relative phenomena, occurring only after pairs of stimuli have been experienced as one being 'more' than the other, whereas proper synesthetic experiences involve an absolute mapping between the inducer and the concurrent stimulus. In this regard, it is remarkable that the effects of crossmodal correspondences between complementary prothetic cues only occur when two stimuli can be clearly identified as one being 'more' and the other being 'less' along a given sensory dimension, but not when congruent and incongruent stimulus pairs are presented in different blocks of trials (Bernstein & Edelstein, 1971; Gallace & Spence, 2006).

Third, while novel crossmodal correspondences can be experimentally induced after a relatively short (<1 hour) exposure to correlated crossmodal stimuli (see Ernst, 2007), synesthetic concurrents cannot be elicited in non-synesthetes even after many tens of thousands of presentations of arbitrary crossmodal stimulus pairings (e.g., Howells, 1944; Kelly, 1934). That is, while it seems clear that perceptual learning plays a key role in the

acquisition of crossmodal correspondences, it can hardly account for the development of full-blown synesthesia (though one might postulate the existence of critical periods after which synesthesia cannot be induced anymore).

With this in mind, it can be argued that in spite of their superficial similarities, synesthesia and crossmodal correspondences might well constitute two different, and possibly entirely unrelated, perceptual phenomena (Spence, 2011). That said, the question is not yet fully resolved. More studies will clearly be needed in order to investigate the precise nature of the relationship between these two perceptual phenomena. In this regard, it would be interesting to directly compare the effects of crossmodal correspondences on normal and synesthetic populations on both the speed of information processing and multisensory integration (see Spence, 2011). Moreover, the application of functional neuroimaging techniques to investigate the neural substrates underlying synesthesia and crossmodal correspondences might also provide valuable insights to help decide whether the two phenomena constitute instances of a single continuum or not.

7.5. Concluding remarks

In the previous chapters, the effects of crossmodal correspondences on sensory processes have been widely documented through novel experiments and a review of the current literature. In Chapter 2 it has been claimed that at least some crossmodal correspondences, namely the statistical correspondences, might reflect the learning of the natural co-occurrence of certain stimulus properties in the real world. Future research should ground this claim with direct measurements of the actual statistics of the environment. Indeed, while sometimes the existence of natural correlations between stimulus properties it is quite obvious, for example because they reflect the laws of physics (e.g., the relation between the size of an object and both its resonance frequency and mass), in other cases such natural mapping between stimulus properties deserves a closer inspection. This is the case for the association between pitch and elevation, one of the most robust and investigated examples of crossmodal correspondences (see Chapter 2). Where exactly does this correspondence come from? It might be argued that small objects, that as we know resonate at higher frequencies,

are more likely to be found at high elevations (e.g., birds have often shrill voices), but this claim is admittedly rather speculative. A simple solution to directly prove the existence of a natural mapping between pitch and elevation would be to collect a large number of environmental sounds using two directional microphones, one pointing up and the other down. If it is true that in the world there is a natural relationship between pitch and elevation, this should emerge from a comparison of the spectra of the environmental sounds recorded by the two microphones. More generally, investigating the statistics of the real world would provide useful insights with regard to the origin of crossmodal correspondences and eventually would allow for a more reliable and detailed taxonomy of crossmodal correspondences.

A second open question relates to the level of details of the internalization of the statistics of the environment. Do humans, for example, have rule of thumb knowledge of the relation between pitch and size or do they know the exact relation between these two physical properties? That is, do they just know that the resonance frequency roughly

scales with size, or do they know the exact shape of such relation? These questions closely relate to the relative vs. absolute nature of crossmodal correspondences discussed in Chapter 2. Surely, given the size of an object it is not possible to determine unequivocally its resonance frequency, because it depends also on the density and the tension of the resonator. But what if such information is provided? With this respect it would possibly be sufficient to give observers a standard, an audiovisual stimulus displaying a resonating object, and ask them to adjust the resonance frequency of a sound given both the standard and a probe visual stimulus of variable size (or vice versa to adjust the size of a visual stimulus given the standard and a probe auditory stimulus of variable frequency). Such a technique would allow for the estimation of the shape of the internal function that defines the mapping between multiple sensory signals and a comparison of such internalized function with the physical one. This comparison would provide further insights on the important question on whether humans simply possess a rough knowledge of the natural statistics of the environment, or else on whether instead they faithfully extract the actual laws of physics.

