THE NEURAL BASIS OF AUDIOVISUAL INTEGRATION

by

ALEXANDRA KRUGLIAK

A thesis submitted to the University of Birmingham
for the degree of DOCTOR OF PHILOSOPHY

School of Psychology
College of Life and Environmental Sciences
University of Birmingham
March 2019

ABSTRACT

Our perception is continuous and unified. Yet, sensory information reaches our brains through different senses and needs to be processed in order to create that unified percept. Interactions between sensory modalities occur already at primary cortical levels. The purpose of such interactions and what kind of information they transmit is still largely unknown. The current thesis aimed to reveal the interactions between auditory pitch and visual size in polar coordinates, two modality specific stimulus features that have robust topographic representations in the human brain. In *Chapter 1*, I present the background of cross-modal interactions in early sensory cortices and of the pitch-size relationship. In *Chapter 2*, we explored the pitch-size relationship in a speeded classification task and, in *Chapter 3*, at the level of functional Magnetic Resonance Imaging activation patterns. In *Chapter 4*, we investigated the effects of actively learning a specific pitch-size mapping during one session on the speeded classification task. In *Chapter 5*, we extended learning over multiple sessions and examined learning effects with behavioral and neural measures. Finally, in *Chapter 6*, I summarize the findings of the thesis, its contributions to the literature, and outline directions for future research.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Uta Noppeney for the remarkable support and guidance throughout my PhD. I greatly admire her scientific rigour, the depth of her knowledge, and that she was always available whenever it was necessary and patiently discussed any detail, however small or self-evident it seemed. I have learned a lot from her and she has greatly shaped my professional and personal development.

I would also like to thank the members of the Computational Cognitive Neuroimaging Group (in no particular order): Remi Gau, Ágoston Mihalik, Mate Aller, Arianna Zuanazzi, David Meijer, Sam Jones, Steffen Burgers, Johanna Zumer, Tom White, Ambra Ferrari and Joana Leitao, for their support and the great atmosphere in and outside of the lab.

Furthermore, I would like to thank Tristan Bekinschtein and Anat Arzi, and the other members of the Consciousness and Cognition Laboratory at the University of Cambridge, for providing me with the great opportunity to work as Research Assistant and to explore new scientific opportunities while juggling part-time work with finishing the last analyses and writing the thesis.

The time of my PhD would not have been such an unforgettable experience without the many incredible people who I met, and my wonderful friends who supported me in many ways. I would like to say a special thank you to those who were there for me throughout all those years and were happy enough to put up with me and this PhD (in order of appearance): Stanimira Georgieva, Risa Sawaki, Jess Kerlin, Ilja Sligte, Benjamin Crossey, Peter Winn, Fiona Lerigo, Uzma Satti,

Sophia Arutunov, Dante Abel Culcuy, Pablo Daniel Martinez, Verónica Rueco and last but not least Daniel Rivero, who helped me to stay sane during the last intense months of writing.

Last but not least, I would like to thank my family. They were always there for me and were my greatest motivation for going ahead and completing this PhD. Especially during the tough times they shielded me and allowed me to continue working. And finally, I would like to thank my father who is unfortunately not here anymore to see me completing this PhD.

These were some incredible years of immense professional and personal growth by the side of many amazing people.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

.

CHAPTER 1: GENERAL INTRODUCTION

Imagine standing at a busy crossroad. People are hurrying in all directions, cyclists are ringing their bells, cars are stopping at a traffic light, other cars are accelerating, dogs are barking. For you, this is a normal situation. You are easily able to discriminate people from cars, if a dog is walking quietly at his owner's side or is approaching you. You can even estimate the velocity at which each car is driving. Your experience of that crossroad is unified and continuous; you can easily navigate in this situation. However, what seems like a simple everyday perceptual experience to you is, in fact, a result of a multitude of computations that happen in your brain. The information that we extract from the environment enters the brain via different sensory channels. How is this information subsequently reconstructed to give us that uniform experience? What are the mechanisms in the brain that allow to extract information and to decide how to combine it? Which sound belongs to which object? Is that object approaching or withdrawing? These are just a few of those questions that the research into multisensory processing aims to answer.

Early accounts proposed a hierarchical perceptual process in which information from different sensory modalities is first processed separately in highly specialized areas of the brain and only later combined in convergence zones (Felleman & van Essen, 1991). However, this approach has been greatly challenged. More and more evidence was accumulating indicating that sensory brain areas are not as specialized and isolated as traditionally thought, that instead, interactions between

senses occur in many parts of the neocortex. In 2006, Ghazanfar and Schroeder even went so far as to propose that the whole neocortex is multisensory. Nowadays, it is considered established that interactions between different sensory modalities in the neocortex occur as early as at the primary cortical level (Noppeney, Jones, Rohe & Ferrari, 2018). The new challenge is to define the nature of these interactions and to understand how they contribute to creating a uniform percept.

In this thesis, I will describe how we used auditory pitch and visual size in polar coordinates to investigate the interplay between the auditory and visual sensory modalities at the behavioural level and at the neural level using functional Magnetic Resonance Imaging (fMRI).

In this chapter, I will introduce the foundations on which our research is based. First, I will present the evidence for cross-modal influences in the brain with the main emphasis on interactions at the primary cortical level. Then, I will motivate the choice of our specific stimuli, and finally present what is known so far about the interactions of pitch and size at the behavioural and neural level.

Cross-modal influences in early and primary auditory and visual cortices

In the past, sensory research was dominated by studying sensory modalities separately from each other. Multisensory integration was considered to occur at a relatively late stage of perceptual processing in higher-order association areas, convergence zones that receive input or respond to more than one sensory

modality (Felleman & van Essen, 1991). While the critical role of such classic multisensory areas like the superior temporal sulcus, the inferior parietal sulcus and frontal cortex was confirmed, there has also been accumulating evidence that the interplay between sensory modalities occurs already much earlier than that, even at the level of primary sensory cortices. Inspired by reports of cross-modal connections and interactions at multiple stages of perceptual processing in the brain, Ghazanfar and Schroeder (2006) even questioned the approach to study sensory modalities separately, as if they were operating independently at the neocortical level, and proposed instead that the whole neocortex contributes to sensory processing and is therefore multisensory. For further reading on this topic, I would like to refer to some extensive reviews (Ghazanfar & Schroeder, 2006; Driver & Noesselt, 2008; Klemen & Chambers, 2012; Noppeney et al., 2018; for a focus on anatomic pathways see: Cappe, Rouiller & Barone, 2009).

In the next part, I will focus on cross-modal interactions in early sensory cortices and their underlying mechanisms in the auditory and visual sensory modalities. Even though it is nowadays established that cross-modal influences occur at several levels of perceptual processing, there is a general consensus that especially early and primary sensory cortices exhibit a clear preference for a specific sensory modality. Therefore, I will from now on refer to preferred and non-preferred sensory modality and corresponding and non-corresponding cortices. Preferred sensory modality refers to the sensory modality that is predominantly processed in an area, the corresponding cortices. The non-preferred modality is any other sensory modality.

Cross-modal interactions can drive or modulate responses in non-preferred sensory areas. Visual or auditory stimuli presented unimodally elicit under certain circumstances activation in the preferred, and deactivation in the non-preferred sensory areas (Laurienti et al., 2002; Leitao, Thielscher, Werner, Pohmann & Noppeney, 2013). However, pairing stimuli from different modalities leads to facilitation in activations i.e. unimodal stimuli that previously induced deactivations in the non-corresponding cortices subsequently activate both sensory areas, cortices that correspond to the sensory modality of that stimulus and those cortices that correspond to the sensory modality with which it was linked (PET: McIntosh, Cabeza & Lobaugh, 1998; Zangenehpour & Zatorre, 2010; fMRI: Tanabe, Honda & Sadato, 2005; Baier, Kleinschmidt & Müller, 2006; Martuzzi et al., 2007; Meyer, Baumann, Marchina & Jancke, 2007).

Sometimes, even if the non-preferred modality itself does not elicit a response, it can, if presented concurrently with the preferred modality, modulate the response or information in the non-corresponding cortices (Allman & Meredith, 2007; Kayser, Petkov & Logothetis, 2008; Kayser, Logothetis & Panzeri, 2010). It has been demonstrated that in the extrastriate visual cortices of cats and ferrets there are besides classical unimodal neurons, also some bimodal neurons, neurons that respond to both auditory and visual stimulation, and, most interestingly, a large proportion of supposedly unimodal neurons that are influenced by subthreshold modulation caused by stimuli presented in a non-preferred modality (Allman & Meredith, 2007; Allman et al., 2008; Allman, 2009). Such modulations can be enhancing or suppressing, depending on the context i.e.

if the stimuli in the different modalities are congruent or incongruent, if they occur within a window of integration of further apart (Kayser et al., 2008).

It is important to note that even though both neuroimaging studies in humans (e.g. McIntosh et al., 1998; Laurienti et al., 2002; Tanabe et al., 2005; Baier et al., 2006; Martuzzi et al., 2007; Meyer et al., 2007; Zangenehpour & Zatorre, 2010; Leitao et al., 2013) and electrophysiological experiments in animals (e.g. Kayser et al., 2008, 2010; Iurilli et al., 2012; Ibrahim et al., 2016) reveal facilitation and suppression effects in response to cross-modal interactions, these methods do not necessarily measure the same underlying mechanisms (Kayser, Petkov & Logothetis, 2009). Activations and deactivations of BOLD-responses relate to a baseline that is specific for an experiment i.e. in terms of an audiovisual experiment a deactivation means that the BOLD-response in a given brain area was lower than during a baseline condition. In contrast, electrophysiological recordings provide a direct measure of neural activity. Kayser and colleagues (2009) dissociated in a careful study responses in single-unit recordings, local field potentials (LFP) and fMRI in auditory cortices of macaque monkeys during auditory, visual and concurrent audio-visual stimulation. LFP and BOLD-responses displayed weak but relatively widespread and consistent responses to unimodal visual stimuli. In contrast, only a small proportion of individual neurons showed multisensory modulation or responded to visual stimuli alone. Moreover, while BOLD and low-frequency LFPs elicited by visual stimuli were mostly enhancing, the responses of most individual neurons were suppressive. The latter mechanism

is consistent with reports of synaptic inhibition in primary visual cortices of mice during auditory stimulation (Iurilli et al., 2012; Ibrahim et al., 2016).

At the neocortical level, modulations between sensory modalities could, for instance, be mediated via anatomical connections. Tracing studies in non-human primates revealed direct connections from primary auditory cortices to peripheral visual field representations in the anterior portions of the primary visual cortices (Falchier, Clavagnier, Barone & Kennedy, 2002; Rockland & Ojima, 2003) and from V2 to caudal auditory cortices (Falchier et al., 2010).

Now, that the existence of cross-modal interactions at primary sensory level is widely confirmed, the next challenge for research on this topic is to investigate what the purpose of such interactions is and what kind of information is being exchanged at the different levels of sensory processing.

One way of addressing this question in humans is to use multivariate pattern analysis to decode the content of a unimodally presented sensory stimulus from the BOLD-response pattern that it elicited in non-corresponding sensory cortices. So far, it has been demonstrated that muted videos showing animals, objects or instruments can be distinguished from each other, not only across categories but also individual exemplars, based on BOLD-response pattern that they elicited in auditory cortices (Meyer at al., 2010). In a similar vein, videos depicting manmade objects were successfully decoded from auditory cortices and portions of the posterior superior temporal sulcus, but the isolated sounds corresponding to those videos were not decodable from visual cortices (Man, Kaplan, Damasio & Meyer, 2012). Furthermore, both natural and imagined sounds were decoded from early

visual cortices of blindfolded individuals (Vetter, Smith & Muckli, 2014) and the spatial location of visual stimuli was reliably decoded from auditory cortices (Liang, Mouraux & Iannetti, 2013).

Importantly, all of these studies used stimuli that included amodal properties like spatial or semantic congruency. These are properties that are shared between different modalities. It is still unknown if also modality-specific information is exchanged between sensory areas. This is a question that we decided to tackle experimentally using the modality-specific features auditory pitch and visual size in polar coordinates. In the next parts of this chapter, I will motivate this specific choice of stimulus features and present what is known about the interplay between them so far.

Auditory pitch and visual size in polar coordinates or eccentricity

In the previous paragraph, I introduced our intention to investigate audio-visual interactions in the auditory and visual sensory cortices using auditory pitch and visual size in polar coordinates. In this part, I will motivate the advantages of choosing these particular stimulus features.

Firstly, previous research has shown that overall cross-modal effects elicit weaker or even nearly sub-threshold responses compared to stimuli of the preferred sensory modality (e.g. Martuzzi et al, 2007, Man et al., 2012). Consequently, in order to increase the chances of capturing modality specific influences in a different modality, our choice of stimuli was determined by identifying modality specific auditory and visual features with robust neural

representations that are not driven by top-down effects following any of the classical multisensory integration cues. Critically, both auditory pitch and visual eccentricity have robust large-scale topographic representations in the neocortex. Frequency is the main dimension that defines the topographic maps in the auditory cortices (Formisano et al., 2003; Da Costa et al., 2011). Eccentricity, on another hand, defines together with spatial angle the retinotopic maps in the visual cortices (Wandell, Dumoulin & Brewer, 2007). If any relationship or even a mapping exists between these two stimulus features it is likely to be metaphoric i.e. pitch does not exist in the visual sensory modality and eccentricity in the classical sense does not exist in the auditory sensory modality. This makes auditory pitch and visual eccentricity good candidate stimulus features for exploring modality specific cross-modal interactions between the early auditory and visual cortices using fMRI.

Secondly, the continuous topographic representations allow exploring the stimulus spaces within these topographies. Based on the tonotopic and retinotopic maps we can investigate if there is a specific mapping between the two modalities. In a first step, we would like to find out if the maps are related in any specific way. Does low pitch correspond to small or large eccentricity? Is there a linear relationship? Is it possible to answer such questions by studying the BOLD-response patterns elicited by auditory and visual stimuli in non-corresponding cortices?

In a second step, inspired by the work of Baier and colleagues (2006) who demonstrated co-activation of the auditory and visual sensory cortices after previously unrelated auditory and visual stimuli were paired in a short learning

phase, we would like to look closer into BOLD-response profiles within the auditory and visual processing streams after creating an artificial mapping between auditory pitch and visual eccentricity. If learning is made topographically specific, can we induce an audiovisual mapping specific to topographic locations within the tonotopic map spanned by auditory pitch and the retinotopic mapping dimension that is spanned by eccentricity?

The classic stimuli for localizing eccentricity maps in the visual cortices are checkerboard rings, centered around fixation (Wandell et al., 2007). We are interested in studying eccentricity not only at the neural level but also in behavioral paradigms. Outside of the imaging context, rings that are centered around fixation and cover different eccentricities are more intuitively described simply as circles of different sizes. Therefore, in order to do justice to both the imaging and the behavioral context, I will formally refer to our visual stimuli as size in polar coordinates. When discussing experiments within the field-specific literature, I will refer to the visual stimuli simply as ‘size’ in the behavioral context, and as ‘eccentricity’ in the neural imaging context.

In the next two parts of this chapter, I will first introduce the literature dedicated to studies of the pitch-size relationship at the behavioral level, and in the second part, the literature about representations of pitch and eccentricity in non-corresponding sensory cortices.

Pitch and size mappings in behavioral experiments

Auditory pitch and visual size have been studied in a variety of contexts. The relationship between these two modality-specific stimulus features can be considered metaphoric because pitch does not exist in vision and size is not an auditory property. In the natural environment, a combination of pitch and size occurs predominantly in two contexts: static and dynamic. Importantly, abstract size stimuli like disks or circles, especially if fixation is directed towards the centre of the visual stimulus, can be interpreted differently: either as objects of different sizes (static context) or as the same object at different distances from the observer (dynamic context). In the natural environment, for example, the frequency of an animal call depends largely on the size of its resonance body i.e. its body size (von Kriegstein, Smith, Patterson, Ives, & Griffiths, 2007). The biological relevance of correctly interpreting changes in pitch and size within a dynamic context is essential for determining if an object is approaching or withdrawing. Does the sound you hear tell you that you should run away from an approaching predator or is this prey that you want to follow?

The relationship between pitch and size has been studied in both contexts with variable results. Several studies reported that high pitch is associated with small size and low pitch with large size in a variety of tasks e.g. speeded classification task (Gallace & Spence, 2006; Evans & Treisman, 2010; Eitan, Schupak, Gotler & Marks, 2014), temporal order judgment task (Parise & Spence, 2009), pitch-size ventriloquist paradigm (Bien, ten Oever, Goebel & Sack, 2012), free drawing of circles to match sounds of variable pitch (Tonelli, Cuturi & Gori, 2017; Ueda, Mizuguchi, Yakushijin & Ishiguchi, 2018), and also in children (Marks, Hammeal,

Bornstein & Smith, 1987; Mondloch & Maurer, 2004; Fernandez-Prieto, Navarra & Pons, 2015). However, also the opposite mapping has been reported: in a dynamic context (Eitan et al., 2014), and in a music notation context (Antovic, 2009). To add further to the controversy, it has to be noted that compared to other metaphoric mappings, for example like pitch-elevation (Bernstein & Edelstein, 1971; Melara & O'Brien, 1987; Ben-Artzi & Marks, 1995; Patching & Quinlan, 2002; Evans & Treisman, 2010), the pitch-size mapping develops later in life (Marks et al., 1987; Fernandez-Prieto et al, 2015). Furthermore, some studies did not find significant effects at all (Haryu & Kajikawa, 2012) or only after testing a different sample of participants (Mondloch & Maurer, 2004).

Taken together, these findings suggest the existence of a metaphoric mapping between pitch and size that points towards an association of high pitch/small size and low pitch/large size but that this mapping can be reversed depending on the context and that it is not as robust as other metaphoric mappings. Alternatively, it cannot be excluded that there are in fact two different mechanisms that mediate the metaphoric relationship between pitch and size in different ways for static context and for dynamic context.

We decided to use the speeded classification task to investigate how auditory pitch is related specifically to visual circles centered around fixation, stimuli that when presented in an fMRI experiment can elicit retinotopically specific activations in visual cortices. Such visual stimuli can be easily interpreted either as objects of different sizes or the same object at different distances of the observer.

Considering the controversial findings it is not straightforward which mapping had to be expected.

Human neural imaging of pitch and eccentricity representation in non-corresponding cortices

In this paragraph, I will briefly discuss studies that investigated the neural representation of pitch in the visual areas and visual eccentricity in auditory areas. According to our knowledge, no study so far has successfully demonstrated pitch representations in visual areas or visual eccentricity representations in auditory cortices of participants with intact hearing and vision. However, two studies of sensory deprived patients present evidence for the existence of such cross-modal representations. For example, in congenitally blind participants, sound stimuli activated visual cortices and, in a few of the participants, the area V5 even contained a tonotopic map (Watkins et al., 2013). Furthermore, the temporal cortices of congenitally deaf participants contain information that allows significant decoding of spatial position and eccentricity (Almeida et al., 2015). Both of these studies were not able to replicate their results in control participants with intact senses.

Importantly, these studies provide us with evidence that there are pre-existing pathways that can be strengthened under certain sensory conditions and allow cross-sensory representations of our stimulus features of choice. We would like to use a wider range of stimulus exemplars of pitch and size to revisit their cross-modal neural representations in sensory healthy participants and further attempt to

strengthen the mapping between pitch and size artificially in order to enhance the chances of finding topographically specific co-activations within the auditory and visual cortices.

Summary

In this chapter, I presented the literature and concepts that are essential for the understanding of the following empirical chapters. First, I reviewed the evidence for audio-visual interactions at the early and primary cortical level. Subsequently, I motivated our choice of auditory pitch and visual size in polar coordinates for the research described in the following empirical chapters. Finally, I reviewed the research into the relationship between pitch and size at the behavioral and at the neural level.

CHAPTER 2: SYNAESTHETIC INTERACTIONS ACROSS VISION AND AUDITION

The work presented in this chapter is a product of a collaboration between Alexandra Krugliak and Uta Noppeney. It was published in Krugliak and Noppeney (2016). The experiments were designed by AK and UN, the data was collected and analysed by AK (Supervised by UN), the manuscript was written by AK and UN.

Introduction

In daily life our brains are bombarded with myriad of signals perceived through different sensory modalities. Signals originating from a common event need to be integrated into one coherent percept and separated from other signals.

Temporal, spatial and semantic congruency are important cues that inform the brain whether signals originate from a common source and should be integrated (Wallace, Wilkinson & Stein, 1996; Laurienti, Kraft, Maldjian, Burdette & Wallace, 2004; van Atteveldt, Formisano, Goebel & Bloemert, 2004; Wallace et al., 2004; Macaluso & Driver, 2005; van Atteveldt, Formisano, Bloemert & Goebel, 2007; Adam & Noppeney, 2010; Lewis & Noppeney, 2010; Vroomen & Keetels, 2010; Donohue, Roberts, Grent-'t-Jong & Woldorff, 2011; Lee & Noppeney, 2011, 2014). In addition to these classical congruency cues more abstract feature correspondences can also influence the binding of signals from multiple sensory modalities.

The most pronounced examples are synaesthetic experiences binding letters with colours or colours with sounds (Rich, Bradshaw & Mattingley, 2005; Brang, Rouw, Ramachandran & Coulson, 2011). Yet, even in non-synesthetic individuals perceptual experiences and decisions are influenced by a wide range of multisensory metaphoric mappings including frequency-size (Marks, Hammeal, Bornstein & Smith, 1987; Mondloch & Maurer, 2004; Gallace & Spence, 2006; Antovic, 2009; Evans & Treisman, 2010; Parise & Spence, 2009; Bien, ten Oever, Goebel, & Sack, 2012; Parise & Spence, 2012; Eitan, Schupak, Gotler & Marks, 2014), dynamic pitch–dynamic size (Kim & Iwamiya, 2008; Eitan et al., 2014; Fernandez-Prieto, Navarra & Pons, 2015), and dynamic pitch–directional motion (Sadahiani, Maier & Noppeney, 2009) (for reviews see Marks, 2004; Spence, 2011; Spence & Deroy, 2013). For example, human observers perceive bright objects as louder than dark objects (loudness-brightness: Marks, 1987). They also tend to associate high-pitch sounds predominantly with visual objects at higher elevation (frequency-elevation: Bernstein & Edelstein, 1971; Melara & O'Brien, 1987; Ben-Artzi & Marks, 1995; Patching & Quinlan, 2002; Evans & Treisman, 2010).

Multiple mechanisms have been proposed to mediate metaphoric relationships. One account posits that metaphoric mappings are mediated via shared semantics or language. For instance, pitch is referred to by words such as 'high' or 'low'. Moreover, musical notation relies on spatial concepts (Martino & Marks, 1999; Ashley, 2004). Hence, interactions between pitch in the auditory sense and elevation in the visual sense may be mediated via a common conceptual reference

frame. Alternatively, seemingly arbitrary metaphoric mappings may in fact be grounded in natural environmental statistics. In line with this conjecture, a recent elegant study by Parise, Knorre, and Ernst (2014) revealed that the mapping between frequency and elevation is grounded in auditory scene statistics where high-frequency sounds tend to originate from elevated sources. Moreover, the filtering characteristics of the outer ear also contributed to the mapping between elevation and frequency.

In a similar vein, the metaphoric relationship between auditory frequency and visual object size has been proposed to emerge from the fact that the frequency of sounds made by animals or musical instruments depends on the size of the resonator (von Kriegstein, Smith, Patterson, Ives, & Griffiths, 2007). In other words, high-pitched sounds should be associated with small objects and low-pitched sounds with large objects (Marks et al., 1987; Mondloch & Maurer, 2004; Gallace & Spence, 2006; Parise & Spence, 2009; Evans & Treisman, 2010; Bien et al., 2012; Parise & Spence, 2012; Eitan et al., 2014; except: Antovic, 2009).

While accumulating evidence associates high-frequency sounds with small objects and vice versa in a static context, controversial evidence has been provided for dynamic contexts. Here, ascending pitch has surprisingly been associated with growing size (Eitan et al., 2014). Amongst other mechanisms, the authors attributed this opposite pattern for dynamic stimuli to the Doppler Effect whereby an approaching object induces a change in pitch. This experiment suggests an ambivalent association between pitch and size in our natural dynamic world. In dynamic contexts, the brain would need to dissociate whether the size as

estimated from a retinotopic representation derives predominantly from the constant size of the object in the natural world or its distance from the observer. This more complex relationship between constant and dynamic size-pitch relationship may explain why the correspondence between pitch and size develops relatively late in life (in dynamic context at the age of 6 months but not 4 months: Fernandez-Prieto et al., 2015; in static context gradually emerging from the age of 9 years: Marks et al., 1987) and has been found only inconsistently (Mondloch & Maurer, 2004; Haryn & Kajikawa, 2012).

This study revisits the pitch-size relationship in a static context. Participants were presented with large or small circles/discs in synchrony with high- or low-frequency sounds in an auditory or visual selective attention paradigm (Bernstein & Edelstein, 1971). In the visual modality, they discriminated between large and small visual size. In the auditory modality, they discriminated between high- and low-pitched tones. As luminance may be a confounding factor when varying the size of a visual stimulus, the visual discs were either brighter or darker than the background colour. Likewise, loudness and sound amplitude can be potential confounds that we evaluated by equating the sounds either with respect to their physical sound amplitude or their perceptual loudness.

Experiment 1 & 2

Methods

Participants

After giving written informed consent, 16 participants (12 female, mean age: 24 years) took part in Experiment 1 and 10 participants (4 female, mean age: 23 years) in Experiment 2. Each had normal or corrected-to-normal vision, reported normal hearing, and had no history of neurological or psychiatric illness. Participation was rewarded with course credits. The study was approved by the local research and ethics committee.

Stimuli

Visual stimuli were either circles (line thickness: 0.5° visual angle) or discs. The radius of both circles and discs was either 2.8° or 7.7° visual angle. Experiment 1 presented circles or discs in lighter grey (mean luminance: 50.08 cd/m^2) than the grey shade of the background (mean luminance: 33.58 cd/m^2). Experiment 2 presented circles or discs in darker grey (mean luminance: 33.58 cd/m^2) than the grey shade of the background (mean luminance: 50.08 cd/m^2). The comparison between Experiment 1 and 2 allows us to assess confounding effects of luminance variation on the pitch-size association. This is important, because previous studies have demonstrated that pitch is not only associated with size but also with brightness (Marks, 1987; Marks et al., 1987). Yet, overall brightness differs between (i) circles and discs and (ii) in particular discs of different sizes.

Auditory stimuli were pure tones of 120 ms duration with linear onset and offset ramps of 1 ms to avoid auditory clicks (sampling rate 44100 Hz). The frequency was either 1250 Hz (low pitch) or 3000 Hz (high pitch).

Experimental design

The 3 x 2 factorial design manipulated: (i) visual stimuli (circles or discs), (ii) task-relevant modality (respond to the auditory or to the visual stimuli), and (iii) mapping (mapping 1: low pitch, large size and high pitch, small size; mapping 2: low pitch, small size and high pitch, large size).

On each trial participants were presented with an audiovisual stimulus (120 ms duration, SOA 1500 ms) defined by pitch (high, low) and size (large, small). Thus, four audiovisual stimulus combinations were presented with equal probability: low pitch/large visual size, low pitch/small visual size, high pitch/large visual size and high pitch/small visual size. We will refer to the stimulus combinations low/large and high/small as mapping 1 and to the stimulus combinations low/small and high/large as mapping 2 (Figure 2.1*).* In Experiment 1, the visual disc was brighter than the background. In Experiment 2, the visual disc was darker than the background. In a selective attention paradigm, participants performed a two-choice discrimination task that focussed either on the auditory frequency or the visual size dimension. Participants discriminated between small and large size in the visual task or high and low pitch in the auditory task as fast and accurately as possible. Further, they were instructed to fixate a central fixation cross throughout the entire experiment.

The main experiment included two runs presenting ‘circles’ or ‘discs’ respectively as visual stimuli. Each run consisted of 12 auditory and 12 visual attention task blocks that were presented in permuted order to facilitate interference effects. The task-relevant sensory modality was indicated at the

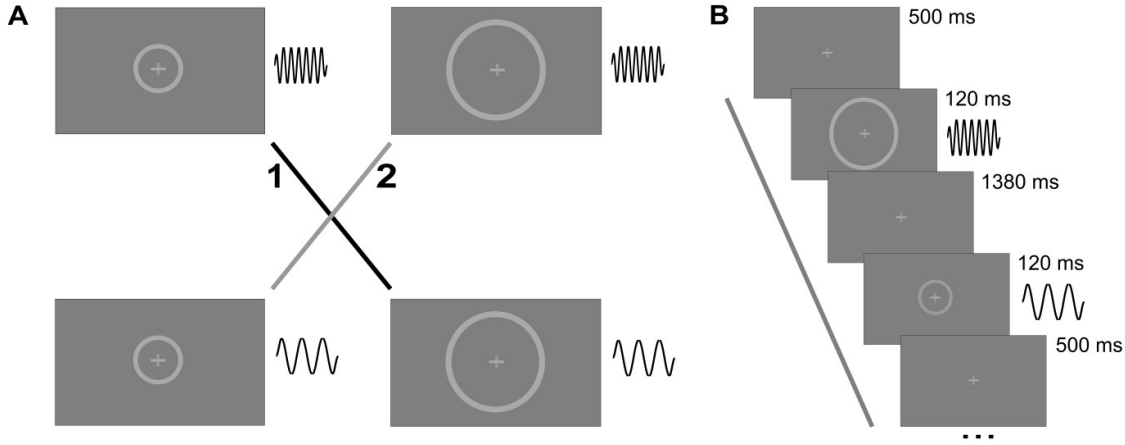


Figure 2.1 Experimental design and example trial. A. Experimental design: Each experiment compared two mappings: mapping 1: high pitch with small size and low pitch with large size; mapping 2: high pitch with large size and low pitch with small size. B. Example trial: On each trial participants were presented with a visual circle (or disc) and an auditory pure tone. In the auditory task, they discriminated between high and low pitched tones. In the visual task, they discriminated between small and large sized visual stimuli.

beginning of each block. Within each block each of the four possible audiovisual stimulus combinations were presented twice in random order. The order of auditory/visual tasks and circle/disc runs was counterbalanced across participants. The start of each block was initiated by button press in order to allow participants to switch between the different response-mappings for the auditory and visual task.

Responses were given via four different buttons: ‘A’, ‘D’, ‘J’ and ‘K on a conventional keyboard. The buttons were chosen to ensure that participants used

different hands and fingers to respond in order to avoid inducing a mapping between auditory and visual parameters at the response level. The mapping of hands and response buttons was counterbalanced across participants resulting in eight response mapping options:

1. auditory left hand and visual right hand: (i) ‘A’ = low (auditory), ‘D’ = high (auditory), ‘J’ = small (visual), ‘K’ = large (visual), (ii) ‘A’ = low (auditory), ‘D’ = high (auditory), ‘J’ = large (visual), ‘K’ = small (visual), (iii) ‘A’ = high (auditory), ‘D’ = low (auditory), ‘J’ = small (visual), ‘K’ = large (visual), (iv) ‘A’ = high (auditory), ‘D’ = low (auditory), ‘J’ = large (visual), ‘K’ = small (visual)
2. auditory right hand and visual left hand: (i) ‘J’ = low (auditory), ‘K’ = high (auditory), ‘A’ = small (visual), ‘D’ = large (visual), (ii) ‘J’ = low (auditory), ‘K’ = high (auditory), ‘A’ = large (visual), ‘D’ = small (visual), (iii) ‘J’ = high (auditory), ‘K’ = low (auditory), ‘A’ = small (visual), ‘D’ = large (visual), (iv) ‘J’ = high (auditory), ‘K’ = low (auditory), ‘A’ = large (visual), ‘D’ = small (visual)

Apparatus

The experiment was conducted in a dimly lit experimental room. Constant viewing distance was ensured by stabilizing the participant’s head on a chin rest at a distance of 50 cm from a LED monitor (1920 × 1080 resolution, 60 Hz refresh rate, iiyama Proline, Japan). Auditory stimuli were presented through headphones (Sennheiser HD 555MR, Germany) at approximately 75 dB SPL. Experimental sessions were presented using Cogent 2000 v1.25 (developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John

Romaya at the LON at the Wellcome Department of Imaging Neuroscience, UCL, London, UK; http://www.vislab.ucl.ac.uk/cogent.php) running under MATLAB (Mathworks Inc., Natick, MA, USA) on a Windows PC. The responses were given via a conventional keyboard.

Analysis

Reaction times (based on within-subject median after excluding incorrect trials and trials with reaction times shorter than 200 ms or longer than 1000 ms) and accuracy (Table 2.1) were entered into independent 2 (visual stimulus: circle vs. disc) x 2 (task-relevant modality: auditory vs. visual) x 2 (mapping: 1 vs. 2) factorial repeated-measures ANOVAs.

Results

Experiment 1

Participants responded to audiovisual stimuli comprised of a high- or a low-pitched tone and a small or large circle/disc. In an auditory task they classified the tones as high or low. In a visual task they classified circles/discs as small or large. In Experiment 1, two runs of the speeded classification task were presented: The visual stimuli in one run were circles and in the other run they were bright discs.

A 2 (visual stimulus: circle vs. disc) x 2 (task-relevant modality: auditory vs. visual) and 2 (mapping: 1 vs. 2) repeated-measures ANOVA on reaction times

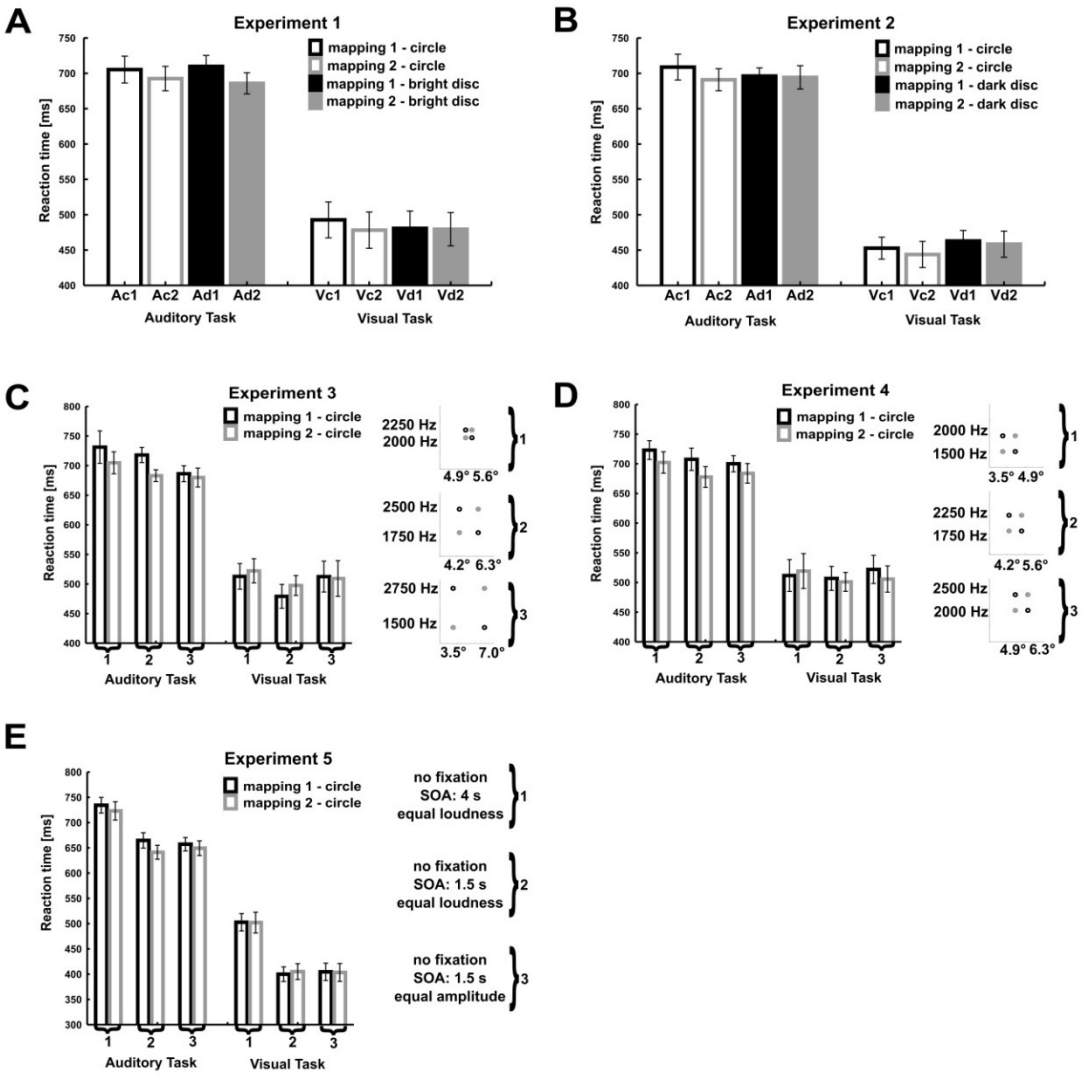


Figure 2.2 Results. Experiment 1 (A) and 2 (B): Bar plots showing reaction times (across subjects' mean +/- *SEM)* for circle or disc stimuli from mapping 1 or 2. A = auditory task, V = visual task, c = circle, d = disc, 1 or 2 = mapping 1 or 2. Experiment 3 (C) and 4 (D): Bar plots showing reaction times (across subjects' mean +/- *SEM*) for circle stimuli from mapping 1 or 2. The inserts show the configuration of audio-visual combinations of the chosen two pitch and two size parameters in a space spanned by size on the x-axis and frequency on y-axis. Experiment 3 manipulated the similarity between the two stimuli: 1 = small, 2 =

medium, 3 = large. Experiment 4 manipulated the relative mapping between pitch and size as shown in the inset. Experiment 5 (E). Experiment 5 manipulated the parameters SOA duration and sound equalization as shown in the insert.

revealed a significant main effect of mapping and task-relevant modality (Table 2.2). Further, it identified a significant three-way interaction between visual stimulus type, task-relevant modality and mapping.

A follow-up 2 (visual stimulus: circle or disc) x 2 (mapping: 1 vs. 2) repeated-measures ANOVA for the auditory task demonstrated that participants responded significantly faster to stimuli from mapping 2 than 1 irrespective of whether the visual stimuli were discs or circles (Table 2.3). For the visual task we observed a significant two-way interaction between visual stimulus and mapping (Table 2.3).

A follow-up 2 (mapping: 1 vs. 2) repeated-measures ANOVA per visual stimulus revealed that participants were faster for mapping 2 than mapping 1 predominantly when the stimuli were circles ($F(1,15) = 8.64$, $p = .010$) but not discs ($F(1,15) = 0.04$, $p = .836$).

A 2 (visual stimulus: circle vs. disc) x 2 (task-relevant modality: auditory vs. visual) and 2 (mapping: 1 vs. 2) repeated-measures ANOVA on % discrimination accuracy did not reveal any significant effects (Table 2.2).

In summary, Experiment 1 provides initial evidence that participants in our study associated low pitch with small size and high pitch with large size.

Table 2.1 Summary of mean reaction time (in ms) and accuracy (% correct) and their standard mean errors for Experiment 1 and 2.

	mapping 1		mapping 2	
	reaction time	accuracy	reaction time	accuracy
Experiment 1				
auditory				
circle	692.53 (17.28)	0.93 (0.03)	705.28 (19.86)	0.93 (0.01)
disc	685.84 (14.98)	0.93 (0.02)	709.75 (15.55)	0.92 (0.02)
visual				
circle	478.19 (25.69)	0.95 (0.01)	492.72 (25.36)	0.94 (0.02)
disc	479.50 (23.61)	0.95 (0.01)	480.72 (24.56)	0.95 (0.01)
Experiment 2				
auditory				
circle	690.95 (15.64)	0.87 (0.04)	708.85 (18.29)	0.86 (0.06)
disc	694.30 (16.52)	0.89 (0.04)	696.20 (11.54)	0.89 (0.04)
visual				
circle	443.80 (18.46)	0.95 (0.02)	452.75 (15.56)	0.94 (0.02)
disc	458.45 (18.45)	0.92 (0.04)	462.90 (14.90)	0.94 (0.03)

Table 2.2 Statistical results of Experiment 1 and 2.

	Experiment 1 (df: 1,15)		Experiment 2 (df: 1,9)	
	reaction time	accuracy	reaction time	accuracy
visual stimulus	F = 0.25	F = 0.00	F = 0.36	F = 0.38
	p = .623	p = .993	p = .566	p = .552
task-relevant modality	F = 255.79	F = 2.12	F = 1444.61	F = 4.77
	p < .001*	p = .166	p < .001*	p = .057
mapping	F = 11.83	F = 0.21	F = 5.75	F = 0.06
	p = .004*	p = .653	p = .040*	p = .817
visual stimulus x task-relevant modality	F = 0.18	F = 0.52	F = 0.90	F = 2.67
	p = .678	p = .480	p = .367	p = .137
visual stimulus x mapping	F = 0.07	F = 0.03	F = 0.76	F = 0.55
	p = .803	p = .867	p = .407	p = .476
task-relevant modality x mapping	F = 4.42	F = 0.08	F = 0.23	F = 0.56
	p = .053	p = .785	p = .640	p = .473
visual stimulus x task-relevant modality x mapping	F = 7.34	F = 0.59	F = 0.26	F = 0.25
	p = .016*	p = .453	p = .625	p = .628

*$p < 0.05$

Table 2.3 Statistical results of follow-up 2-way ANOVAs on reaction time for auditory and visual task of Experiment 1.

	Experiment 1 (df: 1,15)	
	auditory task	visual task
visual stimulus	F = 0.02 p = .884	F = 0.37 p = .552
mapping	F = 15.22 p < .001*	F = 3.21 p = .093
visual stimulus x mapping	F = 3.29 p = .090	F = 4.57 p = .049*

*$p < 0.05$

Experiment 2

Experiment 2 differed from Experiment 1 in that we presented dark discs instead of bright discs in order to access confounding effects of brightness on the pitch-size association.

A 2 (visual stimulus: circle vs. disc) x 2 (task-relevant modality: auditory vs. visual) and 2 (mapping: 1 vs. 2) repeated-measures ANOVA on reaction times revealed a significant main effect of mapping and task-relevant modality (Table 2.2). Participants responded faster during the visual than the auditory task. Most importantly, they responded faster to stimuli that combined low pitch tones with small circles/discs or high pitch tones with large circles/discs (mapping 2), than stimuli that combined low pitch tones with large circles/discs or high pitch tones with small circles/discs (mapping 1).

A 2 (visual stimulus: circle vs. disc) x 2 (task-relevant modality: auditory vs. visual) and 2 (mapping: 1 vs. 2) repeated-measures ANOVA on % discrimination

accuracy did not reveal any significant effects, only a marginally significant effect of task-relevant modality (Table 2.2).

Thus, Experiment 1 and 2 provide convergent evidence that participants associate low pitch with small size and high pitch with large size.

Comparison Experiment 1 vs. 2: Visual discs

To further investigate whether differences in overall luminance between large and small visual stimuli may contribute to the pitch-size mapping, we directly compared Experiments 1 and 2 in a 2 (stimulus luminance: high vs. low) x 2 (mapping: 1 vs. 2) x 2 (task-relevant modality: auditory vs. visual) repeated-measures ANOVA limited to the conditions where discs were presented with different luminance (n.b. the circle stimuli were identical in the two Experiments). This repeated-measures ANOVA replicated the main effects of task-relevant modality ($F(1,24) = 510.32$, $p < .001$) and mapping ($F(1,24) = 4.37$, $p = .047$). In addition, we also observed a marginally significant two-way interaction between task-relevant modality and mapping ($F(1,24) = 3.13$, $p = .090$). These results raise the possibility that changes in overall luminance may interfere with size-pitch relationships.

Experiment 3 & 4

Introduction

Surprisingly, the results of Experiment 1 and 2 provided consistent evidence for a congruency pattern that is opposite to the profile previously described in the literature (Marks et al., 1987; Mondloch & Maurer, 2004; Gallace & Spence, 2006;

Parise & Spence, 2009; Evans & Treisman, 2010; Bien et al., 2012; Parise & Spence, 2012; Eitan et al., 2014). In contrast to previous reports we observed faster responses for low pitch/small visual size and high pitch/large visual size stimuli (mapping 2). In order to test whether the opposite results might be caused by our choice of size and pitch parameters, we explored the space of parameters governing the pitch-size relationship. First, we manipulated the similarity between the two size-pitch stimuli that needed to be discriminated. We hypothesized that congruency/interference effects would be stronger when stimuli are more difficult to discriminate (i.e. the stimuli of the two classes differ less in pitch/size). Second, we varied the relative mapping between pitch and size across different runs. As a consequence, pitch-size pairings that are classified as low/small in one condition are classified as high/large in a different condition. If human observers associate pitch and size in an absolute fashion, some mappings may be more effective than others.

As Experiment 1 and 2 did not reveal any significant differences between circles and discs, we focussed on circles to limit the associated changes in overall luminance.

Methods

Participants

Ten participants (6 female, mean age: 24 years) with no history of neurological or psychiatric illness participated in both Experiment 3 and 4 after giving written informed consent. All of them had normal hearing and normal or corrected-to-

normal vision. Participation was rewarded with course credits. One participant was excluded from the analysis because he/she pushed all buttons of the keyboard in a random fashion. The study was approved by the local research and ethics committee.

Stimuli

Visual stimuli were circles (line thickness: 0.5° visual angle) with radii of 3.5°, 4.2°, 4.9°, 5.6°, 6.3° and 7°. Auditory stimuli were pure tones of 120 ms duration with frequencies of 1500 Hz, 1750 Hz, 2000 Hz, 2250 Hz, 2500 Hz and 2750 Hz, sampled at 44100 Hz, with linear onset and offset ramps of 1 ms to avoid auditory clicks.

Experimental design

This experimental series included two experiments with 3 runs. The experimental design in each run conformed to 2 (task-relevant modality: auditory vs. visual) by 2 (mapping: 1 vs. 2) factorial design.

Experiment 3

Experiment 3 investigated the effect of similarity between the two pitch-size stimuli that needed to be discriminated in separate runs. Specifically, we included runs with three different class similarities: (i) small with sound frequency: 2000 Hz vs. 2250 Hz and circle radius: 4.9° vs. 5.6°, (ii) medium with sound frequency: 1750

Hz vs. 2500 Hz and circle radius: 4.2° vs. 6.3°, and (iii) large with sound frequency: 1500 Hz vs. 2750 Hz and circle radius: 3.5° vs. 7°.

Experiment 4

Experiment 4 manipulated the relative mapping of high/low sound frequency and the small/large visual size in separate runs while holding the similarity between the two stimulus classes constant. Specifically, we included runs where the high/low sound frequencies and small/large circles were sampled from three different ranges: (i) sound frequency: 1500 Hz vs. 2000 Hz and circle radius: 3.5° vs. 4.9°, (ii) sound frequency: 1750 Hz vs. 2250 Hz and circle radius: 4.2° vs. 5.6°, and (iii) sound frequency: 2000 Hz vs. 2500 Hz and circle radius: 4.9° vs. 6.3°. Hence, the stimulus pairing 2000 Hz / 4.9° was classified as high in condition 1, but as low in condition 3. If participants consider size-pitch not only in relative terms, but also have some absolute scale, the results may depend on the exact pitch-size pairing.

At the beginning of the experimental session, participants were familiarized with 16 trials of the first parameter setting. Afterwards each new parameter setting was introduced by sequentially displaying the auditory and visual stimuli to ensure that the labels for ‘low’ and ‘high’ pitch and ‘small’ and ‘large’ circles were assigned correctly. Each run included 10 auditory and 10 visual task blocks. The order of the blocks was permuted. The starting modality of the first block and the order of the parameter settings were counterbalanced across participants. Unlike in Experiment 1 and 2, we reduced the number of mapping options to allow for counterbalancing with a smaller number of subjects. Subjects responded with the

left hand on the visual task and with the right hand on the auditory task. The mapping of response buttons was fully counterbalanced across participants resulting in four response mapping options. Otherwise, the experimental procedures and apparatus were identical to Experiment 1 and 2.

Analysis

For each Experiment, reaction times (based on within-subject median, after excluding incorrect trials and trials with reaction times shorter than 200 ms or longer than 1000 ms) and accuracy (Table 2.4) were entered into independent n (n x parameter settings) x 2 (task-relevant modality: auditory vs. visual) x 2 (mapping: 1 vs. 2) factorial repeated-measures ANOVAs.

Results

Experiment 3

In Experiment 3, we varied the similarity between the two pitch-size stimuli that needed to be discriminated.

A 3 (similarity between stimulus classes: small, medium, large) x 2 (task-relevant modality: auditory vs. visual) and 2 (mapping: 1 vs. 2) repeated-measures ANOVA on reaction times revealed a significant main effect of task-relevant modality and a significant two-way interaction between task-relevant modality and mapping (Table 2.5). Moreover, we observed a marginally significant main effect of class similarity indicating that stimulus discriminability influences audiovisual congruency/interference effects and a marginally significant two-way interaction

between parameter options and mapping. A follow-up 3 (class similarity: small, medium, large) x 2 (mapping: 1 vs. 2) repeated-measures ANOVA for the auditory task showed that participants responded significantly faster to stimuli from mapping 2 than to stimuli from mapping 1 (Table 2.7). The main effect of class similarity was only marginally significant (Table 2.7). For the visual task these effects were only marginally significant (Table 2.7).

A 3 (class similarity: small, medium, large) x 2 (task-relevant modality: auditory vs. visual) and 2 (mapping: 1 vs. 2) repeated-measures ANOVA on % discrimination accuracy did not reveal any significant effects (Table 2.5).

In summary, Experiment 3 provides further evidence that participants in our study responded faster to stimuli from mapping 2 than to stimuli from mapping 1.

Experiment 4

In Experiment 4, we manipulated the relative mapping of high/low pitch and small/large visual size.

A 3 (parameter mapping: 1, 2, 3) x 2 (task-relevant modality: auditory vs. visual) and 2 (mapping: 1 vs. 2) repeated-measures ANOVA on reaction times revealed a main effect of task-relevant modality and a marginally significant effect of mapping (Table 2.6). These results suggest that irrespective of the exact mapping between pitch and size at least within the range tested participants were again slower to respond to stimuli from mapping 1 than mapping 2.

A 3 (parameter mapping: 1, 2, 3) x 2 (task-relevant modality: auditory vs. visual) and 2 (mapping: 1 vs. 2) repeated-measures ANOVA on % discrimination accuracy did not reveal any significant effects (Table 2.6).

Taken together, Experiment 4 revealed that participants responded faster to stimuli with relative low pitch/small visual size and relative high pitch/large visual size (mapping 2), irrespective of the absolute values of the pitch and the size parameters.