A third open question relates to the crossmodal nature of sensory correspondences: Are the compatibility effects described in the present thesis fundamentally multisensory? Literature on human perception provides numerous examples of compatibility between different sensory within-modality features. This is the case for objects' size and weight, two haptic (though size can also be estimated through vision) properties that are normally correlated in nature. Interestingly, it has been widely demonstrated that both manipulatory actions and weight perception are influenced by this natural correlation (e.g., see Cole, 2008; Ellis, & Lederman, 1993; Flanagan, & Beltzner 2000; Flanagan, Bittner, & Johansson, 2008; Johansson, & Westling, 1988). Other examples of within modality correspondences in touch are shape and size (Kahrimanovic, Bergmann Tiest, & Kappers, 2010), temperature and weight (J.C. Stevens, & Hopper, 1982), and weight and shape (Kahrimanovic, Bergmann Tiest, & Kappers, 2011), to name but a few. Nevertheless, to the best of my knowledge there are no studies directly comparing the effects of unimodal and crossmodal sensory correspondences. It will be a matter of future research to assess whether unimodal and crossmodal sensory

correspondences are two instances of the same phenomenon or not, and whether they play a similar role on feature integration.

Finally, the very existence of crossmodal correspondences begs the question of their biological relevance: What is the adaptive value of knowing the mapping between multiple sensory signals? This is a critical point, because on top of allowing for a better perception (and prediction) of the world under most circumstances, crossmodal correspondences can also be exploited in the animal world for 'dishonest signaling'. Some animals, for example, can modulate their voices in order to fake on their body size (Fitch & Hauser, 2003). This is the case for the male red deer (*Cervus elaphus*), which can lengthen their vocal tracts in order to lower the frequencies of their bells and 'appear' larger. In the red deer, this nifty strategy uncovers neat advantages in both mating and agonistic behavior (Charlton, Reby, & McComb, 2007; Reby et al., 2005). Likely, this signaling behavior triggers an evolutionary race between signalers and receives to better produce, and spot-out, dishonest signaling (Krebs & Davies, 1978). Therefore, in the long run, in species like the red deer

exploiting dishonest signals, the natural correlation between pitch and size would likely break down (though see Zahavi, 1975). So how do they balance the costs and benefits of dishonest signaling and still maintain a crossmodal correspondence between sensory properties that sometimes come to be uncorrelated? In order to answer to all these questions, and to better understand the evolutionary advantage of crossmodal correspondences, it would be especially interesting to study the effects of crossmodal correspondences in dishonest signaling species like the red deer.

8. References

Adams, W. J., Graf, E. W., & Ernst, M. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience, 7*(10), 1057-1058.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*(3), 257-262.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioual Brain Science, 33*(4), 245-266; discussion 266-313.

Antovic, M. (2009). Musical metaphors in Serbian and Romani children: an empirical study. *Metaphor and Symbol, 24*(3), 184-202.

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

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science, 13*(3), 99-102.

Baier, B., Kleinschmidt, A., & Müller, N. G. (2006). Cross-modal processing in early visual and auditory cortices depends on expected statistical relationship of multisensory information. *Journal of Neuroscience, 26*(47), 12260-12265.

Bermant, R. I., & Welch, R. B. (1976). Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition. *Perceptual and Motor Skills, 42*, 487-493.

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology, 87*(2), 241-247.

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

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

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

Bozzi, P., & Flores D'Arcais, G. (1967). Experimental research on the intermodal relationships between expressive qualities. *Archivio di Psicologia, Neurologia e Psichiatria, 28*(5), 377-420.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433-436.

Bremner, A., Lewkowicz, D. J., & Spence, C. (Eds.). (in press). *Multisensory development*. Oxford: Oxford University Press.

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

Bresciani, J. P., Ernst, M. O., Drewing, K., Bouyer, G., Maury, V., & Kheddar, A. (2005). Feeling what you hear: Auditory signals can modulate tactile tap perception. *Experimental Brain Research, 162*(2), 172-180.

Burr, D., Silva, O., Cicchini, G. M., Banks, M. S., & Morrone, M. C. (2009). Temporal mechanisms of multimodal binding. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 276*, 1761-1769.

Bushara, K. O., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *Journal of Neuroscience, 21*(1), 300-304.

Cabrera, D., & Morimoto, M. (2007). Influence of fundamental frequency and source elevation on the vertical localization of complex tones and complex tone pairs. *Journal of the Acoustical Society of America, 122*, 478-488.

Caclin, A., Soto-Faraco, S., Kingstone, A., & Spence, C. (2002). Tactile 'capture' of audition. *Perception, & Psychophysics, 64*(4), 616-630.

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

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*(12), 535-543.

Charlton, B. D., Reby, D., & McComb, K. (2007). Female red deer prefer the roars of larger males. *Biology Letters, 3*(4), 382-385.

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*(3), 389-404.

Chen, Y. C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity.

Journal of Experimental Psychology: Human Perception and Performance, Advance online publication. doi: 10.1037/a0024329.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*: Academic Press New York.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: L Erlbaum Associates.

Cohen Kadosh, R., Cohen Kadosh, K., & Henik, A. (2007). The neuronal correlate of bidirectional synesthesia: A combined event-related potential and functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience, 19*(12), 2050-2059.

Cohen Kadosh, R., & Henik, A. (2007). Can synaesthesia research inform cognitive science? *Trends in Cognitive Sciences, 11*(4), 177-184.

Cohen Kadosh, R., Henik, A., & Walsh, V. (2007). Small is bright and big is dark in synaesthesia. *Current Biology, 17*(19), R834-R835.

Cohen, N. (1934). Equivalence of brightness across modalities. *American Journal of Psychology, 13*(3), 117-119.

Cole, K. J. (2008). Lifting a familiar object: visual size analysis, not memory for object weight, scales lift force. *Experimental Brain Research, 188*(4), 551-557.

Cowey, A., & Weiskrantz, L. (1975). Demonstration of cross-modal matching in rhesus monkeys, Macaca mulatta. *Neuropsychologia, 13*(1), 117-120.

Cowles, J. T. (1935). An experimental study of the pairing of certain auditory and visual stimuli. *Journal of Experimental Psychology, 18*(4), 461-469.

Crisinel, A. S., & Spence, C. (2009). Implicit association between basic tastes and pitch. *Neuroscience Letters, 464*(1), 39-42.

Crisinel, A. S., & Spence, C. (2010). A sweet sound? Food names reveal implicit associations between taste and pitch. *Perception, 39*(3), 417-425.

Davis, R. (1961). The fitness of names to drawings: A cross-cultural study in Tanganyika. *British Journal of Psychology, 52*(3), 259-268.

De Jong, R., Liang, C. C., & Lauber, E. (1994). Conditional and unconditional automaticity: A dual-process model of effects of spatial stimulus-response correspondence. *Journal of Experimental Psychology: Human Perception and Performance, 20*(4), 731.

de Lange Dzn, H. (1954). Relationship between critical flicker-frequency and a set of low-frequency characteristics of the eye. *Journal of the Optical Society of America, 44*(5), 380-388.

Demattè, M. L., Sanabria, D., & Spence, C. (2006). Cross-modal associations between odors and colors. *Chemical Senses, 31*(6), 531-538.

Demattè, M. L., Sanabria, D., & Spence, C. (2007). Olfactory-tactile compatibility effects demonstrated using a variation of the Implicit Association Test. *Acta Psychologica, 124*(3), 332-343.

Di Luca, M., Ernst, M. O., & Backus, B. (2010). Learning to use an invisible visual signal for perception. *Current Biology, 20*(20), 1860-1863.

Diffloth, G. (1994). i:big, a:small. In L. Hinton & J. Nichols (Eds.), *Sound symbolism* (pp. 107-114). Cambridge: Cambridge University Press.