Table 2.4 Summary of mean reaction time (in ms) and accuracy (% correct) and their standard mean errors for Experiment 3 and 4.

	mapping 1		mapping 2	
	reaction time	accuracy	reaction time	accuracy
Experiment 3				
auditory				
1	704.78 (8.59)	0.93 (0.04)	731.28 (27.41)	0.89 (0.05)
2	683.00 (9.98)	0.94 (0.02)	718.00 (12.85)	0.92 (0.02)
3	679.89 (16.01)	0.92 (0.03)	686.33 (13.49)	0.93 (0.03)
visual				
1	522.22 (20.49)	0.95 (0.02)	512.83 (21.63)	0.97 (0.01)
2	497.50 (16.78)	0.95 (0.02)	479.11 (20.16)	0.94 (0.02)
3	509.28 (30.13)	0.94 (0.02)	512.50 (26.07)	0.93 (0.02)
Experiment 4				
auditory				
1	702.22 (18.11)	0.94 (0.02)	723.06 (15.80)	0.93 (0.03)
2	677.72 (17.58)	0.92 (0.03)	707.56 (18.81)	0.94 (0.02)
3	683.89 (16.52)	0.95 (0.03)	700.06 (13.64)	0.92 (0.03)
visual				
1	519.22 (29.36)	0.94 (0.02)	511.72 (26.83)	0.97 (0.01)
2	501.00 (15.97)	0.96 (0.02)	506.89 (20.20)	0.97 (0.01)
3	505.78 (22.06)	0.97 (0.01)	522.00 (23.73)	0.97 (0.01)

Table 2.5 Statistical results of Experiment 3.

	Experiment 3	
	reaction time	accuracy
parameter options	F(1.68,13.45) = 3.69 p = .059	F(1.82,14.56) = 0.09 p = .904
task-relevant modality	F(1,8) = 144.16 p < .001*	F(1,8) = 1.02 p = .325
mapping	F(1,8) = 4.65 p = .063	F(1,8) = 2.23 p = .173
parameter options x task-relevant modality	F(1.96,15.68) = 3.61 p = .052	F(1.61,12.91) = 0.78 p = .454
parameter options x mapping	F(1.74,13.93) = 0.07 p = .912	F(1.96,15.67) = 0.32 p = .724
task-relevant modality x mapping	F(1,8) = 15.11 p = .005*	F(1,8) = 1.12 p = .321
parameter options x task-relevant modality x mapping	F(1.40,11,18) = 2.78 p = .116	F(1.48,11.83) = 1.89 p = .196

*$p < 0.05$

Table 2.6 Statistical results of Experiment 4.

	Experiment 4	
	reaction time	accuracy
parameter options	F(1.56,12.37) = 1.07 p = .356	F(1.29,10.31) = 0.06 p = .866
task-relevant modality	F(1,8) = 132.93 p < .001*	F(1,8) = 2.40 p = .160
mapping	F(1,8) = 5.33 p = .050*	F(1,8) = 0.17 p = .694
parameter options x task-relevant modality	F(1.77,14.19) = 0.70 p = .496	F(1.56,12.51) = 0.19 p = .778
parameter options x mapping	F(1.85,14.78) = 0.51 p = .596	F(1.30,10.39) = 1.05 p = .352
task-relevant modality x mapping	F(1,8) = 2.29 p = .169	F(1,8) = 1.79 p = .218
parameter options x task-relevant modality x mapping	F(1.70,13.59) = 1.63 p = .226	F(1.28,10.17) = 1.05 p = .350

*$p < 0.05$

Table 2.7 Statistical results of follow-up 2-way ANOVAs on reaction time for auditory and visual task of Experiment 3.

	Experiment 3	
	auditory task	visual task
similarity	F(1.05,8.39) = 4.15 p = .073	F(1.70,13,62) = 3.17 p = .080
mapping	F(1,8) = 14.12 p = .006*	F(1,8) = 3.79 p = .087
similarity x mapping	F(2.00,16.00) = 1.80 p = .196	F(1.59,12.69) = 0.97 p = .387

*$p < 0.05$

Experiment 5

Introduction

In a series of control experiments we investigated whether the fixation instructions, perceptual loudness or stimulus onset asynchrony could explain the discrepancy between our results and previous reports (Marks et al., 1987; Mondloch & Maurer, 2004; Gallace & Spence, 2006; Parise & Spence, 2009; Evans & Treisman, 2010; Bien et al., 2012; Parise & Spence, 2012; Eitan et al., 2014). First, we removed the fixation cross and fixation instructions, while all other factors were identical to Experiment 1. Next, we equated the two sounds with respect to loudness (Suzuki & Takeshima, 2004). Finally, we increased the SOA to 4000 ms to prevent participants from perceiving or interpreting the change in size/pitch across successive stimuli as dynamic motion.

Like in Experiment 3 and 4 we focussed on circles to control for changes in overall luminance.

Methods

Participants

Sixteen participants (10 female, mean age: 26 years) with no history of neurological or psychiatric illness participated in this study after giving written informed consent. All of them had normal hearing and normal or corrected-to-normal vision. Participation was rewarded with course credits. The study was approved by the local research and ethics committee.

Stimuli

Visual stimuli were circles (line thickness: 0.5° visual angle) with radii of 2.8° and 7.7° - identical to Experiment 1 and 2 except for absence of fixation cross. Auditory stimuli were pure tones of 120 ms duration with frequencies of 1250 Hz and 3000 Hz, sampled at 44100 Hz, with linear onset and offset ramps of 1 ms to avoid auditory clicks - identical to Experiment 1 and 2. The sounds were corrected for equal loudness by presenting the 1250 Hz tone at 70 dB and the 3000 Hz tone at 65 dB (Suzuki & Takeshima, 2004).

Experimental design

The experimental design conformed to 3 (parameter options: long SOA and equal loudness, short SOA and equal loudness, short SOA and equal amplitude), 2 (task-relevant modality: auditory vs. visual) by 2 (mapping: 1 vs. 2) factorial design.

At the beginning of the experiment, participants were familiarized with the task in 16 example trials. Each run included 12 auditory and 12 visual task blocks. The order of the blocks was permuted. The starting modality of the first block was counterbalanced across participants. The parameter options were presented in the following order: first, long SOA and equal loudness, second, short SOA and equal loudness, and third, short SOA and default loudness setting. This particular order was chosen to avoid carry-over effects (e.g. dynamic perception for short SOA may be transferred to long SOA). None of the experimental runs presented a fixation cross or instructed participants to fixate. The experimental procedures and apparatus were otherwise identical to Experiments 1 and 2.

Analysis

Reaction times (based on within-subject median, after excluding incorrect trials and trials with reaction times shorter than 200 ms or longer than 1000 ms) and accuracy (% correct) (Table 2.8) were entered into independent 3 (3 x parameter options: long SOA and equal loudness, short SOA and equal loudness, short SOA and equal amplitude) x 2 (task-relevant modality: auditory vs. visual) x 2 (mapping: 1 vs. 2) factorial repeated-measures ANOVAs.

Results

In Experiment 5, we investigated if removing fixation instructions, increasing the stimulus onset asynchrony and equating the auditory stimuli for perceptual loudness has an effect on participants pitch-size association.

A 3 (parameter options: long SOA and equal loudness, short SOA and equal loudness, short SOA and equal amplitude) x 2 (task-relevant modality: auditory vs. visual) and 2 (mapping: 1 vs 2) repeated-measures ANOVA on reaction times revealed a significant main effects of parameter settings and task-relevant modality and a marginally significant main effect of mapping. Furthermore, it revealed a significant two-way interaction between task-relevant modality and mapping (Table 2.9).

A follow-up 3 (parameter options: long SOA and equal loudness, short SOA and equal loudness, short SOA and equal amplitude) x 2 (mapping: 1 vs. 2) repeated-measures ANOVA for the auditory task showed that participants responded significantly faster to stimuli from mapping 2 than to stimuli from mapping 1 (Table 2.10). For the visual task these effects were not significant (Table 2.10).

Table 2.8 Summary of mean reaction time (in ms) and accuracy (% correct) and their standard mean errors for Experiment 5.

	mapping 1		mapping 2	
	reaction time	accuracy	reaction time	accuracy
Experiment 5				
auditory				
1	734.43 (61.76)	1.00 (0.00)	723.12 (72.40)	1.00 (0.00)
2	664.72 (60.72)	1.00 (0.00)	641.31 (54.97)	1.00 (0.00)
3	657.25 (52.74)	0.95 (0.06)	649.31 (57.97)	0.97 (0.04)
visual				
1	502.78 (68.79)	1.00 (0.00)	502.09 (82.00)	1.00 (0.00)
2	399.97 (57.77)	1.00 (0.00)	405.03 (62.59)	1.00 (0.00)
3	404.63 (68.54)	0.97 (0.03)	403.34 (70.40)	0.98 (0.02)

Furthermore, in both the auditory and visual task participants responded faster for short than long SOAs (Table 2.10).

A 3 (parameter options: long SOA and equal loudness, short SOA and equal loudness, short SOA and equal amplitude) x 2 (task-relevant modality: auditory vs. visual) and 2 (mapping: 1 vs. 2) repeated-measures ANOVA of % discrimination accuracy revealed significant main effects of parameter options and mapping. Furthermore, it revealed a significant interaction between task and mapping (Table 2.9). In a follow-up 2 (task-relevant modality: auditory vs. visual) x 2 (mapping: 1 vs. 2) repeated measures ANOVA revealed a significant main effect of mapping indicating that participants responded more accurately to stimuli from mapping 2 than to stimuli from mapping 1 (Table 2.11).

Table 2.9 Statistical results of Experiment 5.

	Experiment 5	
	reaction time	accuracy
parameter options	F(1.29,19.33) = 70.61 p < .001*	F(1,15) = 21.43 p <.001*
task-relevant modality	F(1,15) = 912.66 p <.001*	F(1,15) = 2.81 p = .115
mapping	F(1,15) = 4.45 p = .052	F(1,15) = 5.57 p = .032*
parameter options x task-relevant modality	F(1.61,24.17) = 3.20 p = .068	F(1,15) = 2.81 p = .115
parameter options x mapping	F(1.70,25.50) = 0.56 p = .563	F(1,15) = 5.57 p = .032*
task-relevant modality x mapping	F(1,15) = 4.86 p = .044*	F(1,15) = 0.07 p = .790
parameter options x task-relevant modality x mapping	F(1.78,26.67) = 1.88 p = .175	F(1,15) = 0.07 p = .790

*$p < 0.05$

Thus, Experiment 5 provides further evidence that participants in our experiment respond faster to stimuli from mapping 2 than to stimuli from mapping 1. The response profiles in reaction times were reversed neither by the absence of fixation nor by a longer stimulus onset asynchrony nor after equating perceptual loudness of the auditory stimuli.

Table 2.10 Statistical results of follow-up 2-way ANOVAs on reaction time for auditory and visual task of Experiment 5.

	reaction time	
	auditory task	visual task
parameter options	F(1.37,20.47) = 35.6 p < .001*	F(1.39,20.83) = 68.83 p < .001*
mapping	F(1,15) = 7.12 p = .018*	F(1,15) = 0.07 p = .794
parameter options x mapping	F(1.58,23.68) = 2.57 p = .107	F(1.71,25.78) = 0.402 p = .672

*$p < 0.05$

Table 2.11 Statistical results of follow-up 2-way ANOVAs on accuracy per parameter option of Experiment 5.

	accuracy	
	short SOA, equal amplitude	short and long SOA, equal loudness
task-relevant modality	F(1,15) = 2.81 p = .115	No effect: 100% correct
mapping	F(1,15) = 5.57 p = .032*	
task-relevant modality x mapping	F(1,15) = 0.07 p = .790	

*$p < 0.05$

Discussion

This study revisited the metaphoric relationship between auditory pitch and visual size. In previous research participants were faster to discriminate between different sizes in the visual modality, when small-sized stimuli were presented with high-pitched tones and large-sized stimuli with low-pitched tones. Yet, a recent study challenged this generic pitch-size mapping by demonstrating the reverse relationship for dynamic stimuli (Eitan et al., 2014). In the dynamic context, increases in size were associated with rising pitch. To shed further light on this seemingly paradoxical finding, we have investigated several factors that can potentially influence the size-pitch mapping in static contexts.

First of all, we investigated whether stimulus luminance may have contributed to the size-pitch association. In past research luminance and size were correlated, because the overall luminance of the presentation screen will decrease for larger-grating or grey-disc stimuli when presented on a white background (Gallace & Spence; 2006; Evans & Treisman, 2010). To dissociate the effects of luminance and size, Experiments 1 and 2 compared discs that were either brighter or darker than the colour of the background. Moreover, we included circles that limit changes in overall luminance induced by changes in stimulus size. Irrespective of the stimulus (i.e. disc or circle) or the relative luminance between stimulus and background we observed faster reaction times when small size was associated with low pitch and large size with high pitch. A direct comparison between Experiment 1 and 2 raised the possibility that changes in overall luminance play a role in the pitch-size relationship. Nevertheless, changes in luminance did not

revert the profile. Instead, both Experiments 1 and 2 provided convergent evidence for a pitch-size mapping that is opposite to the one previously reported in the literature (Marks et al., 1987; Mondloch & Maurer, 2004; Gallace & Spence, 2006; Parise & Spence, 2009; Evan & Treisman, 2010; Bien et al., 2012; Parise & Spence, 2012; Eitan et al., 2014).

In Experiment 3 and 4 we therefore aimed to identify additional factors that may influence how participants associate pitch and size during speeded reaction time tasks. In particular, we asked in Experiment 3 whether the pitch-size mapping depends on similarity between the two stimulus classes. We expected that the congruency effects would be stronger when the discriminability and similarity between the two stimuli is smaller. Indeed, a concurrent visual stimulus exerted a stronger influence on participants' auditory discrimination when the stimulus classes were closer together. This finding reflects the fact that multisensory influences are most pronounced and relevant when participants' perceptual and decisional uncertainty is high. For instance, participants will be more uncertain on their auditory discrimination judgment, when the two auditory signals are close in frequency space (Grinband, Hirsch & Ferrera, 2006). Critically, however, we still observed a marginally significant effect of mapping. In particular, in the auditory discrimination task participants were slower to respond to stimuli pairing low pitch with large size or high pitch with small size. Most likely, the reaction time effects were less reliably found when the visual modality is relevant, because the overall processing times were shorter. Thereby, the interfering or facilitating auditory stimulus exerted only limited impact on the visual discrimination tasks.

In Experiment 4, we finally manipulated the relative mapping between size and pitch, as human observers may potentially have an absolute AV mapping. In that case, congruency/interference effects may not only depend on the relative size and pitch of the two stimuli that need to be discriminated, but also on the absolute pitch and size pairings. However, for our parameter selection we did not observe any evidence in favour of an absolute pitch-size pairing. Replicating the results of our previous studies, we again found a significant main effect of mapping. Participants were faster to respond to stimuli pairing low pitch/small size or high pitch/large size than the opposite pairing.

In Experiment 5, we investigated the effects of SOA and perceptual loudness and fixation instruction. Even though all these additional three experiments did not instruct participants to fixate the centre of the screen, the three experiments again revealed faster response times for the small/low-pitch and large/high-pitch mapping when the auditory modality was task-relevant. Moreover, we hypothesized that stimuli with short SOA may generate a dynamic setting and thereby influence participants' preferred mapping. Yet, SOA did not influence participants' preferred mapping. Likewise, equating stimuli with respect to their perceptual loudness (Suzuki & Takeshima, 2004) did not affect participant's response time profile.

In summary, all experimental series provided convergent evidence for a pitch-size mapping that pairs low pitch with small size and high pitch with large size. This is a surprising finding as it contradicts previous findings in the literature.

Moreover, it is inconsistent with the natural association between the size of a resonance body and the frequency of the sounds it produces.

We suggest that the key for understanding these seemingly contradictory results lies in how participants interpret stimulus size. Crucially, retinotopic size is determined by two key factors. First, it depends on the constant size of the stimulus in the environment. Second, it depends on the distance of the stimulus from the observer. The stimuli in our study – in particular the circles - were less likely to be associated with different objects, but rather with one object at different distances from the observer. Participants may also have performed the task by comparing the current stimulus implicitly to previous ones and judging whether it was closer or farther away. In this way, our study links the previous findings on the pitch-size mapping under static and dynamic contexts. If size is interpreted as the size of an object or a resonance body, large size is associated with low pitch as previously reported in the literature for static contexts. However, if size is interpreted as distance from the observer as in the current study and previous dynamic contexts (Eitan et al., 2014), large size is associated with high pitch. Future studies are needed to carefully manipulate participants' interpretation of 'size' as object size or distance from observer. For instance, experiments may manipulate instructions, background story or change the stimuli to guide participants' interpretation towards either object size or distance from the observer, or contrast the different interpretations directly in a three-dimensional virtual-reality setting.

Moreover, even though the absolute pitch and size values did not significantly affect participants' response time profile in Experiment 4, this finding may not generalize to the entire range of pitch frequency and size values. For instance, it is conceivable that small circles map to high pitch and large circles to low pitch outside the tested range of values. Anecdotally, some of our participants mentioned that they perceived both sounds as high-pitched in our experiments. In other words, even though participants do not have absolute pitch, they may still be endowed with some coarse pitch classification scheme. If both the pitch-levels chosen violate participants' coarse pitch classification, audiovisual interference experiments may be attenuated or even reverted.

Finally, future research also needs to further investigate the role of sound amplitude and perceived loudness. In our experiments we equated sounds of different frequencies with respect to their physical sound amplitude (Experiment 1-4) or perceived loudness (Experiment 5, run 1-2) based on published equal loudness contours (Suzuki & Takeshima, 2004). However, equal loudness contours may differ between subjects. Hence, future studies are required that carefully equate sound loudness individually for each participant (e.g. using adaptive staircases).

In conclusion, this series of AV interference experiments showed that participants map small size onto low pitch and large size onto high pitch under specific parameter settings when the auditory modality was task relevant. These results suggest that the pitch-size mapping may be less generic and stable than

previously assumed. It may depend on the exact stimulus parameters, task-context and potentially prior experience of the participant.

CHAPTER 3: THE NEURAL BASIS OF THE RELATIONSHIP BETWEEN AUDITORY PITCH AND VISUAL SIZE IN POLAR COORDINATES

The work presented in this chapter is part of a collaboration between Alexandra Krugliak and Uta Noppeney. It is currently being prepared for publication. The experiment was designed by AK and UN, the data was collected and analysed by AK (Supervised by UN), the introduction, methods and results sections of this manuscript were written by AK and UN, the discussion was written by AK.

Introduction

Since Ghazanfar and Schroeder (2006) proposed that the whole neocortex is multisensory, a large body of evidence has accumulated, confirming that crossmodal influences are prominent as early as at the primary cortical level (Driver & Noesselt, 2008; Klemen & Chambers, 2012; Noppeney, Jones, Rohe & Ferrari, 2018; for a focus on anatomic pathways see: Cappe, Rouiller & Barone, 2009).

Signals from non-preferred modalities have been shown to modulate the responses of the preferred modality and to drive responses themselves. For example, retrograde tracing studies in the macaque monkey revealed direct connections from primary auditory cortices to anterior portions of the primary visual cortices (Falchier, Clavagnier, Barone & Kennedy, 2002; Rockland & Ojima, 2003) and from V2 to caudal auditory cortices (Falchier et al., 2010). On the neural level, Allman and colleagues discovered in the extrastriate visual

cortices of cats and ferrets a mix of bimodal neurons, neurons that respond to both auditory and visual stimulation, as well as a large proportion of unimodal neurons that are modulated at the subthreshold level by signals from another sensory modality (Allman & Meredith, 2007; Allman et al., 2008; Allman, 2009).

In human neuroimaging, researchers succeeded in decoding stimulus features presented in one sensory modality from brain areas of non-preferred modalities (Meyer et al., 2010; Man, Kaplan, Damasio & Meyer, 2012; Liang, Mouraux & Iannetti, 2013; Vetter, Smith & Muckli, 2014). For example, muted videos depicting animals, objects or instruments have been successfully decoded from BOLD-response patterns in the auditory cortices (Meyer et al., 2010), similarly also videos of manmade objects were decodable in auditory cortices and posterior superior temporal sulcus but not isolated sounds corresponding to those videos in visual cortices (Man et al., 2012). Furthermore, successful decoding has been demonstrated for natural and imagined sounds in early visual cortices (Vetter et al., 2014), and spatial location of visual stimuli in primary auditory cortices (Liang et al., 2013).

However, these were all environmentally realistic stimuli where auditory and visual signals are naturally linked in our everyday experience. Consequently, it cannot be excluded that successful decoding of amodal stimulus properties is in fact facilitated by top-down effects from higher order association areas that triggered a representation in another sensory modality or via imagery.

This raises the critical question of whether stimulus features that do not share a natural mapping between the senses can be decoded from sensory cortices. For

instance, can we decode size and colour from visual cortices and pitch and timbre from auditory cortices? Importantly, because sensory features do not share correspondences across the senses, potential successful decoding results from BOLD-responses in sensory cortices are unlikely to be attributed to mental imagery.

Mappings between seemingly arbitrary sensory features are mostly known from synaesthesia. However, multisensory metaphoric mappings have also been demonstrated in non-synaesthetic individuals. For example, when presented with a high-pitched tone participants intuitively prefer a visual object at a higher elevation (Bernstein & Edelstein, 1971; Melara & O'Brien, 1987; Ben-Artzi & Marks, 1995; Patching & Quinlan, 2002; Evans & Treisman, 2010). In a similar vein, when two objects of equal size are presented interleaved with a high-intensity sound, participants perceive the object following the sound as larger in size than if a low-intensity sound was presented (Takeshima & Gyoba, 2013). It has been proposed that such mappings have their origin in natural environmental statistics (Spence, 2011). The pitch-size mapping, for instance, has been associated with the fact that the frequency of a sound produced by an animal is related to its vocal tract size, large animals like elephants tend to produce lower pitch sounds than small animals like birds (von Kriegstein, Smith, Patterson, Ives, & Griffiths, 2007). Hence, high-pitched sounds are often associated with small objects and low-pitched sounds with large objects (Marks, Hammeal, Bornstein & Smith, 1987; Mondloch & Maurer, 2004; Gallace & Spence, 2006; Parise & Spence, 2009; Evans & Treisman, 2010; Bien, ten Oever, Goebel & Sack, 2012;

Parise & Spence, 2012; Eitan, Schupak, Gotler & Marks, 2014; Tonelli, Cuturi & Gori, 2017; Brunetti, Indraccolo, Del Gatto, Spence & Santangelo, 2018; Ueda, Mizuguchi, Yakushijin & Ishiguchi, 2018). However, the reverse relationship has also been demonstrated. In a dynamic context, a rising pitch has been associated with an approaching object (Eitan et al., 2014). Similarly, participants responded faster when small-sized circles that were centred at fixation were accompanied by a low-pitched tone compared to when it was accompanied by a high-pitched tone (Krugliak & Noppeney, 2016).

In this study, we investigated auditory pitch and visual size as circles centred around fixation covering different eccentricities. Importantly, both stimulus dimensions are represented topographically in the neocortex. Frequency defines the main gradients of tonotopic maps in auditory cortices (Formisano et al., 2003; Da Costa et al., 2011). Eccentricity defines, together with spatial angle, retinotopic maps in visual cortices (Wandell, Dumoulin & Brewer, 2007). Therefore, these stimulus features are good candidates for exploring auditory and visual interactions in auditory and visual cortices. Furthermore, due to their gradual topographic representations, we can also explore if the dimensions of these two stimulus features are related, and if so how: Do the maps align in a particular fashion e.g. are small circles associated with high-frequency tones or with low-frequency tones? Is there an absolute magnitudinal relationship between the two e.g. are the stimuli related very roughly, in a one-to-one mapping, or are there certain bands of stimuli that are related in a particular fashion?

Here, we presented participants with pure tones of different frequencies and circles centred around fixation, covering different eccentricities. We used support vector regression to decode both stimulus dimensions from regions of interests in the auditory cortices and along the ventral and dorsal visual pathways. Critically, if synaesthetic mapping between neural representations is established already at the primary cortical level, the mapping from activation pattern to pitch should be related to the mapping from activation pattern to size. In other words, the mapping from activation pattern to pitch in audition should generalize to the mapping from activation pattern to size in vision.

Methods

Participants

11 participants (5 female, mean age: 23 years) took part in this experiment after giving written informed consent. Each had normal or corrected-to-normal vision, reported normal hearing, and had no history of neurological or psychiatric illness. Participants received a monetary reward. The study was approved by the human research ethics committee at the University of Birmingham.