Dixon, W., & Mood, A. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the American Statistical Association, 43*(241), 109-126.

Doehrmann, O., & Naumer, M. J. (2008). Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain research, 1242*(25), 136-150.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Paper presented at the Society for Industrial and Applied Mathematics, Philadelphia.

Eitan, Z., & Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition, 114*(3), 405-422.

Ellis, R. R., & Lederman, S. J. (1993). The role of haptic versus visual volume cues in the size-weight illusion. *Attention, Perception, & Psychophysics, 53*(3), 315-324.

Ernst, M. O. (2005). A Bayesian view on multimodal cue integration. In G. Knoblich, I. Thornton, M. Grosejan & M. Shiffrar (Eds.), *Perception of the human body from the inside out* (pp. 105–131). New York Oxford University Press.

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

Ernst, M. O. (2008). Multisensory integration: A late bloomer. *Current Biology, 18*(12), R519-R521.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429-433.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*(4), 162-169.

Ernst, M. O., & Di Luca, M. (2011). Multisensory perception: From integration to remapping. In J. Trommershäuser, M. Landy & K.

Körding (Eds.), *Sensory cue integration.* (pp. 224-250). New York, NY: Oxford University Press.

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

Faragó, T., Pongrácz, P., Miklósi, Á., Huber, L., Virányi, Z., & Range, F. (2010). Dogs' expectation about signalers' body size by virtue of their growls. *PLoS ONE, 5*(12), e15175.

Fendrich, R., & Corballis, P. M. (2001). The temporal cross-capture of audition and vision. *Perception & Psychophysics, 63*(4), 719-725.

Fitch, W., & Hauser, M. (2003). Unpacking "honesty":Vertebrate vocal production and the evolution of acoustic signals. In A. Simmons, A. Popper & R. Fay (Eds.), *Acoustic communication* (pp. 65–137). New York: Springer.

Flanagan, J. R., & Beltzner, M. A. (2000). Independence of perceptual and sensorimotor predictions in the size-weight illusion. *Nature Neuroscience, 3*, 737-741.

Flanagan, J. R., Bittner, J. P., & Johansson, R. S. (2008). Experience can change distinct size-weight priors engaged in lifting objects and judging their weights. *Current Biology, 18*(22), 1742-1747.

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7*(7), 773-778.

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics, 68*(7), 1191-1203.

Garner, W. R. (1974). *The processing of information and structure.* Potomac, MD: Lawrence Elrbaum Associates.

Gebels, G. (1969). An investigation of phonetic symbolism in different cultures. *Journal of Verbal Learning and Verbal Behavior, 8*(2), 310-312.

Gepshtein, S., Burge, J., Ernst, M. O., & Banks, M. S. (2005). The combination of vision and touch depends on spatial proximity. *Journal of Vision, 5*(11), 1013-1023.

Gori, M., Del Viva, M., Sandini, G., & Burr, D. C. (2008). Young children do not integrate visual and haptic form information. *Current Biology, 18*(9), 694-698.

Green, A. M., & Angelaki, D. E. (2010). Multisensory integration: Resolving sensory ambiguities to build novel representations. *Current Opinion in Neurobiology, 20*(3), 353-360.

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

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*(6), 1464-1480.

Grossenbacher, P. G., & Lovelace, C. T. (2001). Mechanisms of synesthesia: Cognitive and physiological constraints. *Trends in Cognitive Sciences, 5*(1), 36-41.

Grusser, O. (1983). Multimodal structure of the extrapersonal space. In A. Hein & M. Jeannerod (Eds.), *Spatially oriented behavior* (pp. 327-352). New York, NY: Springer-Verlag.

Haijiang, Q., Saunders, J. A., Stone, R. W., & Backus, B. T. (2006). Demonstration of cue recruitment: Change in visual appearance by means of Pavlovian conditioning. *Proceedings of the National Academy of Sciences of the United States of America, 103*(2), 483-488.

Hegelmaier, F. (1852). Über das Gedächtnis für Linear-Anschauungen (On the memory for the length of a line). *Archiv für physiologische Heilkunde, 11*, 844-853.