Stimuli

Visual stimuli were circles (line thickness: 0.7° visual angle) of 15 sizes (i.e. radius sampled linearly in steps of 0.7° from 0.7° to 10.5° visual angle, 500 ms duration). Irrespective of size the circle's centre was fixed to the centre of the screen where

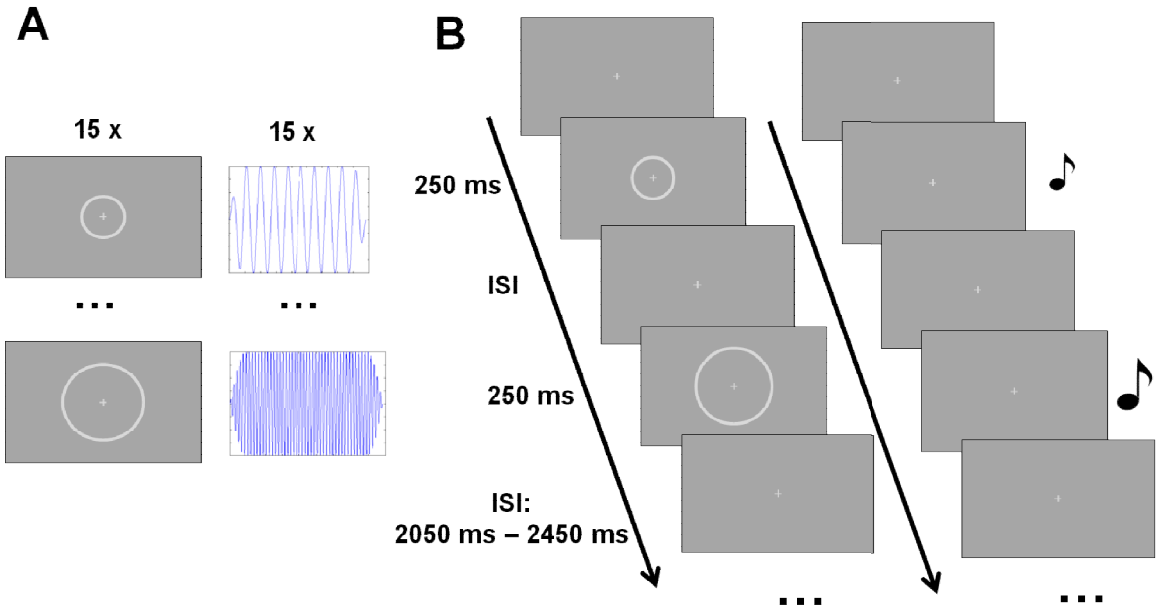


Figure 3.1 Experimental design and example trial. A. Left. Visual stimuli: 15 circles with radii ranging from 0.7° to 10.5°, sampled linearly in steps of 0.7°. Right. Auditory stimuli: 15 pure tones with frequencies ranging from 500 Hz to 4000 Hz, sampled linearly in steps of 250 Hz. B. Left. Visual run. Visual circles of variable sizes were presented for 250 ms separated by an inter-stimulus-interval (ISI) that was jittered between 2050 ms and 2450 ms. Right. Auditory run. Auditory pure tones of variable frequencies were presented for 250 ms separated by an inter-stimulus-interval (ISI) that was jittered between 2050 ms and 2450 ms.

the fixation cross was presented. To enable a gap detection task (see below) half of the circles were presented with a gap (i.e. a small disk with radius = 0.35° coloured as the background) that was located with equal probability anywhere on the circle.

Auditory stimuli were 15 pure tones of variable pitch (500 ms duration with linear onset and offset ramps of 10 ms to avoid auditory clicks; sampling rate 44100 Hz). The frequency of the tone was sampled linearly in steps of 250 Hz from 500 Hz and 4000 Hz. To enable a gap detection task (see below) half of the pure tones were interrupted by 100 ms silence that was inserted randomly between 20 ms and 380 ms after sound onset. The sound pressure level was adjusted to the loudest comfortable level for each participant at the beginning of the scanning session.

Main experimental design

In a 15 (stimulus size respectively pitch: 1-15) x 2 (modality: A or V) factorial design observers were presented with either unisensory visual circles of variable sizes or auditory pure tones of variable pitch in separate runs (Figure 3.1). Half of the circles and tones were interrupted by a spatial respectively temporal gap. On each trial, observers indicated whether the stimulus included a gap via a two choice key press as fast and accurately as possible. Throughout the entire experiment observers were instructed to fixate a central cross.

The trial onset asynchrony was jittered between 2300 ms and 2700 ms. The order of stimuli were pseudo-randomized within each run. Further, 6 % of the events were ‘null events’ (i.e. fixation with no stimulus presentation).

The fMRI data were acquired on two separate days. On each day 8 auditory and 8 visual runs were acquired. The order of auditory and visual runs was counterbalanced within and across subjects.

Experimental setup

Visual and auditory stimuli were presented using Cogent 2000 v1.25 (developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON at the Wellcome Department of Imaging Neuroscience, UCL, London, UK; http://www.vislab.ucl.ac.uk/cogent.php) running on Matlab R2012a (MathWorks Inc.) on a Windows PC. The visual stimuli were back-projected onto a Plexiglass screen at the end of the scanner bore using a D-ILA projector (JVC DLA-SX21). The screen was visible to the participant via a mirror that was mounted on the MR head coil. The auditory stimuli were delivered via AVOTEC SS-3100 headphones (Avotec Inc.) at a maximum comfortable sound level which was established individually for each participant.

MRI data acquisition

The scanning sessions were conducted in a 3 Tesla Philips Achieva scanner with a 32-channel head coil at Birmingham University Imaging Centre. On the first day, T1-weighted anatomical images (TR = 8.4 s, TE = 3.8 ms, TI = 540 ms, 175 slices, image matrix = 288 x 232, spatial resolution: 1 x 1 x 1 mm^3) were acquired, and in two subsequent sessions, T2*-weighted echo-planar images (EPI) (TR = 2.6 s, TE = 0.4 ms, 38 axial slices acquired in ascending order without gaps covering the whole brain, image matrix = 80 x 80, spatial resolution: 3 x 3 x 3 mm^3). The first 4 scans of each run were acquired to allow for T1 saturation effects and discarded immediately. They were followed by 230 volumes. Each EPI run had a duration of 10.14 min.

fMRI analysis:

Pre-processing

The EPI images were pre-processed with Statistical Parametric Mapping (SPM8; Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm/; Friston, Holmes, Worsley, Frith & Frackowiak, 1995) running on Matlab R2012a. Scans from each participant were realigned to the first scan as reference and residual motion-related deformations were corrected using an unwarping-function. The time-series in each voxel were high-pass filtered to 1/128 Hz. The EPI images were analysed in participant's native space. The high-resolution T1-weighted anatomical image was coregistered to the mean EPI image.

ROI definition

Auditory regions of interest (ROI) were defined based on the Brainnetome atlas (Fan et al., 2016) (http://atlas.brainnetome.org/). We defined two ROIs: (i) primary auditory cortex (PAC) including bilateral areas TE1.0 and TE1.2 on the Heschl's gyrus and (ii) STG including superior temporal gyrus (STG), bilateral area 41-42 (~ Planum Temporale), and the caudal and rostral area 22. Visual ROIs were defined based on probabilistic retinotopic maps (Wang, Mruczek, Arcaro & Kastner, 2015; http://scholar.princeton.edu/napl/resources0, using a 80% overlap threshold). We defined four visual ROIs: (i) prob-V1, (ii) combined prob-V2-V3, (ii) a ventral ROI combining hV4, VO1 and VO2, and (iv) a dorsal ROI consisting of V3A, V3B, IPS0, IPS1 and IPS2. The masks were first inverse-normalized from MNI standard

space (Ashburner & Friston, 2005) into native space for each individual participant. Then the masks were resampled to 2 x 2 x 2 mm^3 voxels.

fMRI analysis

The data was modelled in an event-related fashion including one regressor for each of the 15 auditory and visual stimuli. The regressors were entered into a design matrix after convolving each event-related unit impulse (representing a single trial) with a canonical hemodynamic response function and its first temporal derivative. The realignment parameters were included as nuisance parameters in order to account for residual motion artefacts.

Support Vector Regression

We trained linear Support Vector Regression models (libSVM 3.20; Chang & Lin, 2011) as implemented in The Decoding Toolbox (Hebart, Görgen, Haynes & Dubois, 2015) to predict the stimulus labels within each of the six ROIs. First, we extracted response patterns for each voxel within the ROI from the parameter estimate image corresponding to the magnitude of the BOLD-response for each run and condition. The resulting parameter estimate images were then masked with the corresponding binary ROI mask and pre-whitened runwise (Walther et al. 2012). The parameter estimate images for training and test data were normalized independently using euclidean normalization (Schrouff et al., 2013). Before training the SVR models, we standardized the stimulus parameters: first labels were sorted (pure tone frequency from low to high, visual circle radius from small

to large) and then z-normalized. In a leave-one-run-out cross-validation procedure, the support vector regression models were trained to learn the mapping from condition-specific fMRI BOLD-response patterns to the 15 pure tones for the auditory runs or the 15 circles for the visual runs from all but one run. The SVR's parameters C and nu were standard fixed parameters (C = 1, nu = 0.5). The model then used this learned mapping to decode the stimulus codes from the voxel response patterns of the remaining (left-out) run. In a leave-one-run-out cross-validation scheme, the training-test procedure was repeated for all runs. For cross-modality decoding cross-validation was not necessary, all auditory runs were assigned to the training set and all visual runs to the test set (AV), and vice versa (VA).

Statistical Inference

To perform within-subject statistics we used the decoded labels as predictors for the true stimulus labels in general linear regression models, computed separately for each auditory and each visual run. The run-specific parameter estimates were then entered into one sample t-tests individually for each participant.

To enable generalization to the population level we used the decoded labels (averaged across cross-validation folds) as predictors for the true stimulus labels in a general linear regression model for auditory and visual runs within each participant. The participant-specific parameter estimates were then entered into one sample t-tests at the group level.

Results

Behavioural Results - Gap Detection Task

We analyzed the responses of the gap detection task as percentage correct responses for individual runs. All participants performed consistently above 75% correct. From participant S03 we are missing behavioural data from the eighth auditory run because of technical issues with the response box. However, the experimenter was monitoring the behavioural responses during scanning and did not notice any changes in responding. Therefore, we conclude that our participants were awake and attended the stimuli throughout the whole experiment.

A 2 x (modality: auditory vs. visual) one-sample t-test on average percentage correct per run confirmed that there was no significant difference in gap-detection accuracy between the auditory and visual runs ($t(10) = -0.30$, $p = .773$).

Multivariate analyses

Group-level statistics: At the group level, we were unsurprisingly able to decode circle size from BOLD-response pattern across all visual regions and pure tone frequency from BOLD-response pattern from auditory regions (Table 3.1). Critically, however, we were also able to decode circle size from BOLD-response pattern significantly from PAC and with a marginally significance from STG (Table 3.1). We also performed cross-modality decoding by training on BOLD-response patterns of circle size and training on BOLD-response patterns of pure tone frequency and vice versa, but we did not find any significant results (Table 3.2).

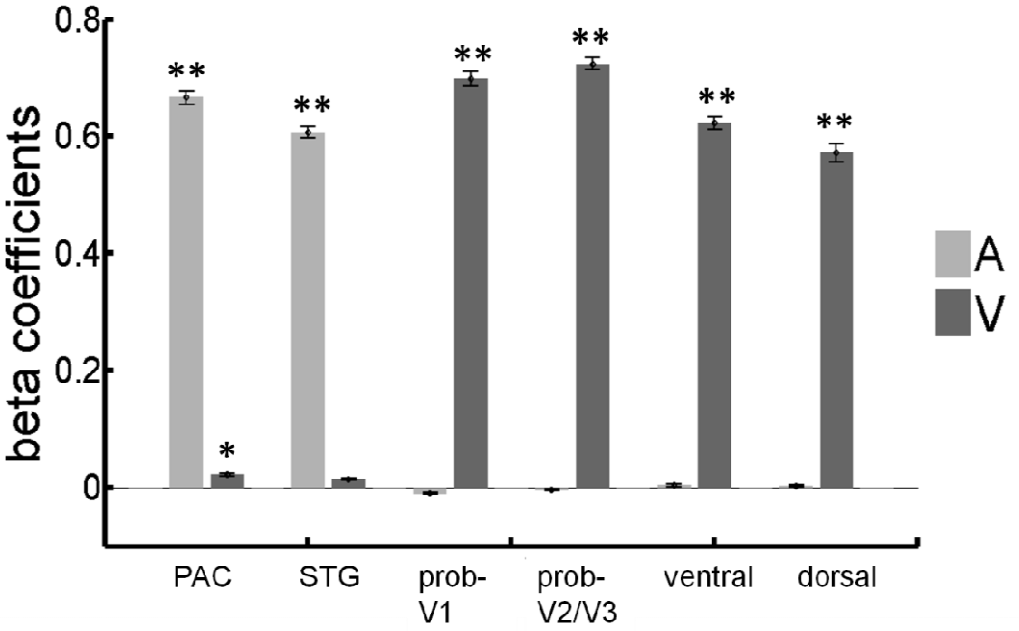


Figure 3.2 Support vector regression results across participants. The beta coefficients were significant for auditory stimuli in both auditory ROIs: PAC and STG. Beta coefficients for visual stimuli were significant in all of the visual ROIs: prob-V1, prob-V2/V3, ventral and dorsal, and also in PAC. In STG there was a trend.

Single-subject statistics: At the single-subject level, we were able to decode circle size from BOLD-response patterns in all visual regions (Tables 3.5, 3.6, 3.7 and 3.8) and pure tone frequency from BOLD-response patterns in auditory regions significantly in each participant (Tables 3.3 and 3.4). However, we were able to decode circle size significantly better than chance from BOLD-response patterns in PAC in 4 participants (Table 3.3) and from STG in 4 participants (2 same as PAC, 2 different) (Table 3.4). Furthermore, we were able to decode pure

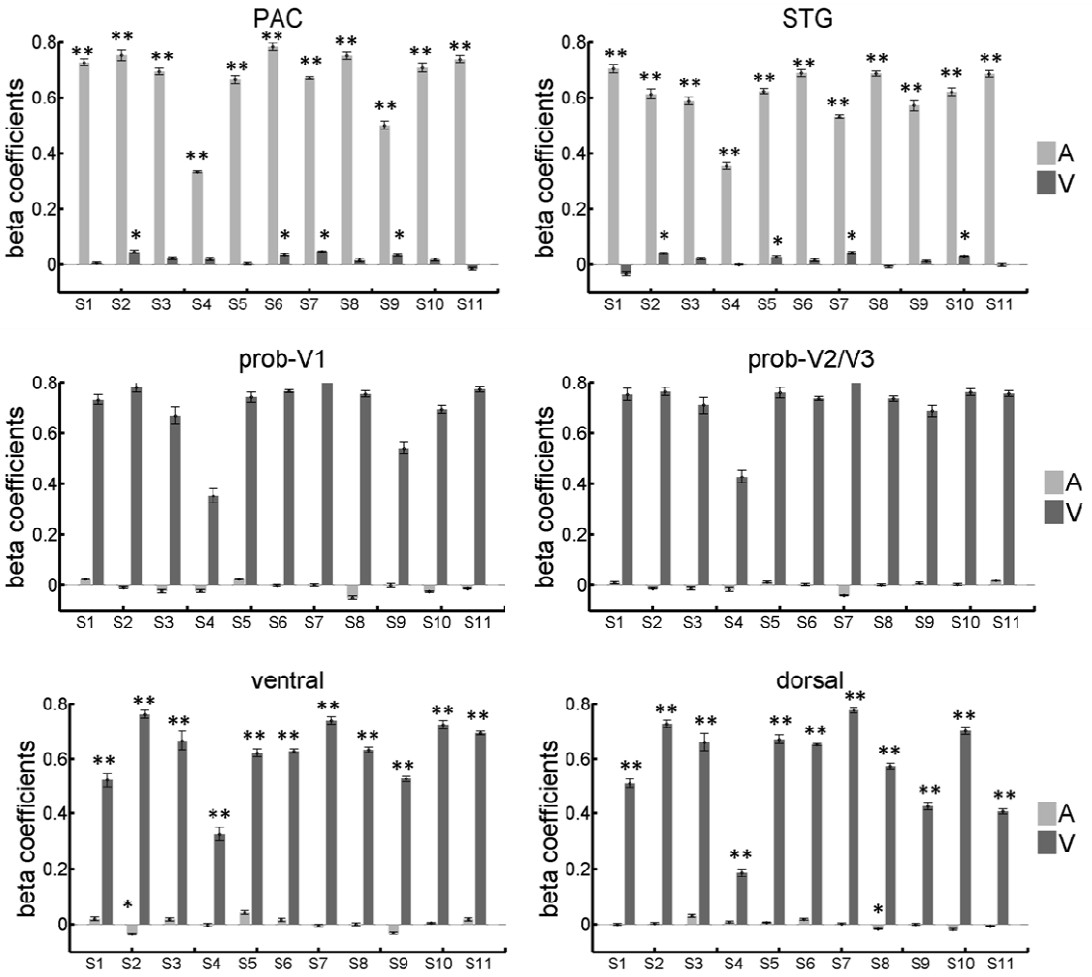


Figure 3.3 Support vector regression results per ROI for individual participants. In all participants the beta coefficients were significant for auditory stimuli in PAC and STG. Beta coefficients for visual stimuli were significant, in all participants, in prob-V1, prob-V2/V3, the dorsal ROI, and the ventral ROI. Additionally, decoding significantly better than chance was observed in PAC in 4 participants, in STG in 4 participants, in prob-V1 in 4 participants, and in one single participant in prob-V2/V3, dorsal and ventral ROIs, respectively.

tone frequency better than chance from BOLD-response patterns in prob-V1 in 4 participants (Table 3.5), and in prob-V2/V3 (Table 3.6), ventral (Table 3.7) and dorsal (Table 3.8) in only one participant respectively.

Table 3.1 Group results of within-modality SVR analysis. Mean beta values for auditory and visual stimuli per ROI.

	auditory stimuli		visual stimuli	
ROI	mean (std)	p-value	mean (std)	p-value
PAC	0.67 (0.13)	p < .001**	0.02 (0.02)	p = .004*
STG	0.61 (0.10)	p < .001**	0.01 (0.02)	p = .064
prob-V1	-0.01 (0.02)	p = .175	0.70 (0.14)	p < .001**
prob-V2/V3	0.00 (0.02)	p = .487	0.72 (0.11)	p < .001**
ventral	0.00 (0.02)	p = .570	0.62 (0.13)	p < .001**
dorsal	0.00 (0.01)	p = .519	0.57 (0.18)	p < .001**

*$p < 0.05$

Table 3.2 Group results of cross-modality SVR analysis. Mean beta values for auditory and visual stimuli per ROI.

	train on visual stimuli decode auditory stimuli		train on auditory stimuli decode visual stimuli	
ROI	mean (std)	p-value	mean (std)	p-value
PAC	-0.01 (0.06)	p = .471	-0.01 (0.03)	p = .314
STG	-0.01 (0.03)	p = .164	0.00 (0.02)	p = .884
prob-V1	0.00 (0.02)	p = .973	-0.01 (0.07)	p = .571
prob-V2/V3	0.00 (0.03)	p = .705	0.00 (0.05)	p = .948
ventral	0.01 (0.02)	p = .237	0.01 (0.05)	p = .485
dorsal	0.00 (0.02)	p = .767	0.00 (0.02)	p = .847

*$p < 0.05$

Table 3.3 Single-subject results of within-modality SVR analysis in PAC. Mean beta values for auditory and visual stimuli per participant.

	auditory stimuli		visual stimuli	
participant	mean (std)	p-value	mean (std)	p-value
S1	0.73 (0.10)	p < .001**	0.01 (0.02)	p = .490
S2	0.75 (0.16)	p < .001**	0.05 (0.05)	p = .031*
S3	0.69 (0.10)	p < .001**	0.02 (0.04)	p = .200
S4	0.33 (0.03)	p < .001**	0.02 (0.04)	p = .245
S5	0.66 (0.12)	p < .001**	0.00 (0.03)	p = .772
S6	0.78 (0.11)	p < .001**	0.04 (0.03)	p = .022*
S7	0.67 (0.04)	p < .001**	0.05 (0.03)	p = .001*
S8	0.75 (0.11)	p < .001**	0.01 (0.05)	p = .457
S9	0.50 (0.12)	p < .001**	0.03 (0.03)	p = .017*
S10	0.71 (0.12)	p < .001**	0.02 (0.04)	p = .241
S11	0.74 (0.10)	p < .001**	-0.01 (0.04)	p = .327

*$p < 0.05$

Table 3.4 Single-subject results of within-modality SVR analysis in STG. Mean beta values for auditory and visual stimuli per participant.

	auditory stimuli		visual stimuli	
participant	mean (std)	p-value	mean (std)	p-value
S1	0.70 (0.12)	p < .001**	-0.03 (0.05)	p = .135
S2	0.61 (0.14)	p < .001**	0.04 (0.02)	p = .001*
S3	0.59 (0.12)	p < .001**	0.02 (0.04)	p = .131
S4	0.35 (0.09)	p < .001**	0.00 (0.02)	p = .960
S5	0.62 (0.07)	p < .001**	0.03 (0.03)	p = .035*
S6	0.69 (0.09)	p < .001**	0.02 (0.04)	p = .276
S7	0.53 (0.04)	p < .001**	0.04 (0.04)	p = .016*
S8	0.69 (0.08)	p < .001**	-0.01 (0.03)	p = .569
S9	0.57 (0.15)	p < .001**	0.01 (0.02)	p = .145
S10	0.62 (0.11)	p < .001**	0.03 (0.01)	p < .001**
S11	0.69 (0.10)	p < .001**	0.00 (0.04)	p = .936

*$p < 0.05$

Table 3.5 Single-subject results of within-modality SVR analysis in prob-V1. Mean beta values for auditory and visual stimuli per participant.

	auditory stimuli		visual stimuli	
participant	mean (std)	p-value	mean (std)	p-value
S1	0.02 (0.03)	p = .037*	0.73 (0.16)	p < .001**
S2	-0.01 (0.03)	p = .431	0.78 (0.16)	p < .001**
S3	-0.02 (0.05)	p = .204	0.67 (0.27)	p < .001**
S4	-0.02 (0.04)	p = .146	0.35 (0.22)	p = .003*
S5	0.02 (0.02)	p = .014*	0.74 (0.15)	p < .001**
S6	0.00 (0.03)	p = .815	0.77 (0.04)	p < .001**
S7	0.00 (0.04)	p = .945	0.86 (0.10)	p < .001**
S8	-0.05 (0.05)	p = .034*	0.76 (0.10)	p < .001**
S9	0.00 (0.06)	p = .927	0.54 (0.20)	p < .001**
S10	-0.03 (0.02)	p = .011*	0.69 (0.12)	p < .001**
S11	-0.01 (0.02)	p = .052	0.77 (0.09)	p < .001**

*$p < 0.05$

Table 3.6 Single-subject results of within-modality SVR analysis in prob-V2/V3. Mean beta values for auditory and visual stimuli per participant.

	auditory stimuli		visual stimuli	
participant	mean (std)	p-value	mean (std)	p-value
S1	0.01 (0.04)	p = .543	0.76 (0.18)	p < .001**
S2	-0.01 (0.02)	p = .188	0.77 (0.12)	p < .001**
S3	-0.01 (0.04)	p = .413	0.71 (0.26)	p < .001**
S4	-0.02 (0.05)	p = .392	0.43 (0.21)	p < .001**
S5	0.01 (0.03)	p = .349	0.76 (0.16)	p < .001**
S6	0.00 (0.04)	p = .966	0.74 (0.06)	p < .001**
S7	-0.04 (0.02)	p < .001**	0.85 (0.10)	p < .001**
S8	0.00 (0.03)	p = .979	0.74 (0.08)	p < .001**
S9	0.01 (0.04)	p = .498	0.69 (0.17)	p < .001**
S10	0.00 (0.03)	p = .891	0.76 (0.10)	p < .001**
S11	0.02 (0.02)	p = .075	0.76 (0.09)	p < .001**

*$p < 0.05$

Table 3.7 Single-subject results of within-modality SVR analysis in dorsal ROI. Mean beta values for auditory and visual stimuli per participant.

	auditory stimuli		visual stimuli	
participant	mean (std)	p-value	mean (std)	p-value
S1	0.02 (0.05)	p = .299	0.52 (0.21)	p < .001**
S2	-0.04 (0,03)	p = .006*	0.76 (0.11)	p < .001**
S3	0.02 (0.05)	p = .298	0.67 (0.27)	p < .001**
S4	0.00 (0.05)	p = .925	0.33 (0.20)	p = .003*
S5	0.04 (0.05)	p = .064	0.62 (0.11)	p < .001**
S6	0.02 (0.05)	p = .340	0.63 (0.07)	p < .001**
S7	0.00 (0.04)	p = .733	0.74 (0.11)	p < .001**
S8	0.00 (0.04)	p = .977	0.63 (0.08)	p < .001**
S9	-0.03 (0.04)	p = .062	0.53 (0.09)	p < .001**
S10	0.00 (0.02)	p = .460	0.72 (0.12)	p < .001**
S11	0.02 (0.05)	p = .298	0.69 (0.05)	p < .001**

*$p < 0.05$

Table 3.8 Single-subject results of within-modality SVR analysis in ventral ROI. Mean beta values for auditory and visual stimuli per participant.

	auditory stimuli		visual stimuli	
participant	mean (std)	p-value	mean (std)	p-value
S1	0.00 (0.02)	p = .959	0.51 (0.13)	p < .001**
S2	0.00 (0.02)	p = .694	0.73 (0.11)	p < .001**
S3	0.03 (0.04)	p = .073	0.66 (0.25)	p < .001**
S4	0.01 (0.03)	p = .432	0.19 (0.11)	p = .002*
S5	0.01 (0.02)	p = .261	0.67 (0.13)	p < .001**
S6	0.02 (0.03)	p = .087	0.65 (0.03)	p < .001**
S7	0.00 (0.02)	p = .644	0.78 (0.08)	p < .001**
S8	-0.02 (0.01)	p = .006*	0.57 (0.08)	p < .001**
S9	0.00 (0.02)	p = .955	0.43 (0.11)	p < .001**
S10	-0.02 (0.03)	p = .118	0.70 (0.01)	p < .001**
S11	-0.01 (0.20)	p = .364	0.41 (0.07)	p < .001**

*$p < 0.05$

Discussion

In this study, we used support vector regression (SVR) to explore the relationship between auditory pitch and visual eccentricity (as circles centred around fixation). These stimulus features are known to have robust topographic representations in

the neocortex: frequency defines topographic maps in the auditory cortices and eccentricity defines, together with spatial angle, retinotopic maps in the visual cortices. Accordingly, we found robust decoding of auditory stimuli in the auditory ROIs and of visual stimuli in the visual ROIs. Critically, we also showed significant decoding of visual stimuli from PAC and a trend in STG.