Hevner, K. (1935). Experimental studies of the affective value of colors and lines. *Journal of Applied Psychology, 19*(4), 385-398.

Hillis, J., Ernst, M. O., Banks, M., & Landy, M. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science, 298*(5598), 1627-1630.

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

Hornbostel, E. (1938). The unity of the senses. In M. V. Ellis & D. Willis (Eds.), *A source book of Gestalt psychology* (pp. 210-216). New York: The Gestalt Journal Press.

Howells, T. (1944). The experimental development of color-tone synesthesia. *Journal of Experimental Psychology, 34*(2), 87-103.

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

Jaekl, P. M., & Harris, L. R. (2007). Auditory-visual temporal integration measured by shifts in perceived temporal location. *Neuroscience Letters, 417*(3), 219-224.

Johansson, R., & Westling, G. (1988). Coordinated isometric muscle commands adequately and erroneously programmed for the weight during lifting task with precision grip. *Experimental Brain Research, 71*(1), 59-71.

Johnson, A., Jepma, M., & De Jong, R. (2007). Colours sometimes count: Awareness and bidirectionality in grapheme-colour synaesthesia. *The Quarterly Journal of Experimental Psychology, 60*(10), 1406-1422.

Kahrimanovic, M., Bergmann Tiest, W., & Kappers, A. M. L. (2010). Haptic perception of volume and surface area of 3-D objects. *Attention, Perception, & Psychophysics, 72*(2), 517-527.

Kahrimanovic, M., Bergmann Tiest, W., & Kappers, A. (2011). Characterization of the haptic shape-weight illusion with 3-dimensional objects. *IEEE Transactions on Haptics, 4*(4), 316-320.

Kanai, R., Sheth, B. R., Verstraten, F. A. J., & Shimojo, S. (2007). Dynamic perceptual changes in audiovisual simultaneity. *PLoS ONE, 2*(12), e1253.

Karwoski, T., Odbert, H., & Osgood, C. E. (1942). Studies in synesthetic thinking: II. The role of form in visual responses to music. *Journal of General Psychology, 26*, 212-221.

Keetels, M., Stekelenburg, J., & Vroomen, J. (2007). Auditory grouping occurs prior to intersensory pairing: Evidence from temporal ventriloquism. *Experimental Brain Research, 180*(3), 449-456.

Keetels, M., & Vroomen, J. (2011). No effect of synesthetic congruency on temporal ventriloquism. *Attention, Perception, & Psychophysics, 73*(1), 1-10.

Kelly, E. L. (1934). An experimental attempt to produce artificial chromaesthesia by the technique of the conditioned response. *Journal of Experimental Psychology, 17*(3), 315-341.

King, A., & Palmer, A. (1985). Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Experimental Brain Research, 60*(3), 492-500.

King, R., & Oldfield, S. (1997). The impact of signal bandwidth on auditory localization: Implications for the design of three-dimensional audio displays. *Human Factors, 39*(2), 287–295.

Klein, R., Brennan, M., & Gilani, A. (1987 November). *Covert cross-modality orienting of attention in space.* Paper presented at the Annual meeting of the Psychonomic Society, Seattle, WA.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge, UK: Cambridge University Press.

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

Köhler, W. (1947). *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology*. New York, NY: Liveright Publ. Corporation.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE, 2*(9), 943.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 945-959.

Krebs, J. R., & Davies, N.B. (1978). *Behavioural ecology: an evolutionary approach.* Oxford: Blackwell.

Külpe, O. (1893). *Grundriss der Psychologie: Auf experimenteller Grundlage (Fundamentals of psychology: From an experimental perspective).* Leipzig: W. Engelmann.

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*(4), 405-414.

Levitan, C. A., Zampini, M., Li, R., & Spence, C. (2008). Assessing the role of color cues and people's beliefs about color-flavor associations on the discrimination of the flavor of sugar-coated chocolates. *Chemical Senses, 33*(5), 415-423.

Lewald, J. (1997). Eye-position effects in directional hearing. *Behavioural Brain Research, 87*(1), 35-48.

Lewald, J., & Ehrenstein, W. H. (1998). Auditory-visual spatial integration: A new psychophysical approach using laser pointing to acoustic targets. *Journal of the Acoustical Society of America, 104*(3), 1586-1597.

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological bulletin, 126*(2), 281-308.

Lewkowicz, D. J., Leo, I., & Simion, F. (2010). Intersensory perception at birth: Newborns match non-human primate faces & voices. *Infancy, 15*(1), 46-60.

Long, J. (1977). Contextual assimilation and its effect on the division of attention between nonverbal signals. *The Quarterly Journal of Experimental Psychology, 29*(3), 397-414.

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *American Journal of Psychology, 87*(1/2), 173-188.

Marks, L. E. (1987a). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psycholology: Human Perception and Performance, 13*(3), 384-394.

Marks, L. E. (1987b). On cross-modal similarity: Perceiving temporal patterns by hearing, touch, and vision. *Perception & Psychophysics, 42*(3), 250-256.

Marks, L. E. (1989). On cross-modal similarity: The perceptual structure of pitch, loudness, and brightness. *Journal of Experimental Psycholology: Human Perception and Performance, 15*(3), 586-602.

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of mutisensory processes.* (pp. 85-106). Cambridge, MA: MIT Press.

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception, 28*(7), 903-924.

Martino, G., & Marks, L. E. (2000). Cross-modal interaction between vision and touch: The role of synesthetic correspondence. *Perception, 29*(6), 745-754.

Martino, G., & Marks, L. E. (2001). Synesthesia: Strong and weak. *Current Directions in Psychological Science, 10*(2), 61-65.

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

Maurer, D., Pathman, T., & Mondloch, C. (2006). The shape of boubas: sound-shape correspondences in toddlers and adults. *Developmental Science, 9*(3), 316-322.

Meier, B., & Rothen, N. (2007). When conditioned responses "fire back": Bidirectional cross-activation creates learning opportunities in synesthesia. *Neuroscience, 147*(3), 569-572.

Melara, R. D. (1989). Similarity relations among synesthetic stimuli and their attributes. *Journal of Experimental Psychology: Human Perception and Performance, 15*(2), 212-231.

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General, 116*(4), 323-336.

Melara, R. D., & O'Brien, T. P. (1990). Effects of cuing on cross-modal congruity. *Journal of Memory and Language, 29*(6), 655-686.

Meltzoff, A. N. (1993). Molyneux's babies: Cross-modal perception, imitation and the mind of the preverbal infant. In N. Eilan, R. McCarthy & B. Brewer (Eds.), *Spatial representation: Problems in philosophy and psychology* (pp. 219-235). Oxford: Blackwell.

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

Meyer, G. F., & Noppeney, U. (2011). Multisensory integration: from fundamental principles to translational research. *Experimental Brain Research, 213*(2-3), 163-166.

Meyer, G. F., Wuerger, S. M., Röhrbein, F., & Zetzsche, C. (2005). Low-level integration of auditory and visual motion signals requires spatial co-localisation. *Experimental Brain Research, 166*(3), 538-547.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14*(2), 247-279.

Miller, J. (1991). Channel interaction and the redundant-targets effect in bimodal divided attention. *Journal of Experimental Psychology: Human Perception and Performance, 17*(1), 160-169.

Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch–object correspondences in young children. *Cognitive, Affective, & Behavioral Neuroscience, 4*(2), 133-136.

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: Examining temporal ventriloquism. *Cognitive Brain Research, 17*(1), 154-163.

Murray, M. M., & Wallace, M. T. (Eds.). (2011). *The neural bases of multisensory processes*. Boca Raton, FL: CRC Press.

Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Current Biology, 18*(9), 689-693.

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

O'Boyle, M. W., & Tarte, R. D. (1980). Implications for phonetic symbolism: the relationship between pure tones and geometric figures. *Journal Psycholinguistic Research, 9*(6), 535-544.

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*(3-4), 348-355.

Osgood, C. E. (1960). The cross-cultural generality of visual–verbal synesthetic tendencies. *Behavioral Science, 5*(2), 146-169.