The highly significant results for decoding frequency from the auditory cortices and eccentricity from visual ROIs were expected given that analogous stimuli are used to map the local topographies. The high reliability of these results confirms that our stimuli elicited response patterns sufficiently different from each other to allow optimal decoding.

Significant decoding of visual stimuli from auditory cortices but not of auditory stimuli from visual cortices is consistent with findings of previous studies. For example, Meyer and colleagues (2010) were able to classify muted videos from auditory cortices but not the corresponding sounds from visual cortices. Likewise, Man and colleagues (2012) were only able to classify muted videos of manmade objects and actions from the primary auditory cortices and an area in posterior superior temporal sulcus but not the corresponding auditory sounds from the primary visual cortex. Here, we successfully demonstrated that also a range of very similar modality-specific visual eccentricity stimuli can be decoded from auditory cortices.

Decoding of auditory stimuli from BOLD-response patterns in visual cortices appears to be more challenging than the other way round. To our knowledge, there is only one study that reported successful classification of auditory stimuli

and imagined auditory stimuli from activation patterns in visual cortices (Vetter et al., 2014). When blindfolded participants were presented with natural sounds, the identity and category of those sounds could be reliably decoded from both auditory and visual cortices. Yet, when participants were asked to imagine those sounds, decoding was only successful in the early visual cortices, mostly in areas representing the peripheral part of the visual field. Like in previous studies, Vetter and colleagues (2014) presented natural sounds of everyday situations. The most striking difference with other studies was that participants were blindfolded. Is there such a dominance of the visual modality that only under circumstances of temporary or complete sensory deprivation auditory influences can be decoded from visual cortices? This is not entirely true. The experiment of Vetter and colleagues (2014) was recently replicated (Petro, Paton & Muckli, 2017). Instead of blindfolding, participants were presented with a blank screen and instructed to keep their eyes open. The general findings were replicated but classification accuracies were lower than with blindfolding. The authors agreed that visual stimulation might impair the decoding of auditory influences on BOLD-response patterns in visual cortices and speculated that more demanding visual stimulation might abolish decodability entirely. This factor might explain why decoding of auditory information from visual cortices in sighted participants remains a challenge, yet proves successful under conditions of sensory deprivation. For example, Watkins and colleagues (2013) reported that auditory stimuli activated visual cortices in congenitally blind participants, in some participants activations in area V5 even resembled a tonotopic map. Furthermore, electrophysiology in visual

cortices of two epilepsy patients revealed topographically specific responses to peripheral spatial sounds (Brang et al., 2015). It remains for future studies to optimize and investigate decoding of auditory stimuli in visual areas.

Cross-modality decoding of BOLD-response pattern elicited by visual stimuli after the support vector regression model was trained on BOLD-response patterns of auditory stimuli, and vice versa was not successful in any of the ROIs. Overall, successful cross-modality decoding remains a challenge. For example, Meyer and colleagues (2010) reported a trend for classification of the categories animals and objects but the results for individual exemplars within those categories were not even near statistical significance. Cross-modality classification in the auditory cortices has only been shown to be successful after participants were familiarized with videos and subsequently presented with isolated auditory and visual components of those videos during scanning, additionally, participants were also instructed to imagine the missing sensory component as vividly as possible (Man et al., 2012). At the methodological level, one of the difficulties working on cross-modality decoding is that one dataset is often inherently noisier than the other (e.g. preferred modality vs. non-preferred modality). This aspect is potentially mediated my multisensory integration mechanisms. In an interesting study, de Haas and colleagues (2013) revealed how congruency of audiovisual stimuli affects not the similarity between BOLD-response patterns but their reliability. Response patterns elicited by incongruent stimuli were less reliable than those elicited by congruent stimuli. This is consistent with a view proposing that the main purpose of cross-modal interactions at early and primary cortical levels is to manipulate the saliency

of multisensory stimuli (Noppeney et al., 2018). At this stage, it is speculative to claim that the effect of congruency on the reliability of BOLD-response patterns is a general mechanism, taking into account that the effects were evident only in V2 and V3 but not in V1 or the auditory cortices. Still, it is an idea worth following-up in future studies.

The results of the current study help to shed lights on the findings by Stokes and colleagues (2009). They studied the neural representations of perceived and imagined visual letters 'X' and 'O'. In the imagery condition two pure tones served as cues for imagining the letters. Prior to scanning the letter 'X' was associated with a 600 Hz tone and the letter 'O' with a 200 Hz tone. Both perceptual and imagined letters could be decoded from visual cortices but decoding in auditory cortices was significant only in the imagery condition in which the tones were actually presented. The most likely explanation is that decoding in auditory cortices was driven mainly by the presentation of the tones. However, the design of the experiment does not allow excluding a visual contribution because: (i) auditory tones were linked with visual letters before scanning, (ii) perceptual and imagery blocks were presented in interleaved order, and (iii) the block design did not allow to spatially and/or temporally disentangling responses to the tones and the imagined letters. In fact, all of the decoding results in this study could have been driven not by perceived/imagined letters and pure tones alone but by their serial association. Our study provides supporting evidence that it is indeed possible to decode pure tones of a similar frequency range as used in the study of Stokes and colleagues (2009) from auditory cortices.

The main motivation for choosing a wide range of highly similar stimuli was to reveal the existence and nature of a neural mapping between auditory pitch and visual size in polar coordinates. Our data do not allow us to draw any conclusions regarding these questions so far. The BOLD-response patterns appear to be too noisy to allow cross-modality decoding or to reveal a mapping. Importantly, we have demonstrated that despite weak activations by our stimuli, we were still able to reliably decode the visual stimuli from auditory cortices. In the future, it would be interesting to investigate if fewer but highly salient visual stimuli can reveal more information about the representation of visual eccentricity in the auditory cortices of sighted participants.

In summary, we demonstrated robust decoding of auditory pitch from auditory cortices and of visual size in polar coordinates from early visual ROIs as well of both ventral and dorsal visual areas. Crucially, we are the first study to have successfully decoded a modality-specific stimulus, visual size in polar coordinates, from auditory cortices of sighted participants. Decoding of auditory pitch from visual areas and cross-modality decoding, however, have not yielded significant results. Our results were not able to reveal the nature of a potentially existing mapping between auditory pitch and visual size in polar coordinates.

CHAPTER 4: SEEING PITCH AND HEARING SIZE

Introduction

In the previous two chapters, I described how we used the speeded classification task (Krugliak & Noppeney, 2016) and functional Magnetic Resonance Imaging (fMRI) (Krugliak & Noppeney, in preparation) to investigate the relationship between the stimulus features auditory pitch and visual size in polar coordinates. In the speeded classification task participants classified auditory pitch faster when high-pitched tones were accompanied by large size and low-pitched tones with small size. Using fMRI, we demonstrated that unimodally presented visual size in polar coordinates can not only be decoded from BOLD-response patterns of traditional visual areas but also from BOLD-response patterns elicited in auditory areas. The opposite was not true for unimodally presented tones of varied pitch; we were only able to decode the pitch of the presented tones from auditory regions but not from visual regions.

Taken together, these results confirm an existing pitch-size relationship but one that was not as strong as expected. Overall, the findings in the speeded classification task were not as consistent and the decoding results not as strong as compared to similar metaphoric mappings like pitch-elevation (Bernstein & Edelstein, 1971; Melara & O'Brien, 1987; Ben-Artzi & Marks, 1995; Patching & Quinlan, 2002; Evans & Treisman, 2010), or audiovisual stimuli that contain spatial, temporal or semantic congruency cues to guide multisensory integration (Wallace, Wilkinson, & Stein, 1996; Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004; van Atteveldt, Formisano, Goebel, & Bloemert, 2004; Wallace et al., 2004;

Macaluso & Driver, 2005; van Atteveldt, Formisano, Bloemert & Goebel, 2007; Adam & Noppeney, 2010; Lewis & Noppeney, 2010; Vroomen & Keetels, 2010; Donohue, Roberts, Grent-'t-Jong, & Woldorff, 2011; Lee & Noppeney, 2011, 2014).

One theory states that the origin of metaphoric mappings lies in natural environmental statistics (Spence, 2011). Stimulus features that naturally co-occur together in a specific fashion become bound more easily than a different combination of the same features (Marks, 2004; Spence, 2011; Deroy & Spence, 2013). For example, if you walk in a forest and you hear an animal calling, you will likely rely on your experience of previous walks in the forest. Therefore, if the sound has a relatively high pitch you will be more likely to look for the source of that call in the trees, suspecting a bird, but if that call has a lower pitch you will be more likely to look around you at the ground level, looking for a deer. Consequently, observers are more likely to associate high pitch with high elevation. If indeed metaphoric mappings are formed through statistical co-occurrence, and the pitch-size mapping appears not to underlie a robust relationship already, can we artificially make the pitch-size mapping robust?

In a next step we would like to take advantage of this existing but relatively weak relationship between pitch and size in order to manipulate it using audiovisual perceptual learning and answer the following questions: Can we robustly map small size on high pitch, large size on low pitch and vice versa? Can we even induce a one-to-one mapping between size and frequency?

Previous studies have demonstrated that it is possible to link previously unrelated stimulus features from different modalities. Successful attempts include: (i) associating the stiffness of an object with its brightness, and (ii) affecting the perception of static random dots or dots moving in an ambiguous direction after associating the direction of the movement with an auditory cue (Ernst, 2007; Michel & Jacobs, 2007; Teramoto, Hidaka & Sugita, 2010). Furthermore, even very short passive exposure to newly created multisensory stimuli appears sufficient to induce noticeable learning effects in the neural substrates (Tanabe, Honda & Sudato, 2005; Baier, Kleinschmidt & Müller, 2006; Zangenehpour & Zatorre, 2010; Karunanayaka et al., 2015). Usually, presenting a unimodal stimulus leads to increased activation in the corresponding sensory cortices and to deactivations in sensory cortices that prefer another sensory modality (Laurienti et al., 2002; Leitao, Thielscher, Werner, Pohmann & Noppeney, 2012; Iurilli et al., 2012; Ibrahim et al., 2016). However, after stimuli from different sensory modalities were linked in a multisensory context, either by passive exposure or by active learning, and were presented unimodally again, they evoked increased activation in both sensory cortices, those of the preferred modality and those that prefer the modality with which it was linked (PET: McIntosh, Cabeza & Lobaugh, 1998; Zangenehpour & Zatorre, 2010; fMRI: Tanabe et al., 2005; Baier et al., 2006; Meyer, Baumann, Marchina & Jancke, 2007; Martuzzi et al., 2007).

Previous studies have focused on audiovisual association learning. However, in this study, we are interested in going beyond simple association learning by linking not only a few exemplars of pitch and size but instead induce a profound mapping

that will enable participants to generalize to any stimulus parameter combination that lies along the learned mapping. Therefore, we presented participants with a large variety of stimulus exemplars, too many stimuli to remember individually, in order to encourage a general strategy over stimulus-specific learning strategies (Hussain, Bennett & Sekuler, 2012; Arnold & Auvray, 2017). Furthermore, in order to make learning faster and the task more engaging we introduced active learning which has been shown to induce stronger learning and neural coupling effects than passive learning (in multisensory learning: Butler & James, 2012).

In this between-subject experiment, participants learned one of two anti-correlated mappings that linked 51 auditory tone frequencies and circle size parameters one-on-one in a linear fashion. Mapping 1 related high pitch with small size and low pitch with large size. Mapping 2 on the other hand related high pitch with large size and low pitch with small size. After familiarizing participants with the to be learned mapping in an exploration phase, they were presented with an adjustment task in which one sensory parameter was kept fixed and the other sensory parameter was adjusted to match the fixed parameter as close as possible according to the to be learned mapping. At the end of most trials, we presented the correct audiovisual stimulus as feedback. However, in order to test the stability of the performance e.g. of how well they had learned the mapping we interspersed the session with a few blocks without feedback. Finally, we evaluated the effect of learning in a speeded classification task. We expected successful learning to be reflected: (i) in increased accuracy in the learning task, (ii) to remain robust in the absence of feedback, and (iii) a congruency effect in the speeded

classification task leading to faster reaction times and/or higher accuracies in response to stimulus parameter combinations from the learned mapping compared to stimulus parameter combination from the anti-correlated mapping.

Methods

Participants

After giving written informed consent, 16 participant (7 female, mean age: 23 years) took part in this experiment. Each had normal or corrected-to-normal vision, reported normal hearing, and had no history of neurological or psychiatric illness. Participation was rewarded with course credits. The study was approved by the local research and ethics committee.

Stimuli

The visual stimuli were 51 circles (radius sampled linearly from 0.5° to 10.5 visual angle, line thickness of 0.2°) of light gray colour (mean luminance: 50.08 cd/m^2) presented on a dark gray background (mean luminance: 33.58 cd/m^2).

Auditory stimuli were 51 pure tones (frequencies sampled linearly from 250 Hz to 4000 Hz, with linear onset and offset ramps of 1 ms to avoid auditory clicks, sampling rate 44100 Hz). The tones were presented through headphones at a sound level of approximately 75 dB.

From these 51 auditory and visual stimuli two linear anti-correlated mappings were constructed (Figure 4.1). Each circle was assigned to a particular tone in a

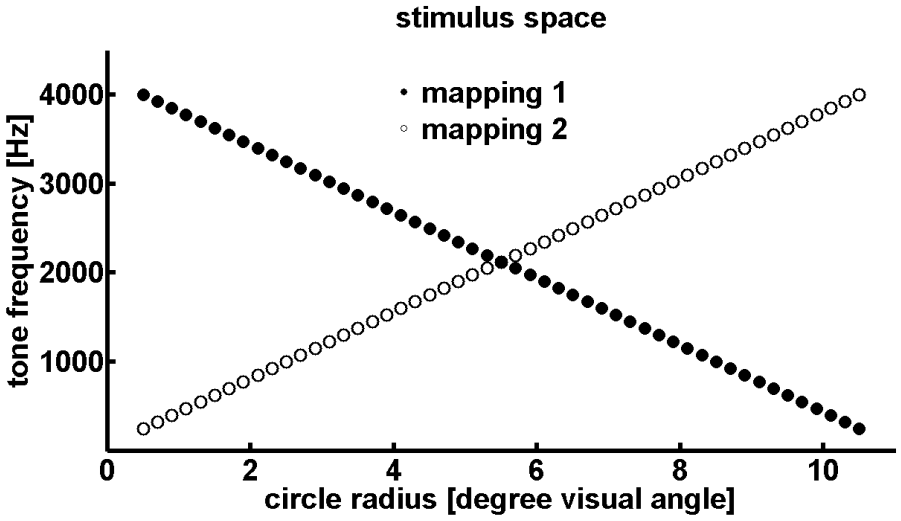


Figure 4.1 Stimulus space. An audiovisual stimulus space spanned by circle radius in degree visual angle (x-axis) and pure tone frequency in Hz (y-axis). Linear combinations of the 51 auditory and visual parameters resulted in two anti-correlated mappings: mapping 1: high pitch with small size and low pitch with large size; mapping 2: high pitch with large size and low pitch with small size.

linear fashion. In mapping 1, the highest frequency was mapped onto the smallest size and the lowest frequency onto the largest size. In mapping 2, high frequency was mapped onto large size and low frequency on small size.

Experimental Procedure

The experiment consisted of two tasks: a speeded classification task and a learning task (Figure 4.2). Using the speeded classification, we first accessed if any baseline differences in pitch-size mapping were present between the groups

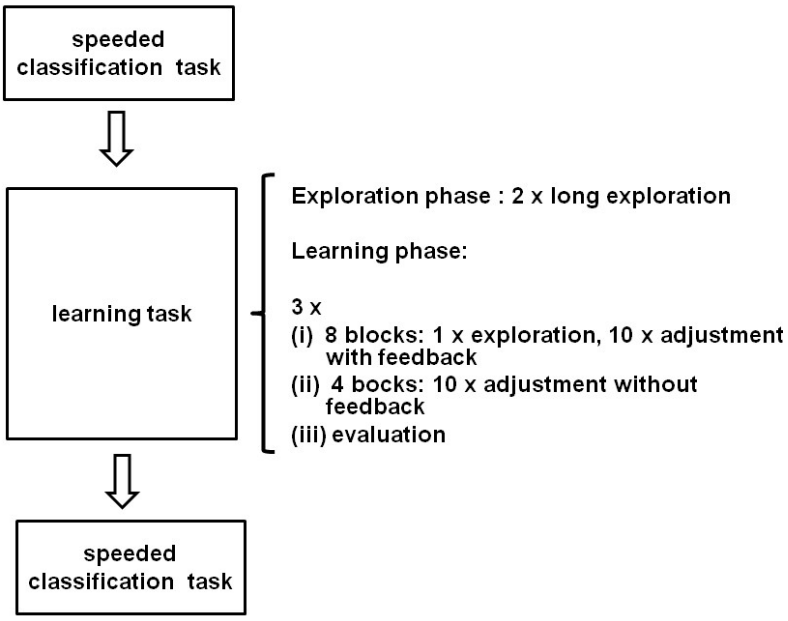


Figure 4.2 Experimental procedure. The experimental session started with 16 blocks (8 auditory task blocks, 8 visual task blocks, in permuted order) of the speeded classification task. In the learning task, first 2 long exploration trials were presented (204 exploration steps per trial), followed by 3 runs of the learning phase. Each run contained: (i) 8 blocks of 1 exploration trial (50 exploration steps) and 10 adjustment trials with feedback, (ii) 4 blocks of adjustment trials without feedback, (iii) in runs 1 and 2 only: evaluation of the learning progress. The session ended with another 16 blocks of the speeded classification task.

prior to learning, and second, after learning, if response patterns of reaction times and accuracy were showing a congruency effect in favor of the learned mapping. In the learning task, participants were trained in an active audiovisual learning paradigm to map the auditory and visual stimulus parameters according to one of

the two mappings. Half of the participants were trained on mapping 1 and the other half on mapping 2.

The session started with the speeded classification task, followed by the learning task, and finished with another run of the speeded classification task. Participants were instructed to fixate a central fixation cross throughout the entire experiment.

Learning task

Participants learned in an active audiovisual learning paradigm one of two anti-correlated mappings spanned by the stimulus parameters auditory pitch in Hz and visual circle radius in degree visual angle. Each mapping contained 51 audiovisual stimuli. The task consisted of two kinds of trials: (i) exploration trails and (ii) adjustment trials (Figure 4.3).

Exploration trials allowed to actively explore the stimulus space and to learn the one-to-one mapping between circle size and tone frequency. Each trial started with a fixation period of 500 ms. It was followed by an audiovisual stimulus (250 ms duration) that was sampled randomly from the to be learned stimulus mapping. From there, participants proceeded - by pressing either the left or the right arrow- to the next or previous audiovisual stimulus of that mapping i.e. increased or decreased a circle radius by 0.2° and a sound by 75 Hz. If no response was given within 1200 ms after stimulus onset, a warning ‘explore faster’ was displayed. The trial terminated after a pre-defined number (see below) of exploration steps was completed.

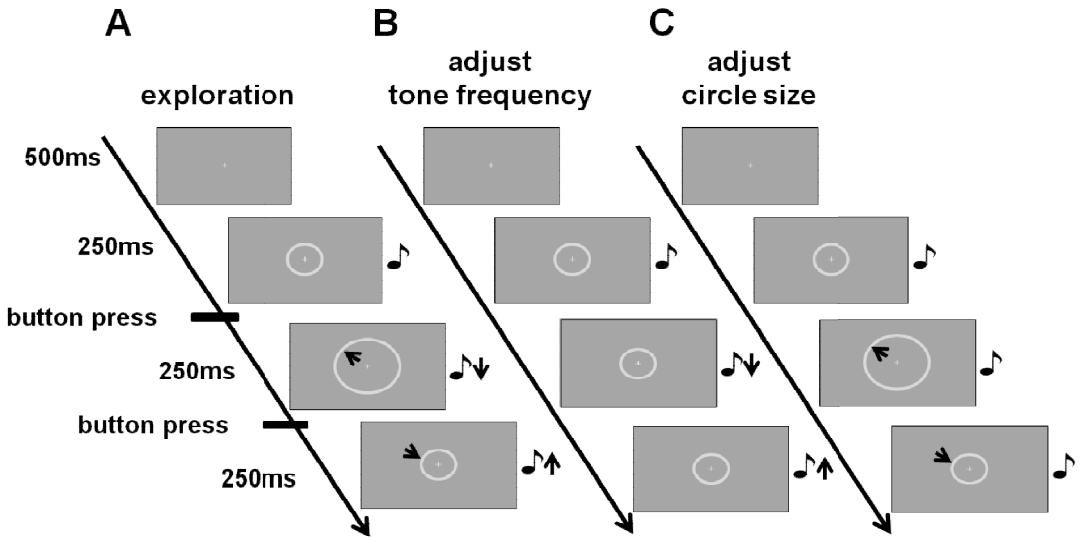


Figure 4.3 Learning task trial types. Each trial started with a fixation period of 500ms. Then, a tone and a circle were presented for 250ms. A button press initiated the presentation of the next stimulus parameter from the to be learned mapping i.e. increased or decreased circle size by 0.2° visual angle and/or tone frequency by 75 Hz. A. Exploration. In exploration trails, the tone frequency and circle size changed together and always corresponded to a stimulus from the to be learned mapping. B. Adjust tone frequency. In auditory adjustment trials, the circle size was fixed and the frequency of the tone had to be matched to the size of the circle. C. Adjust circle size. In visual adjustment trials, the tone frequency was fixed and the circle size had to be matched to the frequency of the tone.

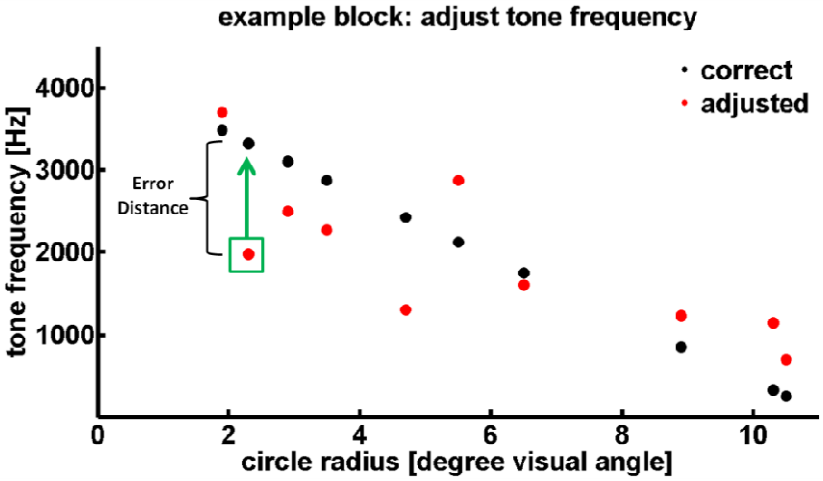


Figure 4.4 Example block: adjust tone frequency. Results of an auditory adjustment block. The stimulus space is spanned by circle radius (x-axis) and tone frequency (y-axis). Each black dot represents one of 10 audiovisual stimulus parameter combinations of auditory pitch and circle size that were sampled for this block from the to be learned mapping. Red dots represent the audiovisual combination of the fixed circle size and the matched tone frequency. The error distance in stimulus space unit between the correct (black) and matched (red) audiovisual stimulus reflects how well participants matched the tone frequency. If the match was perfect, the error distance would be zero. A similar figure was shown to participants during evaluation of the learning progress at the end of run 1 and 2.

Adjustments trials tested how precisely participants were able to reproduce the mapping. There were four kinds of blocks: auditory adjustment, visual adjustment, blocks with feedback, and blocks without feedback. The latter was included to test if the learning performance was stable enough to be maintained without feedback. For each block, 10 audiovisual stimuli were sampled randomly from the to be

learned mapping. Each trial started with a fixation period of 500 ms. It was followed by an audiovisual stimulus (250 ms duration), however, unlike in exploration trials, the two stimulus parameters were not matched but selected semi-randomly from the 10 parameter combinations sampled for that block. One of the sensory parameters was kept fixed and the parameter from the other sensory modality had to be adjusted to match the fixed parameter according to the to be learned mapping. The parameter was adjusted in the same fashion as on exploration trials, by moving through the stimulus space using a button press, just that this time, only one parameter was changing. There were two kinds of adjustment blocks: visual and auditory adjustment. In visual adjustment blocks, the frequency of the tone was kept constant and circle size had to be adjusted. In auditory adjustment blocks the circle size was fixed and the frequency of the tones had to be adjusted. The type of adjustment was indicated by a cue at the start of each block. The distance in stimulus space between the correct and adjusted parameter indicated how well the distribution has been learned (Figure 4.4). In order to allow participants to be as accurate as possible, there was no time or step limit for the adjustment trials, as long as participants kept actively adjusting the stimuli. The trial terminated if no response was given for 2 s. In blocks with feedback, the correct audiovisual stimulus was presented for 500 ms.

The learning phase started with a few practice trials of each kind of trials, which were repeated until we were sure that the participant had understood the task. The task started with the exploration phase. Participants explored the mapping in two blocks of 204 exploration steps respectively. The main experiment consisted of

three runs, each lasting about 30 min. Each run comprised of: (i) 8 blocks, alternating auditory and visual adjustment, presenting one exploration trial limited to 50 exploration steps, in order to refresh the mapping, and 10 adjustment trials with feedback, (ii) 3 blocks (4 blocks in run 3) of 10 adjustment trials without feedback, and (iii) in runs 1 and 2 only, an evaluation of the learning performance during which a graphical image of the correct stimuli and adjustment results of the last 30 adjustment trials were displayed, depicting how close the estimated responses were to the true parameters (Figure 4.4).