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

Otto, T., & Mamassian, P. (2010). Noise vs sensory integration: The return of the race model. *Perception, 39*(ECVP Abstract Supplement), 67.

Parise, C., & Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research. 214*(3), 373-380

Parise, C., & Spence, C. (2008). Synesthetic congruency modulates the temporal ventriloquism effect. *Neuroscience Letters, 442*(3), 257-261.

Parise, C., & Spence, C. (2009). When birds of a feather flock together: Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE, 4*(5), e5664.

Parise, C., & Spence, C. (under review). Audiovisual crossmodal correspondences. In J. Simner & E. M. Hubbard (Eds.), *Oxford handbook of synaesthesia.* Oxford, UK: Oxford University Press.

Parker, A., & Easton, A. (2004). Cross-modal memory in primates: The neural basis of learning about the multisensory properties of objects and events. In G. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisensory processes* (pp. 333-342). Cambridge, MA: MIT Press.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision, 10*(4), 437-442.

Petrini, K., Dahl, S., Rocchesso, D., Waadeland, C. H., Avanzini, F., Puce, A., & Pollick, F. E. (2009). Multisensory integration of drumming actions: musical expertise affects perceived audiovisual asynchrony. *Experimental Brain Research, 198*(2), 339-352.

Poffenberger, A., & Barrows, B. (1924). The feeling value of lines. *Journal of Applied Psychology, 8*(2), 187-205.

Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences, 24*(5), 574-590.

Radeau, M. (1994). Auditory-visual spatial interaction and modularity. *Current Psychology of Cognition, 13*, 3-51.

Radeau, M., & Bertelson, P. (1987). Auditory-visual interaction and the timing of inputs. *Psychological Research, 49*(1), 17-22.

Ramachandran, V., & Hubbard, E. (2001). Synaesthesia: A window into perception, thought and language. *Journal of Consciousness Studies, 8*(12), 3-34.

Ramachandran, V. S., & Oberman, L. M. (2007). Broken mirrors: a theory of autism. *Scientific American Special Edition, 17*(2), 20-29.

Reby, D., McComb, K., Cargnelutti, B., Darwin, C., Fitch, W., & Clutton-Brock, T. (2005). Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proceedings of the Royal Society B: Biological Sciences, 272*(1566), 941-947.

Robson, D. (2011). Language's missing link. *New Scientist, 211*(2821), 30-33.

Robson, J. (1966). Spatial and temporal contrast-sensitivity functions of the visual system. *Journal of the Optical Society of America, 56*(8), 1141-1142.

Rogers, S. K., & Ross, A. S. (1968). A cross-cultural test of the Maluma-Takete phenomenon. *Perception, 4*(1), 105-106.

Root, R., & Ross, S. (1965). Further validation of subjective scales for loudness and brightness by means of cross-modality matching. *American Journal of Psychology, 78*(2), 285-289.

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: A bivariate boxplot. *The American Statistician, 53*(4), 382-387.

Rouw, R., & Scholte, H. S. (2007). Increased structural connectivity in grapheme-color synesthesia. *nature neuroscience, 10*(6), 792-797.

Rouw, R., & Scholte, H. S. (2010). Neural basis of individual differences in synesthetic experiences. *Journal of Neuroscience, 30*(18), 6205.

Rusconi, E., Kwan, B., Giordano, B. L., Umilta, C., & Butterworth, B. (2006). Spatial representation of pitch height: The SMARC effect. *Cognition, 99*(2), 113-129.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12*(3), 225-239.

Scheier, C., Nijhawan, R., & Shimojo, S. (1999). Sound alters visual temporal resolution. *Investigative Ophthalmology & Visual Science, 40*(Suppl 4), 4169.

Seo, H. S., Arshamian, A., Schemmer, K., Scheer, I., Sander, T., Ritter, G., & Hummel, T. (2010). Cross-modal integration between odors and abstract symbols. *Neuroscience Letters, 478*(3), 175-178.

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

Shore, D., Spence, C., & Klein, R. (2001). Visual prior entry. *Psychological Science, 12*(3), 205-212.

Slutsky, D., & Recanzone, G. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport, 12*(1), 7-10.

Smith, E. L., Grabowecky, M., & Suzuki, S. (2007). Auditory-visual crossmodal integration in perception of face gender. *Current Biology, 17*(19), 1680-1685.

Sokal, R., & Rohlf, F. (1995). *Biometry: The principles and practice of statistics in biological research* (3rd ed.). New York, NY: W. H. Freeman and Co.

Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C., & Kingstone, A. (2002). The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities. *Cognitive Brain Research, 14*(1), 139-146.

Spector, F., & Maurer, D. (2009). Synesthesia: A new approach to understanding the development of perception. *Perception, 45*, 175-189.

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology, 28*(2), 61-70.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*(4), 1-25.

Spence, C., Levitan, C. A., Shankar, M. U., & Zampini, M. (2010). Does food color influence taste and flavour perception in humans? *Chemosensory Perception, 3*(1), 68-84.

Spence, C., & Squire, S. (2003). Multisensory integration: maintaining the perception of synchrony. *Current Biology, 13*(13), R519-R521.

Stevens, J. C., & Hooper, J. E. (1982). How skin and object temperature influence touch sensation. *Attention, Perception, & Psychophysics, 32*(3), 282-285.

Stevens, J. C., & Marks, L. E. (1965). Cross-modality matching of brightness and loudness. *Proceedings of the National Academy of Sciences USA, 54*(2), 407-411.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*(3), 153.

Stevens, S. S., & Stevens, G. (1975). *Psychophysics*. New York: NY: Wiley-Interscience.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9*(4), 578-585.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643-662.

Stumpf, K. (1883). *Tonpsychologie*. Leipzig: S. Hirzel.

Taylor, A., Reby, D., & McComb, K. (2010). Size communication in domestic dog, Canis familiaris, growls. *Animal Behaviour, 79*(1), 205-210.

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

Trommershäuser, J., Landy, M., & Körding, K. (Eds.). (2011). *Sensory cue integration.* New York, NY: Oxford University Press.

Tyler, C. W., & Julesz, B. (1978). Binocular cross-correlation in time and space. *Vision Research, 18*(1), 101-105.

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision, 8*(9), 1-11.

Vatakis, A., & Spence, C. (2007a). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics, 69*(5), 744-756.

Vatakis, A., & Spence, C. (2007b). Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica, 127*, 12-23.

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*(4), 1063.

Vroomen, J., & Stekelenburg, J. (2010). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition, 118*(2), 75-83.

Walker, P., Bremner, J., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science, 21*(1), 21-25.

Walker, P., & Smith, S. (1984). Stroop interference based on the synaesthetic qualities of auditory pitch. *Perception, 13*(1), 75-81.

Walker, P., & Smith, S. (1985). Stroop interference based on the multimodal correlates of haptic size and auditory pitch. *Perception, 14*(6), 729-736.

Walker, R. (1987). The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. *Perception, & Psychophysics, 42*(5), 491-502.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences, 7*(11), 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*(2), 264-280.

Watson, A., & Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics, 47*(1), 87-91.

Watson, A., & Pelli, D. (1983). QUEST- A Bayesian adaptive psychometric method. *Perception and Psychophysics, 33*(2), 113-120.

Weiskrantz, L., & Cowey, A. (1975). Cross-modal matching in the rhesus monkey using a single pair of stimuli. *Neuropsychologia, 13*(3), 257-261.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience, 5*(6), 598-604.

Welch, R., & Warren, D. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*(3), 638-667.

Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language, 93*(1), 10-19.

Wichmann, F., & Hill, N. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63*(8), 1293-1313.

Wuerger, S., Meyer, G., Hofbauer, M., Zetzsche, C., & Schill, K. (2010). Motion extrapolation of auditory-visual targets. *Information Fusion, 11*(1), 45-50.

Zahavi, A. (1975). Mate selection--a selection for a handicap. *Journal of theoretical Biology, 53*(1), 205-214.

Zangenehpour, S., & Zatorre, R. J. (2010). Crossmodal recruitment of primary visual cortex following brief exposure to bimodal audiovisual stimuli. *Neuropsychologia, 48*(2), 591-600.