Speeded Classification Task

For the speeded classification task we selected the 2 auditory and 2 visual stimuli that were positioned between the stimuli at a distance of 1 quartile from the end of the stimulus space and the next outer stimulus. These were the two visual circles with radii of 2.9° (small) and 7.6° (large), and the two tones with the frequencies of 1215.5 Hz (low pitch) and 2962.5 Hz (high pitch). Importantly, none of these stimulus parameters was presented during learning.

On each trial, participants were presented with an audiovisual stimulus (120 ms duration, SOA 1500 ms) defined by pitch (high, low) and size (large, small). Thus, four audiovisual stimulus combinations were presented with equal probability: low pitch/large visual size, low pitch/small visual size, high pitch/large visual size and high pitch/small visual size. We will refer to the stimulus combinations low/large and high/small as mapping 1 and to the stimulus combinations low/small and high/large as mapping 2 (Figure 2.1). In a selective attention paradigm,

participants performed a two-choice discrimination task that focused either on the auditory frequency or the visual size dimension. Participants discriminated between small and large size in the visual task or high and low pitch in the auditory task as fast and accurately as possible. Further, they were instructed to fixate a central fixation cross throughout the entire experiment.

Participants were presented at the beginning of the session with 4 practice blocks. The main experiment included two runs, one before and one after the learning task. Each run consisted of 8 auditory and 8 visual attention task blocks that were presented in permuted order to facilitate interference effects. The task-relevant sensory modality was indicated at the beginning of each block. Within each block, each of the four possible audiovisual stimulus combinations was presented twice in random order. The order of auditory/visual tasks was counterbalanced across participants. The start of each block was initiated by button press in order to allow participants to switch between the different response-mappings for the auditory and visual task.

Responses were given via four different buttons: ‘A’, ‘D’, ‘J’ and ‘K on a conventional keyboard. The buttons were chosen to ensure that participants used different hands to respond during the auditory and visual tasks in order to avoid interference and transference effects at the response level. The mapping of response buttons was counterbalanced across participants resulting in eight response mapping options:

auditory right hand and visual left hand: (i) ‘J’ = low (auditory), ‘K’ = high (auditory), ‘A’ = small (visual), ‘D’ = large (visual), (ii) ‘J’ = low (auditory), ‘K’ = high (auditory),

‘A’ = large (visual), ‘D’ = small (visual), (iii) ‘J’ = high (auditory), ‘K’ = low (auditory), ‘A’ = small (visual), ‘D’ = large (visual), (iv) ‘J’ = high (auditory), ‘K’ = low (auditory), ‘A’ = large (visual), ‘D’ = small (visual)

Apparatus

The experiment was conducted in a dimly lit experimental room. Constant viewing distance was achieved by stabilizing the participant’s head on a chin rest at a distance of 50 cm from a LED monitor (1920 × 1080 resolution, 60 Hz refresh rate, iiyama Proline, Japan). Auditory stimuli were presented through headphones (Sennheiser HD 555MR, Germany) at approximately 75 dB SPL. Experimental sessions were presented using Cogent 2000 v1.25 (developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON at the Wellcome Department of Imaging Neuroscience, UCL, London, UK; http://www.vislab.ucl.ac.uk/cogent.php) running under MATLAB (Mathworks Inc., Natick, MA, USA) on a Windows PC. The responses were given via a conventional keyboard.

Analysis

Learning task

Learning performance was measured as error distance in stimulus space units between the correct and adjusted parameter of each adjustment trial. First, the value of the smallest parameter (250 Hz for frequency and 0.5° for circle size) was subtracted from the raw corrected and adjusted parameter values. Then, the error

distance between the stimulus parameters was divided by the distance between two adjacent stimulus parameters in stimulus space (75 Hz for frequency and 0.2° for circle size). Auditory and visual adjustment trials were analyzed separately. For each block, the average absolute error distance was calculated.

Successful learning was defined as a significant difference between the first block and the average of the last two blocks. The difference in error distance between the first block and the average of the last two blocks, per adjustment modality, was entered into a 2 (between subject factor group: learned mapping 1 vs. 2) x 2 (within-subject factor adjustment modality: circle size vs. sound frequency) repeated-measures mixed model ANOVA.

In order to test how well participants were able to maintain the learned mapping without feedback, we compared if the learning performance of the last blocks without feedback was significantly different from the very first block. The difference in error distance between the first block and the average of the last two blocks without feedback, per adjustment modality, was entered into a 2 (between subject factor group: learned mapping 1 vs. 2) x 2 (within-subject factor adjustment modality: circle size vs. sound frequency) repeated-measures mixed model ANOVA.

Speeded classification task

In order to control for any group related differences before learning we entered reaction times (based on within-subject median, after excluding incorrect trials and trials with reaction times shorter than 200 ms or longer than 1000 ms) and

accuracy as % correct discrimination of the baseline collected prior to learning into 2 x (group: learned mapping 1 vs. mapping 2) x 2 (task-relevant modality: auditory vs. visual) x 2 (mapping: 1 vs. 2) repeated-measures mixed model ANOVAs.

The effect of learning was accessed by entering reaction times (based on within-subject median, after excluding incorrect trials and trials with reaction times shorter than 200 ms or longer than 1000 ms) and accuracy of the post-test in 2 (task-relevant modality: auditory vs. visual) x 2 (congruency with learned mapping: congruent vs. incongruent) x 2 group (learned mapping: 1 vs. 2) repeated-measures mixed model ANOVAs.

Results

Learning task

Participants learned to match the pitches of auditory tones with the sizes of visual circles according to one of two anti-correlated linear mappings that related 51 auditory pitch parameters and 51 visual size parameters in a linear fashion. Mapping 1 related small size to high pitch and large size to low pitch. Mapping 2 related small size with low pitch and large size with high pitch. The accuracy of the match was quantified as error distance between the correct and the matched auditory and visual parameters respectively.

First, we accessed if participants have successfully learned their respective mapping by comparing how closely they were able to match the auditory and visual stimulus parameters in the first block and the last two blocks with feedbacks (Table 4.1). A 2 (between subject factor group: learned mapping 1 vs. 2) x 2

(within-subject factor adjustment modality: circle size vs. sound frequency) repeated-measures mixed model ANOVA on the difference in error distance between the first and the average of the last two blocks per adjusted modality did not reveal any significant effects of group and adjustment modality (Table 4.2). Therefore, we pooled the data over these factors and entered the error distance difference between the first and the average of the last two blocks (Table 4.3) into a one-sample t-test. There was a significant effect of learning ($t(15) = 3.71$, $p = .002$) that was further confirmed by a linear regression analysis ($\beta = -0.095$, $p < .000$).

Second, in order to access if participants were able to reproduce the mapping also in absence of feedback, we repeated the analysis with as dependent variable: the difference between the error distance of the first block and the average error distance of the last two blocks without feedback (Table 4.1). Again, there was no significant effect of group or adjustment modality (Table 4.2), therefore, the data was pooled over adjustment modality and group. The error distance between the first and the average of the last two blocks without feedback (Table 4.3) was entered into a one-sample t-test which revealed a trend indicating a learning progress ($t(15) = 1.85$, $p = .085$).

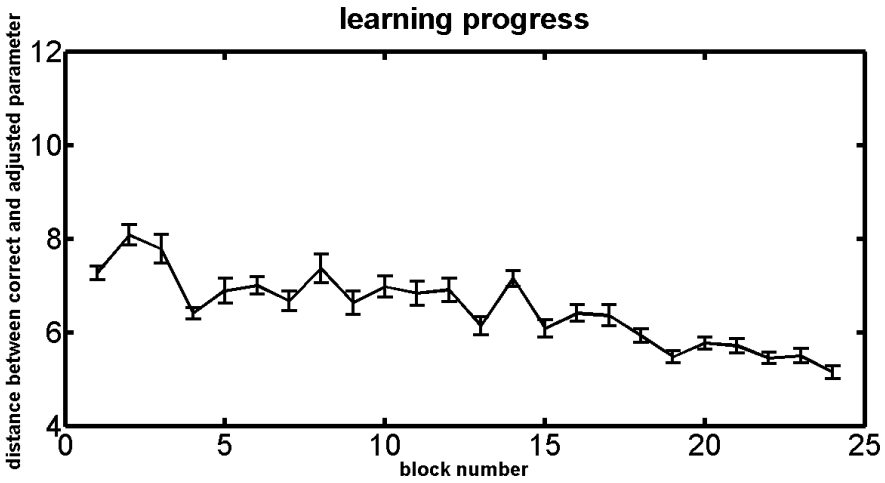


Figure 4.5 Results learning task. Learning progress as error distance between the correct and matched parameters on blocks with feedback.

Table 4.1 Summary of mean error distance and its standard mean errors for the learning task.

group:	learned mapping 1	learned mapping 2
adjust tone		
block 1 (with feedback)	6.95 (0.38)	7.59 (0.18)
last blocks (with feedback)	5.82 (0.36)	5.39 (0.23)
last blocks (no feedback)	5.90 (0.33)	6.59 (0.36)
adjust size		
block 1 (with feedback)	9.8 (0.47)	6.35 (0.27)
last blocks (with feedback)	5.65 (0.25)	4.95 (0.26)
last blocks (no feedback)	6.76 (0.38)	6.53 (0.36)

Table 4.2 Statistical results of the learning task.

	learning task (df; 1,14)	
	feedback trials	no-feedback trials
adjustment-modality	F = 1.43, p = .252	F = 0.15, p = .706
adjustment-modality x group	F = 4.19, p = .060	F = 2.27, p = .154
group	F = 0.48, p = .500	F = 1.58, p = .229

* $p < 0.05$.

Table 4.3 Summary of mean error distance and its standard mean errors for the learning task results pooled over group and adjustment modality.

	pooled over group and adjustment modality
block 1 (with feedback)	7.67 (0.15)
last blocks (with feedback)	5.45 (0.10)
last blocks (no feedback)	6.45 (0.09)

These results indicate that in the absence of feedback participants, on average, were not able to reproduce the mapping as well as when they received feedback.

In summary, the results of the learning task show that participants were able to learn the general mapping between auditory pitch and visual size but that in order to reproduce it they were relying on feedback. Furthermore, participants learned both mapping equally well and it did not make a difference if the frequency of the tone or the size of the circle was adjusted.

Speeded Classification Task

The speeded classification task was presented twice: as pre-test in order to ensure that there were no differences between the groups before learning and

after the learning task was completed (post-test) in order to inspect if learning a specific mapping between auditory pitch and visual size resulted in faster and more accurate responses to audiovisual stimuli that were congruent with the learned mapping compared to stimuli that were incongruent with the learned mapping (Table 4.4).

First, we tested if any group effects were present before learning. A 2 (learned mapping: 1 vs. 2) x 2 (task-relevant modality: auditory vs. visual) x 2 (mapping: 1 vs. 2) repeated-measures mixed model ANOVA at pre-test did not reveal any significant group effects, neither for reaction times nor for accuracy (Table 4.5). Participants from both groups responded faster during the visual task than during the auditory task (main effect of task-relevant modality: Table 4.5). Furthermore, there was a significant two-way interaction between mapping and task-relevant modality (Table 4.5). The follow-up 2 (learned mapping: 1 vs. 2) x 2 (mapping: 1 vs. 2) repeated-measures mixed model ANOVAs for the auditory and for the visual task did not reveal any significant effects (Table 4.6).

Next, we tested if learning has affected reaction times and accuracy for audiovisual stimuli that were congruent with the learned mapping compared to audiovisual stimuli that were incongruent (Table 4.7). A 2 (task-relevant modality: auditory vs. visual) x 2 (congruency with learned mapping: congruent vs. incongruent) x 2 group (learned mapping: 1 vs. 2) repeated-measures mixed model ANOVA of reaction times of the post-test revealed a significant main effect of task-relevant modality and a significant three-way interaction between congruency, task-relevant modality and group (Table 4.8).

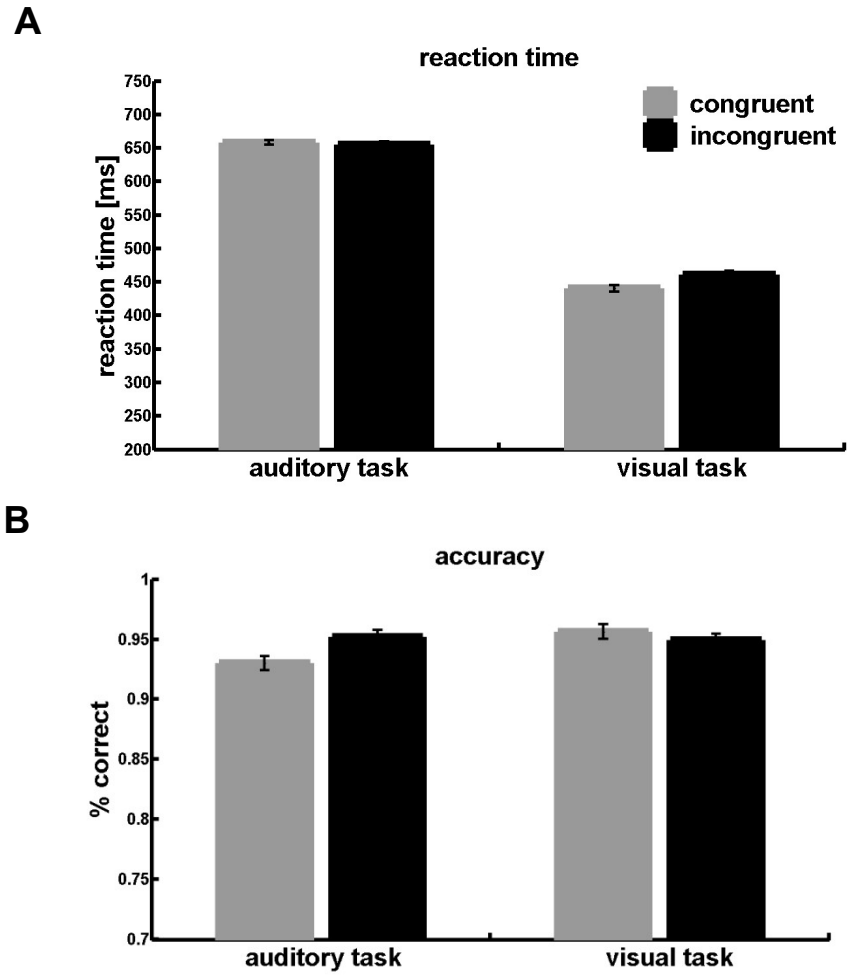


Figure 4.6 Results speeded classification task after learning. A. Bar plot showing reaction times (across participants' mean +/- SEM) for stimuli congruent with the learned mapping and stimuli incongruent with the learned mapping. B. A bar plot showing accuracy (across participants' mean +/- SEM) for stimuli congruent with the learned mapping and stimuli incongruent with the learned mapping.

Table 4.4 Summary of mean reaction time (in ms) and accuracy (% correct) and their standard mean errors for the speeded classification task before learning.

	learned mapping 1		learned mapping 2	
	reaction time	accuracy	reaction time	accuracy
mapping 1				
auditory	706.36 (6.31)	0.94 (0.02)	689.06 (7.27)	0.90 (0.01)
visual	443.38 (8.44)	0.86 (0.02)	454.13 (11.23)	1.00 (0.00)
mapping 2				
auditory	670.63 (10.15)	0.98 (0.01)	630.88 (6.84)	0.93 (0.01)
visual	443.38 (8.44)	0.86 (0.02)	454.13 (11.23)	1.00 (0.00)

Table 4.5 Statistical results of the speeded classification task before learning.

	baseline (df; 1,15)	
	reaction time	accuracy
mapping	F = 0.64, p = .436	F = 0.01, p = .932
mapping x group	F = 14.0, p = .717	F = 2.89, p = .111
task-relevant-modality	F = 135.02, p < .000*	F = 3.92, p = .068
task-relevant-modality x group	F = 0.07, p = .801	F = 0.75, p = .401
mapping x task-relevant-modality	F = 6.64, p = .022*	F = 0.35, p = .854
mapping x task-relevant-modality x group	F = 3.25, p = .093	F = 0.63, p = .806

* $p < 0.05$.

Table 4.6 Statistical results of follow-up ANOVAs on reaction time for auditory and visual task of the speeded classification task before learning.

	auditory task	visual task
mapping	F = 1.92, p = .188	F = 3.45, p = .085
mapping x group	F = 3.84, p = .070	F = 0.50, p = .493
group	F = 0.44, p = .517	F = 0.38, p = .547

* $p < 0.05$.

A follow-up 2 (congruency) x 2 (group) repeated-measures mixed model ANOVA for the auditory task demonstrated a significant two-way interaction between congruency and group (Table 4.9). However, the follow-up one-way ANOVAs on congruency did not reveal any significant effects, neither in the group who learned mapping 1 ($F(1,7) = 1.44$, $p = .250$) nor in the group who learned mapping 2 ($F(1,7) = 1.15$, $p = .302$). A follow-up 2 (congruency) x 2 (group) repeated-measures mixed model ANOVA for the visual task did not reveal any significant effects (Table 4.9). Thus, both groups of participants those who learned mapping 1 and those who learned mapping 2 responded equally to the audiovisual stimuli, irrespective if the stimuli were congruent or incongruent with the learned mapping.

Table 4.7 Summary of mean reaction time (in ms) and accuracy (% correct) and their standard mean errors for the speeded classification task after learning.

	learned mapping 1		learned mapping 2	
	reaction time	accuracy	reaction time	accuracy
congruent				
auditory	627.50 (6.34)	0.96 (0.01)	689.06 (7.27)	0.90 (0.01)
visual	426.63 (7.19)	0.91 (0.02)	454.13 (11.23	1.00 (0.00)
incongruent				
auditory	678.63 (11.37)	0.98 (0.01)	630.88 (6.84)	0.93 (0.01)
visual	439.75 (13.87)	0.92 (0.01)	481.56 (8.69)	0.98 (0.01)

Table 4.8 Statistical results of the speeded classification task after learning.

	post-test (df; 1,15)	
	reaction time	accuracy
congruency	F = 0.38, p = .546	F = 0.19, p =.673
congruency x group	F = 3.07, p = .101	F = 0.14, p = .717
task-relevant-modality	F = 137.54, p < .000**	F = 0.18, p = .676
task-relevant-modality x group	F = 0.62, p = .443	F = 5.06, p = .041*
congruency x task-relevant-modality	F = 0.838, p = .376	F = 0.51, p = .489
congruency x task-relevant-modality x group	F = 5.645, p = .032*	F = 0.312, p = .585
group	F(1,15) = 0.56, p = .466	F(1,15) = 0.139, p = .714

* p < 0.05.

Table 4.9 Statistical results of follow-up 2-way ANOVAs on reaction time for auditory and visual task after learning.

	auditory task (df: 1,15)	visual task (df: 1,15)
congruency	F = 0.4, p = .845	F = 1.05, p = .324
congruency x group	F = 9.55, p = .008*	F = 0.13, p = .723
group	F = 0.06, p = .806	F = 0.86, p = .368

* p < 0.05.

Table 4.10 Statistical results of follow-up 2-way ANOVAs on accuracy for auditory and visual task after learning.

	auditory task (df: 1,15)	visual task (df: 1,15)
task-relevant modality	F = 0.77, p = .396	F = 0.07, p = .799
task-relevant modality x group	F = 0.04, p = .838	F = 0.40, p = .538
group	F = 1.87, p = .193	F = 3.71, p = .075

* p < 0.05.

A 2 (task-relevant modality: auditory vs. visual) x 2 (congruency with learned mapping: congruent vs. incongruent) x 2 group (learned mapping: 1 vs. 2) repeated-measures mixed model ANOVA of accuracy of the post-test demonstrated a significant two-way interaction between task-relevant modality and group (Table 4.8). However, the follow-up 2 (task-relevant modality) x 2 (group) repeated-measures mixed model ANOVAs for the auditory task and visual task did not reveal any significant effects (Table 4.10).

The bar graphs shows only for the visual task a response pattern, both in reaction times and accuracy, that is consistent with a congruency effect induced by learning (Figure 4.6).

Discussion

In this study, we attempted to train participants to reproduce a one-to-one mapping between auditory pitch and visual size. The results of our previous experiments have demonstrated that the pitch-size mapping is not as robust as other cross-modal metaphoric mappings like pitch-elevation or pitch-brightness. One of the theories states that such relationships between stimulus features that do not share any of the classical binding cues like temporal, spatial or semantic congruency become related because they tend to co-occur in some combinations more likely than in other combinations (Marks, 2004; Spence, 2011; Deroy & Spence, 2013). We used the pitch-size stimulus features to artificially induce two specific mappings in two groups of participants.

Our results demonstrate that within one session participants were able to learn the mapping. However, they had not learned it profoundly enough to be able to reproduce it without feedback or to induce congruency effects in a speeded classification task. Participants successfully learned to match the pitch and size parameters according to the mapping that they were trained on. Both groups learned equally well and there was no difference in matching the auditory or the visual parameters. However, the drop in accuracy in the last few blocks without feedback indicated that the learned mapping was not stable yet. Furthermore, in the speeded classification task learning did not produce any decrease in reaction times nor increased accuracy for audiovisual stimulus combinations that were congruent with the learned mapping.

There is a multitude of reasons that have potentially impaired the success of this experiment. First of all, in the learning task, we were surprised to encounter a drop in accuracy in the last blocks without feedback. We had expected that throughout the session the mapping would become stable enough to not depend on regular feedback. Instead, participants seemed confused by the absence of feedback. Retrospectively, we realized that in the adjustment trials most of the stimulus parameter combinations were in fact incongruent with respect to the target mapping. This has very likely interfered with the learning progress because it made the learning task more difficult than we intended. There was also a drop in performance between the first block (that followed a prolonged period of exploration i.e. presentation of congruent stimuli) compared to the second block, which confirms the suspicion that the adjustment task was not as efficient for

learning the mapping as the exploration task. Moreover, the interfering stimuli might have disturbed the still fragile pitch-size mapping and diminished effects of learning before the speeded classification task was presented, which would explain why no significant congruency effects were found. For the future, we shall revise this task and reduce interference from incongruent stimuli.

Furthermore, the learning phase of only 90 minutes was not sufficiently long for learning the mapping well enough to reproduce it accurately. Participants have learned the general direction of the mapping and the rough correspondence between pitch and size parameters relatively quickly. Already, in the very first block, straight after the exploration phase, they achieved, on average, to match parameters within 8 units of the target mapping that consisted of 51 stimulus units in total. However, throughout the experiment, they were not able to reproduce the one-to-one mapping and towards the end of the session they were missing the correct parameters still by more than 5 units. As indicators for a robustly learned mapping, we expected a smaller error distance and no drop in performance in the absence of feedback. However, naturally emerging metaphoric mappings or learning of any new skill takes time. Therefore, in order to reach better specificity, a longer duration of the learning phase is recommended (Jeter, Dosher, Liu & Lu, 2010). In future, we shall allow participants more time to learn the mapping profoundly, splitting learning over several days and increase the duration of the learning phase.

Additional improvements for future versions of this experiment include: (i) adding an informative pre-learning baseline measure and (ii) making the task more

engaging. Firstly, in this study, we used the first block of the learning task as the starting measure of the learning progress, after participants were already extensively exposed to the mapping during the exploration phase. We included the speeded classification task at the start of the experiment in order to capture the implicit pre-existing mapping. However, it would be an informative addition to also obtain a measure of the participants' explicit associations between our stimulus parameters. Secondly, we felt that also improvements in the motivational aspect of the task should be considered. Besides the methodological shortcomings described above, it is very likely that fatigue and loss of motivation towards the end of a long, demanding experimental session contributed to the drop of performance on the last blocks of the learning task. Furthermore, some of the participants commented that the task was abstract and that they experienced difficulties to stay motivated. Therefore, in order to increase the chances of successful learning, it is important to revise the task, to make it more engaging, and successful performance more rewarding.

Taken together, we have demonstrated that within only 90 minutes participants can learn to roughly reproduce a mapping between auditory pitch and visual size in polar coordinates. For future experiments, in order to induce a robust size-pitch mapping, we need to increase the number of learning sessions and to revise the task by reducing interference with the target mapping and increasing the motivation for participants to perform well.

CHAPTER 5: AN ATTEMPT TO INDUCE A SYNAESTHESIA-INSPIRED PITCH-SIZE MAPPING

Introduction

In *Chapter 4*, I presented our attempt to train participants to map auditory pitch and visual size according to an artificially constructed mapping. We created a stimulus space that related 51 auditory and visual parameter to each other in a linear fashion. Mapping 1 related high pitch with small size and low pitch with large size. Mapping 2 related high pitch with large size and low pitch with small size. Participants were divided into two groups, one for each mapping, and trained in an active learning paradigm to reproduce the one-to-one relationship between the pitch and size parameters.

We succeeded to induce learning but it was not sufficient for reproducing the one-to-one mapping between the stimulus parameters. It also failed to produce a congruency effect in a speeded classification task i.e. the reaction times and accuracy in response to audiovisual stimuli from the learned mapping did not differ from responses to stimuli from the anti-correlated mapping.

Two main methodological problems were identified. Firstly, while matching one stimulus parameter to another in an adjustment task, participants were presented with a multitude of incongruent stimuli. This made the learning task much harder than intended and therefore less efficient (Rothen & Meier, 2014). Moreover, it potentially demolished the learning effect in blocks in which no feedback was given, because in those blocks no congruent stimuli were presented at all. Consequently, the presentation of four no-feedback blocks straight before the

speeded classification task might have reduced learning effects dramatically and caused the lack of a congruency effect. Secondly, one single learning session of 90 minutes appears not have been sufficient to achieve a profound mapping. Previous studies demonstrated that participants can learn overall structures rapidly but that achieving greater specificity requires longer periods of time (Jeter, Dosher, Liu & Lu, 2010). Therefore, we should aim to train participants for multiple sessions, ideally until a stable 100% accuracy plateau is reached. Further suggestions for improvements included shorter sessions and an engaging task that encourages participants to stay motivated throughout the whole experiment.

For further improvements of our learning paradigm, we turned for inspiration to research that sought to induce synaesthesia in non-synaesthetic individuals. Synaesthesia, a condition in which individuals perceive additional sensory experiences, like seeing colors when hearing a sound, or seeing a letter 'C' always as blue. Specifically, such additional sensory experiences are consistent within individuals, they occur involuntarily and are triggered automatically by specific stimuli, like in the example above the letter 'C' would always appear blue for one individual and green for another individual (Rothen & Meier, 2014). The causes underlying such sensory cross-mappings are still a matter of debate. Synaesthesia tends to run in families, therefore, it is argued that there is a genetic component involved (Rothen & Meier, 2014). However, there are many forms of synaesthesia that are triggered by cultural artifacts like letters or digits. Therefore, some researchers proposed that this condition is acquired via associative learning (Rich, Bradshaw & Mattingley, 2005; Simner, Harrold, Creed, Monro & Foulkes, 2009),

not unlike the metaphoric mappings that we have discussed in the previous chapters. It has to be noted that there is a fundamental difference in that synaesthetic experiences are highly specific and vivid, while effects of metaphoric mappings are encountered predominantly in experimental settings and participants are usually not aware of them (Deroy & Spence, 2013). The idea that synaesthesia could be learned inspired numerous studies to attempt inducing synaesthesia (Kelly, 1934; Howells, 1944; Meier & Rothen, 2009; Rothen, Wantz & Meier, 2011; Colizoli, Murre & Rouw, 2012; Kusnir & Thut, 2012; Cohen Kadosh et al., 2005; Nunn et al., 2002; Brang, Rouw, Ramachandran & Coulson, 2011; Niccolai, Wascher & Stoerig, 2012; Bor, Rothen, Schwartzman, Clayton & Seth, 2014). None, except of two of these studies have succeeded to re-create synaesthetic-like experiences (Howells, 1944, Bor et al., 2014). However, most of them succeeded in inducing learning effects in independent behavioral measures, for example in the synaesthetic Stroop task or the contextual priming task (Howells, 1944; Meier & Rothen, 2009; Rothen et al., 2011; Colizoli et al., 2012; Niccolai et al., 2012; Bor et al., 2014). This is not far from our aim. In terms of this framework, we are in fact aiming at inducing a synaesthesia-like mapping between pitch and size.

In this chapter, I introduce a revised learning task. First of all, we gave the mapping between auditory pitch and visual size a purpose by making the task to resemble a computer game. The cover story of 'Squeaky Planets' instructed participants to learn by means of exploration and a dual forced-choice task to distinguish audiovisual stimuli from the target mapping, aka galaxy 'Calax', from

audiovisual stimuli originating from other galaxies (distracters). Each planet was unique in size and the sound that it produced. The distracter stimuli were maximally 5 stimulus units orthogonally off the to be learned mapping and narrowed it in 5 difficulty levels down to 1 unit, thereby constantly challenging participants to learn the mapping well enough to level-up to the next difficulty level and to avoid leveling down. Crucially, while we still presented incongruent stimuli they were similar to the target mapping and the increase in difficulty level lead participants gradually towards distinguishing the target stimulus from a distracter at only one unit from the target mapping, indicating successful one-to-one mapping. Furthermore, the experiment consisted of multiple learning sessions. Learning was evaluated in a speeded classification task and a newly added modified version of the adjustment task that aimed at revealing the participant's explicit mapping. After a satisfactory level of over-learning has been reached participants were meant to undergo two functional Magnetic Resonance Imaging (fMRI) sessions. The procedure was meant to be repeated with the other mapping. Our previous experiment (*Chapter 3*) demonstrated high inter-subject variability, therefore, a within-subject comparison of learning effects of the two-anti-correlated mappings presents a more appropriate approach.

This chapter covers a pilot case study in which we tested the revised task. We recruited an experienced participant based on reliability and good performance in pilot versions of this task. The participant was trained on mapping 2 for 9 sessions and then underwent two sessions of fMRI in which auditory and visual stimuli from the learned mapping were presented unimodally. The original experimental

procedure included (i) training until an over-learning plateau was reached on mapping 2, (ii) fMRI, (iii) training until an over-learning plateau was reached on mapping 1, (iv) fMRI. Unfortunately, due to external circumstances, the learning phase was terminated after 9 sessions of learning. At that moment, the over-learning plateau was still not reached, however, the learning and behavioral task were already showing a response pattern consistent with effects of learning. Given such indications, we expected good chances that learning effects could have transferred to the neural level already and, therefore, decided to continue with the fMRI experiment despite not having reached the targets in the behavioral measures. Because the experiment was interrupted we were not able obtain data to contrast the learning effects between the two mappings. Furthermore, we also lack a pre-learning baseline because the participant was not naive to our learning paradigm. Therefore, we are limited in the conclusions that we can draw from this experiment.

Taken together, we expected that ideally, the participant would level-up in the learning task until a stable performance at the highest difficulty level could be maintained - an indication that a robust one-to-one mapping has been achieved. Furthermore, we expected that the participant would be able to successfully reproduce the learned mapping in the adjustment task and in the speeded classification task respond significantly faster and more accurately to audiovisual stimuli congruent with the learned mapping compared to audiovisual stimuli from the anti-correlated mapping. Using fMRI and support vector regression (SVR) we expected to be able to decode both stimulus dimensions from regions of interest in

auditory cortices and along the ventral and dorsal visual pathways. Moreover, after having introduced a topographically specific mapping between auditory pitch and visual size in polar coordinates (eccentricity), we expected generalization from one stimulus dimension to another. In terms of cross-modality decoding this means that we expected to be able to decode auditory pitch after a SVR model was trained on BOLD-responses elicited by visual size in polar coordinates and vice versa.

Methods

Participants

One participant (female, age 21) took part in this experiment after giving written informed consent. She had normal vision, reported normal hearing, and had no history of neurological or psychiatric illness. The participant received a monetary reward. The study was approved by the human research ethics committee at the University of Birmingham.

Stimulus Space

The participant learned a stimulus mapping spanned by 40 visual circles (radius sampled linearly from 0.7° to 10.5° visual angle) and 40 auditory pure tones (frequency sampled logarithmically from 500 Hz to 4000 Hz; with linear onset and offset ramps of 10 ms to avoid auditory clicks; sampling rate 44100 Hz). The auditory and visual dimensions were mapped as following: low-frequency tones/

small circle size and high-frequency tones/ large circle size, equivalent to mapping 2 in previous chapters (Figure 5.1 A).

Experimental Procedure

The participant completed 11 sessions: 9 psychophysics/learning- and two MRI sessions. The sessions were completed on different days (Table 5.1).

In sessions 1-9 the participant was trained in a learning task to map visual and auditory parameters according to mapping 2. Furthermore, two tasks evaluated the learning effects: an adjustment task tested the participants' explicit association of the stimulus parameters, and a speeded classification task that revealed the learning effects implicitly i.e. faster reaction times and higher accuracies in response to audiovisual stimulus parameter combinations that were congruent with the learned mapping compared to audiovisual stimulus parameter combinations that were incongruent. The adjustment task was presented in session 1 in order to obtain a pre-learning baseline and in session 9 in order to access if learning had altered the participants association between the stimulus parameters. We expected the post-test to reflect the learned mapping. The speeded classification task was administered three times: in session 1, session 5 and session 9. Finally, sessions 10 and 11 were conducted in the MRI scanner to investigate neural effects of learning.

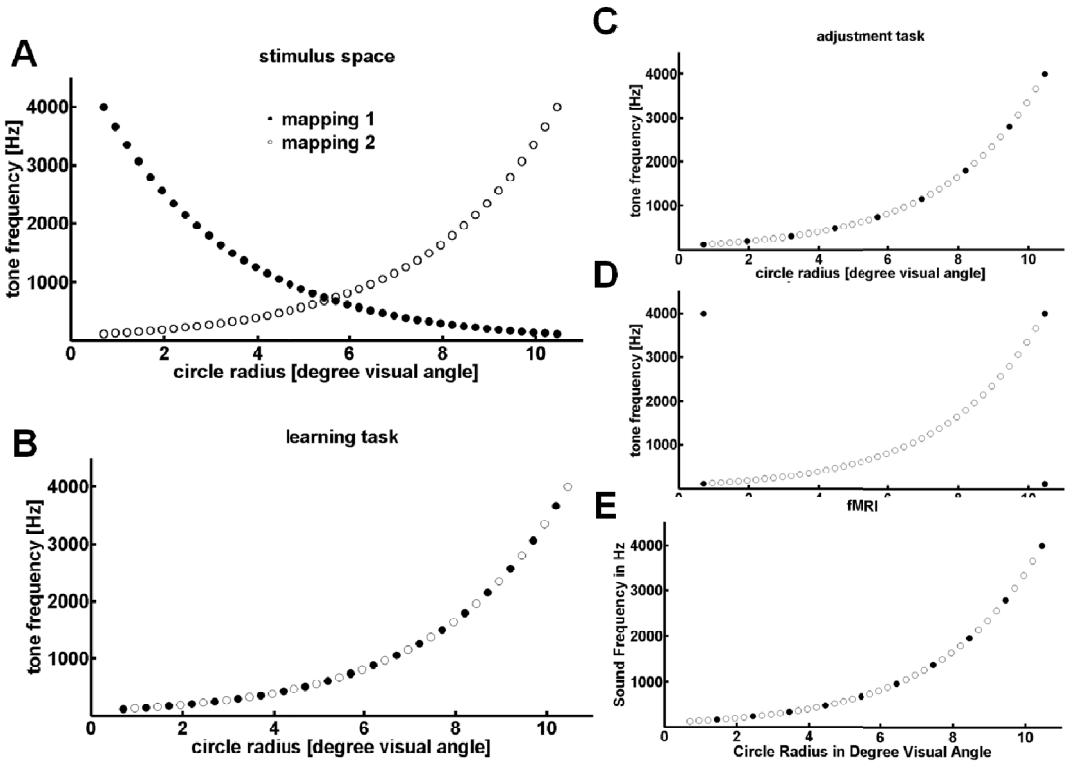


Figure 5.1 Stimulus spaces. A. Stimulus space. On the x-axis circle radius in degree visual angle. On the y-axis sound frequency in Hz. B. Learning task. Audiovisual stimuli presented to the participant in the learning task: only odd stimuli from mapping 2. C. Adjustment task. Every 5th (and the highest auditory and largest visual) auditory and visual parameter was sampled. D. Speeded classification task. E. fMRI. Every 4th stimulus parameter from mapping 2 was sampled. Stimuli were presented unimodally.

Table 5.1 Experimental procedure.

session 1	session 2-4	session 5	session 6-8	session 9	session 10-11
adjustment task speeded classification task learning	learning	learning speeded classification task	learning	learning adjustment task speeded classification task	fMRI

Psychophysics and Learning

Adjustment Task

In the adjustment task every 5th visual and auditory parameter was selected from the stimulus space (9 visual circles with radii of 0.70°, 1.95°, 3.20°, 4.45°, 5.70°, 6.95°, 8.20°, 9.45°, 10.45° visual angle and 9 auditory tones with frequencies of 125 Hz, 195 Hz, 304 Hz, 474 Hz, 739 Hz, 1152 Hz, 1798 Hz, 2803 Hz, 4000 Hz - rounded) (Figure 5.1 C). At the beginning of the task, the participant was familiarized with the stimulus selection. First, the 9 unimodal visual circles were presented in random order, then the 9 unimodal auditory tones. Next, 18 audiovisual trials were presented in which the parameter from one sensory modality was kept fixed and the parameter of the other sensory modality had to be adjusted (sampled semi-randomly from the 9 parameter options). On visual adjustment trials, the tone was kept fixed and the circle size had to be adjusted (Figure 5.2 A). On auditory trials, the circle size was kept constant and the frequency of the tone had to be adjusted (Figure 5.2 B). The visual and auditory

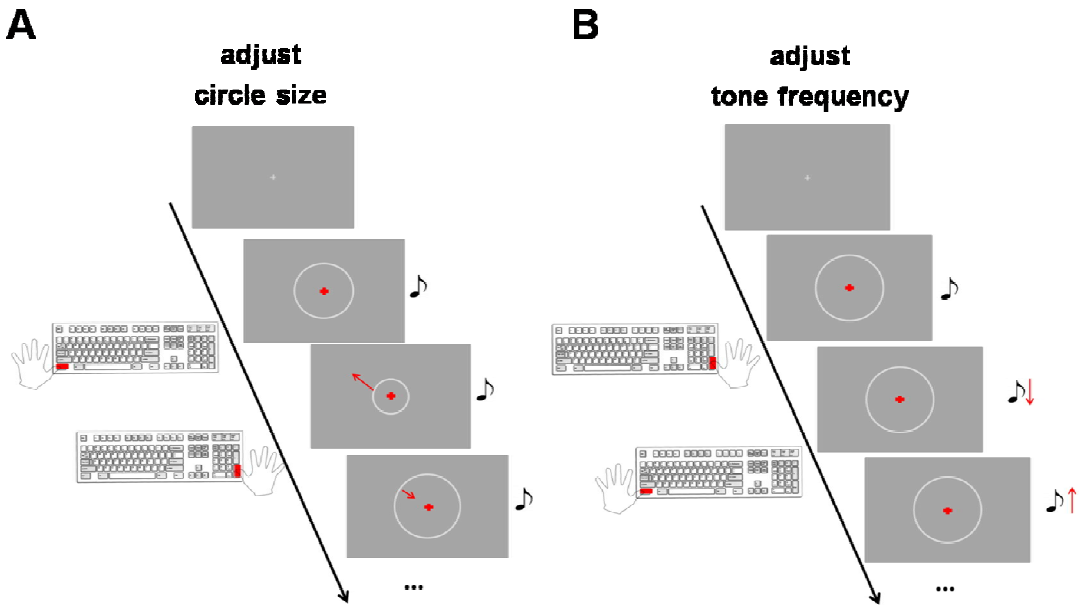


Figure 5.2 Adjustment task. Example trials for visual and auditory adjustment trials. A. Visual adjustment trial. The frequency of the tone was kept fixed and the size of the circle was adjusted via button press. B. Auditory adjustment trial. The size of the circle was kept fixed and the frequency of the tone was adjusted via button press.

adjustment trials were presented in random order. The adjustment modality was indicated at the beginning of each trial. Each trial started with 500 ms fixation, and then the audiovisual stimulus was presented for 250 ms. Next, the participant moved - by pressing a button - through the 9 options and selected one that in her personal opinion matched the fixed parameter. No explicit criteria were given. At the very beginning of the experiment, there was no association between the auditory and visual stimulus parameters. Therefore, the adjustment task before learning should reflect the participant's initial pre-learning baseline stimulus

parameter relationship within the to be learned stimulus space. At the end of the learning phase, the adjustment task was expected to reflect the post-learning concept of the stimulus space.

The results were evaluated as error distance in units of adjustment task stimulus space, separately for the adjusted stimulus parameter from the corresponding parameters in mapping 1 and mapping 2, and separately per adjustment modality.

Speeded Classification Task

The auditory stimuli for the speed classification task were selected to meet three criteria: (i) as close as possible to the extreme end of the mapping, (ii) as similar equal loudness curves as possible, (iii) equal distance from the centre of the mapping. The two tone frequencies meeting these criteria best were 149 Hz and 3348 Hz (rounded). Their matched visual parameters were circle radii of 1.2° and 9.85° visual angle (Figure 5.1 D).

The procedure is identical to the one described in *Chapter 4*.

Responses were given via four different buttons: ‘A’, ‘D’, ‘J’ and ‘K on a conventional keyboard in the following fashion: left hand: ‘A’ = small (visual), ‘D’ = large (visual), right hand: ‘J’ = low (auditory), ‘K’ = high (auditory).

The task was presented three times during the experiment: (i) during the first session, preceded by 4 practice blocks (after the adjustment task and before the first learning task), (ii) during session 5 (after the learning task), and (iii) during session 9 (after the last session of the learning task and the adjustment task).

Learning Task

For the learning task, the stimulus space was divided into odd and even stimuli. One-half of the stimuli (here odd stimuli) was presented during the learning phase (Figure 5.1 B). The task was designed as a computer game called 'Squeaky Planets'. We adopted a cover story and a game-like character for the task in order to keep participants motivated to perform as well as possible. We asked to focus on the size and sound of 'planets' and to learn distinguishing planets from the target galaxy called 'Calax' (stimuli from the to be learned mapping) from planets originating from other galaxies (distracters).

During the first session, after completing the pre-tests, participants were introduced to the cover story and given a few practice trials. The task included two kinds of trials: (i) exploration (Figure 5.3 A) and (ii) dual-forced-choice task (DFC) (Figure 5.3 B).

Exploration trials allowed to actively explore the stimulus space and to learn the one-to-one mapping between circle size and tone frequency. Differently to the exploration task described in *Chapter 4*, we increased the step-size between exploration steps to two stimuli in order to ensure that only odd stimuli were presented during learning. Each trial started with a fixation period of 500 ms. It was followed by an audiovisual stimulus (250 ms duration) that was selected pseudo-randomly from all stimuli that were sampled for the learning task. From

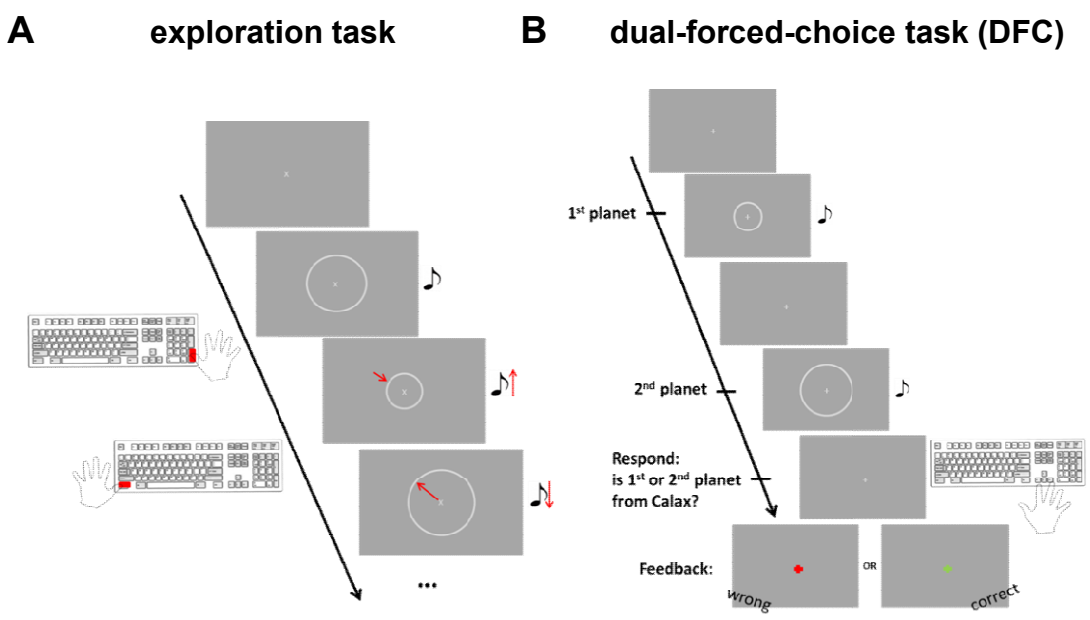


Figure 5.3 Learning task. A. Exploration. After an initial fixation period of 500 ms, a tone and a circle were presented for 250 ms. A button press initiated presentation of the next stimulus parameter from the to be learned mapping. B. Dual-forced choice task. After a 500 ms fixation period, two audiovisual stimuli were presented for 250 ms respectively (ISI 1750 ms). One of the stimuli was from the to be learned mapping, the other a distracter. After a response was detected, feedback was presented for 250 ms.

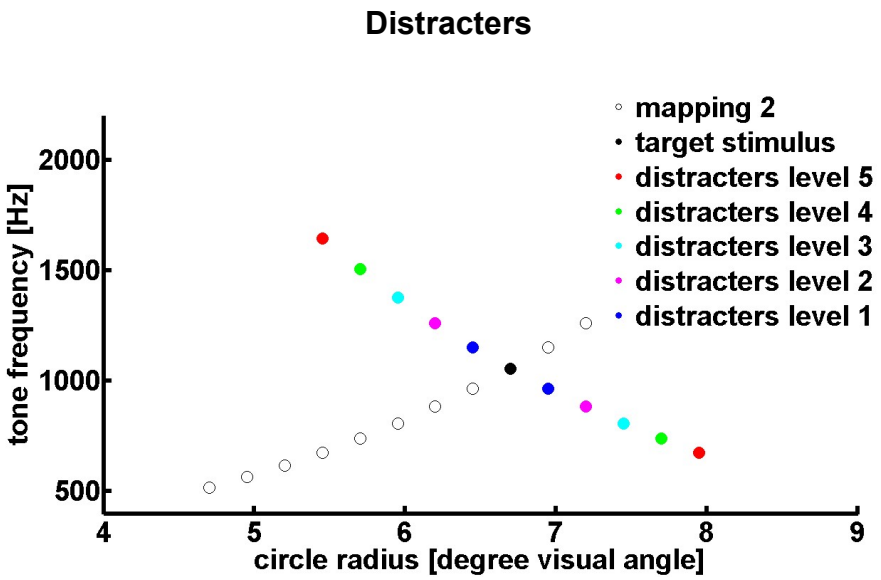


Figure 5.4 Constructing distracters. Distracters of all 5 difficulties level for an example stimulus.

there, participants proceeded - by pressing either the left or the right arrow- to the next or previous audiovisual stimulus in the stimulus space of the to be learned mapping. If no response was given within 1200 ms after stimulus onset, a warning 'explore faster' was displayed. The trial terminated after 100 exploration steps were completed.

In the DFC, two audiovisual stimuli were presented one after the other. One of the stimuli was sampled from the to be learned mapping, the other was a distracter (Figure 5.3 B). The distracters were constructed as follows: two stimuli flanking the target stimulus on the to be learned mapping, at a distance between 5 and 1 units in stimulus space, were selected and their visual and auditory parameters recombined, creating two stimuli flanking the target stimulus in the opposite

direction than the to be learned mapping (Figure 5.4). Then, one of those stimuli was randomly selected as a distracter.

The participant completed 9 learning sessions (20 blocks per session). Each block contained 20 trials of the DFC and one exploration trial (left out in the last block of each session). Within each block, all 20 odd stimuli from mapping 2 were presented once, in random order. Each trial started with 500 ms fixation, followed by the first audiovisual stimulus (250 ms duration), an inter-stimulus-interval (1750 ms duration) and the second audiovisual stimulus (250 ms duration). Participants were asked to indicate via a button response which of the two stimuli was from the to be learned mapping aka 'galaxy Calax'. After the response was given, the fixation cross turned red if the answer was wrong or green if it was correct. At the end of the block, the accuracy was calculated and displayed on an evaluation screen. The accuracy was depicted in a bar graph with markers for the level-up and level-down thresholds. For each block a new bar was added to the graph to demonstrate the participant's learning progress. In order to keep up the spirit of the game, avatars were displayed comparing the performance between the last and the previous block (happy faces if performance improved, a sad looking planet if it worsened, and a neutral sky of stars if it remained the same). If participants reached an accuracy of 85% on a block, they leveled up on the next block i.e. the distance between target and distracter decreased by one unit. If the performance fell below 70% they leveled down i.e. the distance increased by one unit. The lowest possible level was a distance of 5 units, the highest possible level, a

distance of 1 unit. Halfway through the learning session, a break of 2 minutes was offered. Additionally, participants had the option to take a break after any block.

We expected that the participant would spend with every session a higher proportion of blocks at a higher difficulty level and finally, reach a stable performance at the most difficult level 1, an indication of over-learning. A learning progress index (LPI) was calculated in the following fashion: per session, we summed each block multiplied by its level. The LPI was high if the participant spent a large proportion of the session on level 5 or 4 and decreased as a higher proportion of blocks were completed at higher difficulty levels. The LPI was entered into a linear regression analysis.

Apparatus

The experiment was conducted in a dimly lit experimental room. Constant viewing distance was ensured by stabilizing the participant's head on a chin rest at a distance of 50 cm from a LED monitor (1920 × 1080 resolution, 60 Hz refresh rate, iiyama Proline, Japan). Auditory stimuli were presented through headphones (Sennheiser HD 555MR, Germany) at approximately 75 dB SPL. Experimental sessions were presented using Cogent 2000 v1.25 (developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON at the Wellcome Department of Imaging Neuroscience, UCL, London, UK; http://www.vislab.ucl.ac.uk/cogent.php) running under MATLAB (Mathworks Inc., Natick, MA, USA) on a Windows PC. The responses were given via a conventional keyboard.

fMRI

Stimuli

Visual stimuli were circles (line thickness: 0.7° visual angle) of 10 sizes (every fourth radius from mapping 2, 250 ms duration). Irrespective of size the circle's centre was fixed to the centre of the screen where the fixation cross was presented.

Auditory stimuli were 10 pure tones of variable pitch (every fourth frequency from mapping 2, 250 ms duration with linear onset and offset ramps of 10 ms to avoid auditory clicks; sampling rate 44100 Hz). The sound pressure level was adjusted to the loudest comfortable level for each participant at the beginning of the scanning session.

Main experimental design

In a 10 (stimulus size respectively pitch: 1-10) x 2 (modality: A or V) factorial design participants were presented with either unisensory visual circles of variable sizes or auditory pure tones of variable pitch in separate runs. 15% of the events were followed by catch trials consisting of an information mask (500ms) and another stimulus. The visual mask consisted of 10 circles sampled randomly from mapping 2 and presented in rapid succession (50 ms per stimulus) (Figure 5.5 B). The auditory mask contained 20 pure tones of various pitch, sampled uniformly from mapping 2 and presented in random order in rapid succession (25 ms per tone) (Figure 5.5 A). Participants were instructed to indicate via a two-choice key press as fast and accurately as possible whether the two stimuli temporarily

flanking the mask were the same or different. Throughout the entire experiment, participants were instructed to fixate a central cross.

The trial onset asynchrony was jittered between 2300 and 2700 ms. The order of stimuli was pseudo-randomized within each run. Further, 6 % of the events were ‘null events’ (i.e. fixation with no stimulus presentation).

The fMRI data was acquired on two separate days. On the first day, 8 auditory and 8 visual runs were acquired, on the second day 9 runs per modality. The order of auditory and visual runs was counterbalanced across the two days.

Experimental setup

Visual and auditory stimuli were presented using Cogent 2000 v1.25 (developed by the Cogent 2000 team at the FIL and the ICN and Cogent Graphics developed by John Romaya at the LON at the Wellcome Department of Imaging Neuroscience, UCL, London, UK; http://www.vislab.ucl.ac.uk/cogent.php) running on Matlab R2012a (MathWorks Inc.) on a Windows PC. The visual stimuli were back-projected onto a Plexiglass screen at the end of the scanner bore using a D-ILA projector (JVC DLA-SX21). The screen was visible to the participant via a mirror that was mounted on the MR head coil. The auditory stimuli were delivered via AVOTEC SS-3100 headphones (Avotec Inc.) at a maximum comfortable sound level which was established individually for each participant.

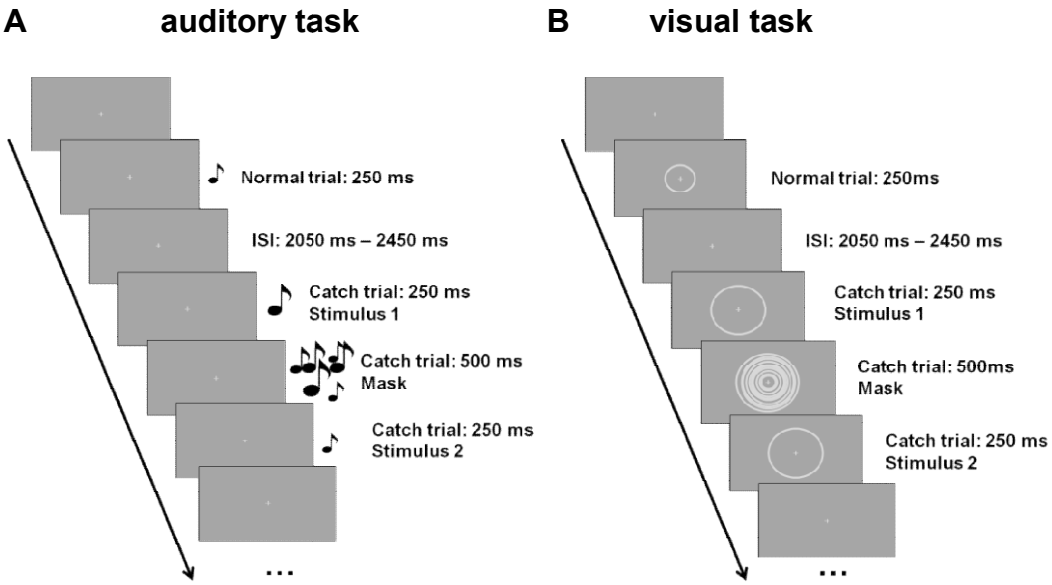


Figure 5.5 fMRI example trials. A. Auditory task. Pure tones (250 ms) were presented individually, interspersed by a jittered ISI (2050 - 2450 ms). 15% of trials were catch trials. After a pure tone (250 ms), an auditory information mask was presented (500 ms), followed by another pure tone (250 ms). B. Visual task. Circles (250 ms) presented individually, interspersed by a jittered ISI (2050 - 2450 ms). 15% of trials were catch trials. After a circle (250 ms), a visual information mask was presented (500 ms), followed by another circle (250 ms).

MRI data acquisition

The scanning sessions were conducted in a 3 Tesla Philips Achieva scanner with a 32-channel head coil at Birmingham University Imaging Centre. On a separate day, T1-weighted anatomical images (TR = 8.4 s, TE = 3.8 ms, TI = 540 ms, 175 slices, image matrix = 288 x 232, spatial resolution: 1 x 1 x 1 mm^3) were acquired, and in two subsequent sessions, T2*-weighted echo-planar images (EPI) (TR = 2.6 s, TE = 0.4 ms, 39 axial slices acquired in ascending order without gaps

covering the whole brain, image matrix = 80 x 80, spatial resolution: 3 x 3 x 3 mm^3). The first 4 scans of each run were acquired to allow for T1 saturation effects and discarded immediately. They were followed by 181 volumes. Each EPI run had a duration of 8.02 min.

fMRI analysis:

Pre-processing

The EPI images were pre-processed with Statistical Parametric Mapping (SPM8; Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm/; Friston, Holmes, Worsley, Frith & Frackowiak, 1995) running on Matlab R2012a. Scans from each participant were realigned to the first scan as reference and residual motion-related deformations were corrected using an unwarping-function. The time-series in each voxel were high-pass filtered to 1/128 Hz. The EPI images were analyzed in participant's native space. The high-resolution T1-weighted anatomical image was coregistered to the mean EPI image.

ROI definition

Auditory regions of interest (ROI) were defined based on the Brainnetome atlas (Fan et al., 2016) (http://atlas.brainnetome.org/). We defined two ROIs: (i) primary auditory cortex (PAC) including bilateral areas TE1.0 and TE1.2 on the Heschl's gyrus and (ii) STG including superior temporal gyrus (STG), bilateral area 41-42 (~ Planum Temporale), and the caudal and rostral area 22. Visual ROIs were defined

based on probabilistic retinotopic maps (Wang et al., 2015, http://scholar.princeton.edu/napl/resources0, using an 80% overlap threshold). We defined four visual ROIs: (i) prob-V1, (ii) combined prob-V2-V3, (iii) a ventral ROI combining hV4, VO1 and VO2, and (iv) a dorsal ROI consisting of V3A, V3B, IPS0, IPS1 and IPS2. The masks were first inverse-normalized from MNI standard space (Ashburner & Friston, 2005) into native space of the participant. Then the masks were resampled to 2 x 2 x 2 mm^3 voxels.

fMRI analysis

The data was modeled in an event-related fashion including one regressor for each of the 10 auditory and visual stimuli. The regressors were entered into a design matrix after convolving each event-related unit impulse (representing a single trial) with a canonical hemodynamic response function and its first temporal derivative. The realignment parameters were included as nuisance parameters in order to account for residual motion artifacts.

Support Vector Regression

We trained linear Support Vector Regression models (libSVM 3.20; Chang & Lin, 2011) as implemented in The Decoding Toolbox (Hebart, Görgen, Haynes & Dubois, 2015) to predict the stimulus labels within each of the six ROIs. First, we extracted response patterns for each voxel within the ROI from the parameter estimate image corresponding to the magnitude of the BOLD-response for each run and condition. The resulting parameter estimate images were then masked

with the corresponding binary ROI mask and pre-whitened runwise (Walther et al. 2012). The parameter estimate images for training and test data were normalized independently using euclidean normalization (Schrouff et al., 2013). Before training the SVR models, we standardized the stimulus parameters: first labels were sorted (pure tone frequency from low to high, visual circle radius from small to large) and then z-normalized. In a leave-one-run-out cross-validation procedure, the support vector regression models were trained to learn the mapping from condition-specific fMRI BOLD-response patterns to the 10 pure tones for the auditory runs or the 10 circles for the visual runs from all but one run. The SVR's parameters C and nu were standard fixed parameters (C = 1, nu = 0.5). The model then used this learned mapping to decode the stimulus codes from the voxel response patterns of the remaining (left-out) run. In a leave-one-run-out cross-validation scheme, the training-test procedure was repeated for all runs. For cross-modality decoding cross-validation was not necessary, all auditory runs were assigned to the training set and all visual runs to the test set (AV), and vice versa (VA).

Statistical Inference

To perform within-subject statistics we used the decoded labels as predictors for the true stimulus labels in general linear regression models computed separately for each auditory and each visual run. The run-specific parameter estimates were then entered into a one sample t-test.

Results

Psychophysics and learning

Learning Task

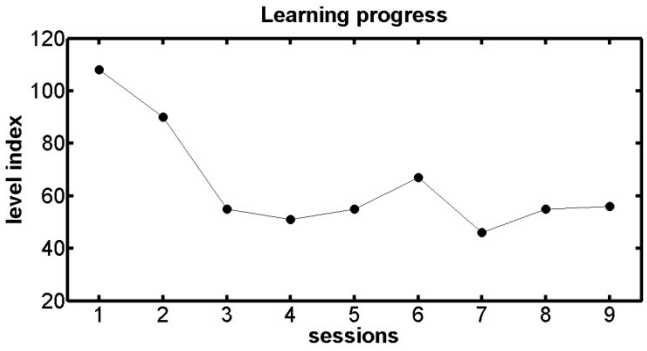
The participant learned in a computer game-like task to discriminate between distracters and audiovisual stimuli originating from a mapping that related high pitch with small size and low pitch with large size in a linear one-to-one fashion. The progress of learning was quantified as a Learning Progress Index (LPI: sum of blocks multiplied by level per session). A regression analysis on the LPI revealed that the participant has made significant progress in learning the mapping between auditory pitch and visual size (β = -5.25, p = .037).

The learning progress was fastest during the first 3 sessions. Level 1 was reached for the first time during session 7 and from then on every session (Figure 5.6 A). However, the participant did not reach a stable performance on the highest level and, therefore, has not achieved over-learning (Figure 5.6 B).

Adjustment Task

The adjustment task measured the participant's explicit association between the auditory and visual stimulus parameters as error distance between the participant's responses and the two mappings. One unit of error distance corresponded to a distance of one stimulus in stimulus space between the parameter that the participant had adjusted and the correct parameter of either mapping (Table 5.2). Before learning, the adjustment task indicated that the

A



B

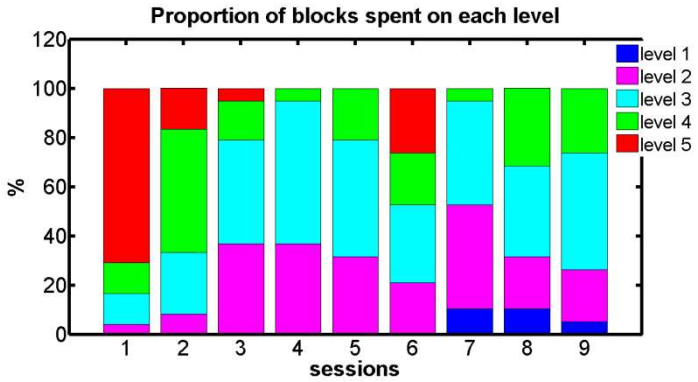
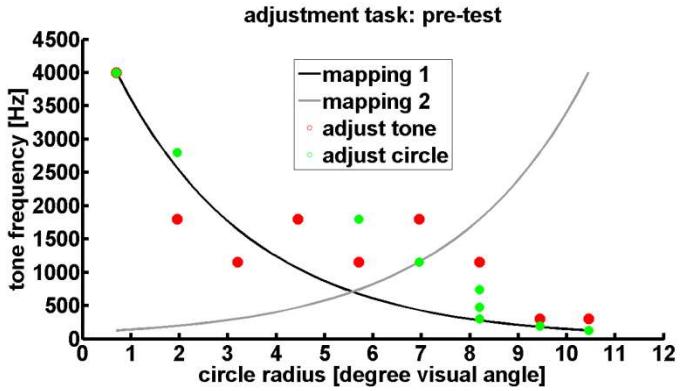


Figure 5.6 Results learning task. A. Learning progress Index (LPI: sum of blocks multiplied by level per session) across sessions. B. Learning progress as proportion of blocks spent on each level per session.

participant had a preference for mapping 1 (small size/high frequency, large size/low frequency), especially for stimulus parameters towards the extreme ends of the stimulus space (Figure 5.7 A). In contrast, after learning the results reflect the learned mapping 2 (small size/low frequency, large size/high frequency) (Figure 5.7 B). Taken together, these results show that the participant's explicit association between auditory pitch and visual size parameters was updated towards the learned mapping.

A



B

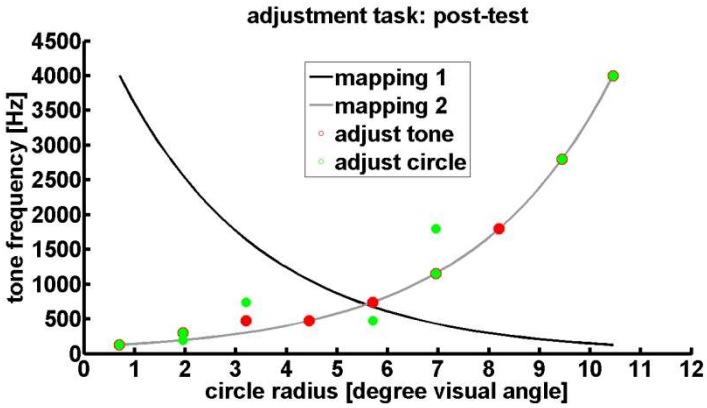


Figure 5.7 Results adjustment task. A. Before learning. B. After learning mapping 2.

Table 5.2 Results adjustment task. Error distance between adjusted parameter and parameters from mapping 1 and mapping 2 respectively, per test moment.

pre-test	mapping 1	mapping 2	post-test	mapping 1	mapping 2
adjust tone	33	13	adjust tone	2	38
adjust circle	39	7	adjust circle	5	41

Speeded Classification Task

The speeded classification task was implemented as an implicit measure of learning effects. We expected that after learning the participant would respond

faster and more accurately to audiovisual stimuli that were congruent with the learned mapping 2 compared to stimuli from anti-correlated mapping 1. The test was administered three times: before learning (pre-test), after session 5 (interim-test) and after learning (post-test). We performed 2 x (mapping: 1 vs. 2) two-sample t-tests for each test-moment, separately per modality-specific task, with the dependent variables: reaction times and accuracy (Table 5.3).

For reaction times, the response patterns before learning (pre-test), reflected a preference for mapping 1 (Figures 5.8 A and 5.8 B). After learning (post-test), the reaction times for the learned mapping 2 compared to reaction times to mapping 1 decreased in both the auditory and the visual task (Figures 5.8 A and 5.8 B), indicating a congruency effect with the learned mapping. However, none of these differences was statistically significant (Table 5.4). Furthermore, in the auditory task and to a lesser extend in the visual task there was an overall increase in reaction times from pre-test to post-test (Figures 5.8 A and 5.8 B). This is likely caused by fatigue because in contrast to the pre-test that was administered at the beginning of the first session the interim- and post-test were presented after the learning task, about 90 minutes after the beginning of the testing session.

Accuracy at post-test was higher for audiovisual stimuli that were congruent with mapping 2 compared to stimuli from mapping 1 (Figures 5.8 C and 5.8 D), yet, this difference was not statistically significant (Table 5.4). Only in the interim-test the accuracy was significantly higher for stimuli from mapping 1 (Table 5.4). However, this effect is weak and would not survive a correction for multiple comparisons.

In summary, the response pattern of both reaction times and accuracy at post-test were consistent with a congruency effect with the learning mapping but these effects were not statistically significant.

Table 5.3 Summary of mean reaction time (in ms) and accuracy (% correct) and their standard mean errors for the speeded classification task.

	mapping 1		mapping 2	
	reaction time	accuracy	reaction time	accuracy
auditory task				
pre-test	408.00 (2.45)	1.00 (0.00)	425.00 (2.64)	0.96 (0.00)
interim-test	511.50 (3.62)	0.96 (0.00)	502.00 (3.92)	0.92 (0.01)
post-test	574.00 (5.00)	0.81 (0.01)	562.00 (4.13)	0.93 (0.01)
visual task				
pre-test	414.00 (0.90)	0.98 (0.00)	424.00 (2.00)	0.98 (0.00)
interim-test	436.00 (2.31)	1.00 (0.00)	448.50 (2.30)	0.92 (0.01)
post-test	474.00 (2.66)	0.96 (0.00)	447.50 (2.16)	0.98 (0.00)

Table 5.4 Statistical results of the speeded classification task.

	reaction times		accuracy	
	auditory task	visual task	auditory task	visual task
pre-test	p = .656	p = .085	p = .156	p = 1.00
interim-test	p = .847	p = .812	p = .435	p = .044*
post-test	p = .701	p = .434	p = .109	p = .562

* $p < 0.05$.

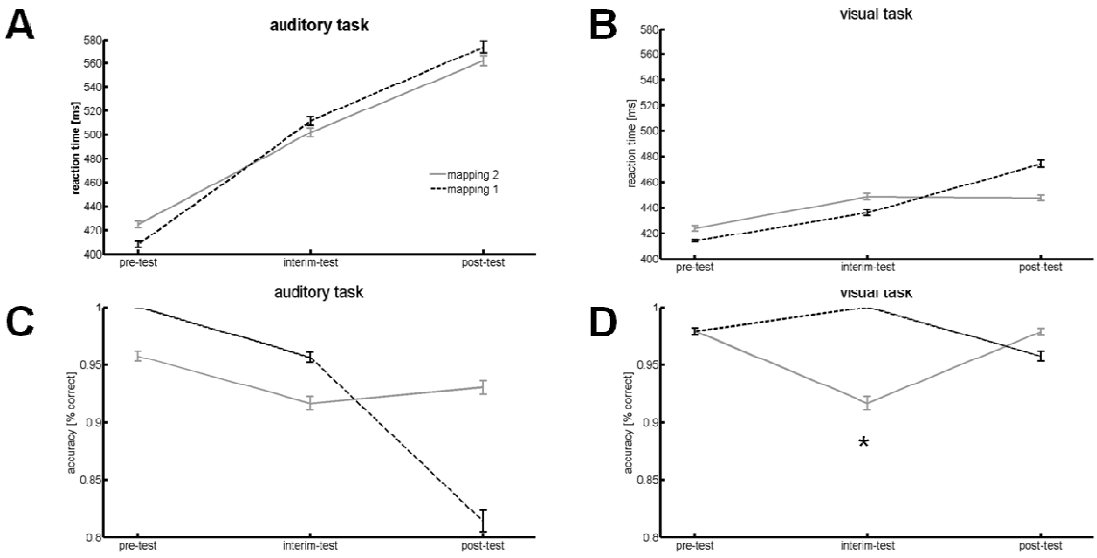


Figure 5.8 Results speeded classification task. A. Reaction times in auditory task. B. Reaction times in visual task. C. Accuracy in auditory task. D. Accuracy in visual task.

fMRI

Behavioural Task - Same-Different Task

We analyzed the responses of the catch trials as percentage of detected catch-trials for individual runs. Our participant consistently detected more than 85% of the catch trials (because of technical problems no data is available for run 1) (Table 5.5). A 2 (modality: auditory vs. visual) factor Wilcoxon signed-ranks test on the percentage of detected catch-trials did not reveal any significant difference in detection of visual and auditory catch trials (Z = 5, p = .375). Therefore, we conclude that our participant was awake and attended the stimuli throughout the whole experiment.

Table 5.5 fMRI - results behavioural task.

run number:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
% catch trials detected:	0	86	95	95	91	100	95	100	100	100	100	100	100	95	95	100	95	100

Multivariate Analysis

The participant was presented with auditory tones of variable frequencies and visual circles of variable radii in unimodal auditory and visual runs respectively.

First, we trained and tested support vector regression (SVR) models on BOLD-response patterns elicited by the same modality, separately for auditory and visual runs. We were able to decode circle size from BOLD-response patterns in all visual regions and pure tone frequency from BOLD-response patterns in auditory regions significantly but not vice versa (Figure 5.9 A).

Second, we tested for generalization effects across sensory modalities by decoding the tone frequency from BOLD-response patterns elicited by auditory tones after SVR models were trained on BOLD-response patterns elicited by visual circles and vice versa. In both primary sensory regions of interest, prob-V1 and PAC, we were able to significantly decode circle size from BOLD-response patterns after the model was trained on BOLD-response patterns for pure tone frequency, and pure tone frequency from BOLD-response patterns after the model was trained on BOLD-response patterns for circle size (Figure 5.9 B). Additionally, cross-modality decoding was significant in STG if auditory stimuli were decoded, and in prob-V2/V3 and dorsal ROI if visual stimuli were decoded.

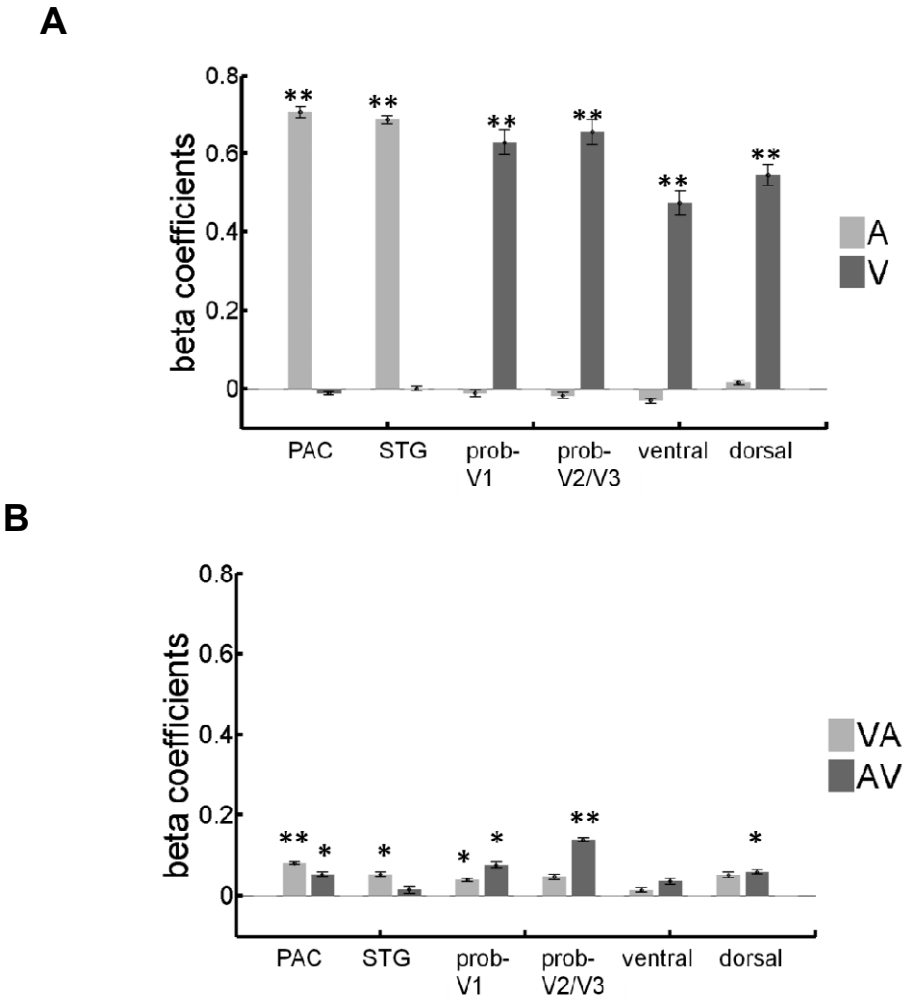


Figure 5.9 Results support vector regression. A. Within-modality decoding. The beta coefficients were significant for auditory stimuli in both auditory ROIs: PAC and STG. Beta coefficients for visual stimuli were significant in all of the visual ROIs: prob-V1, prob-V2/V3, ventral, dorsal. B. Cross-modality decoding. The beta coefficients were significant for auditory stimuli after the model was trained on BOLD-response patterns of visual stimuli (VA) in both auditory ROIs: PAC and STG, and prob-V1. In prob-V2/V3 and dorsal there was a trend. Beta coefficients for visual stimuli after the model was trained on BOLD-response patterns of auditory stimuli (AV) were significant in PAC, prob-V1, prob-V2/V3 and dorsal.

Table 5.6 Single-subject results within-modality SVR analysis. Mean beta values for auditory and visual stimuli per ROI.

	auditory stimuli		visual stimuli	
ROI	mean (std)	p-value	mean (std)	p-value
PAC	0.71 (0.13)	p < .001*	-0.01 (0.03)	p = .321
STG	0.69 (0.10)	p < .001*	0.00 (0.05)	p = .967
prob-V1	-0.01 (0.08)	p = .682	0.63 (0.28)	p < .001*
prob-V2/V3	-0.02 (0.07)	p = .474	0.66 (0.28)	p < .001*
ventral	-0.03 (0.06)	p = .168	0.47 (0.28)	p < .001*
dorsal	0.02 (0.05)	p = .384	0.55 (0.23)	p < .001*

* p < 0.05.

Table 5.7 Single-subject results cross-modality SVR analysis. Mean beta values for auditory and visual stimuli per ROI.

	train on visual stimuli, decode auditory stimuli (VA)		train on auditory stimuli, decode visual stimuli (AV)	
ROI	mean (std)	p-value	mean (std)	p-value
PAC	0.08 (0.03)	p < .001*	0.05 (0.06)	p = .026*
STG	0.05 (0.05)	p = .021*	0.01 (0.08)	p = .603
prob-V1	0.04 (0.04)	p = .016*	0.08 (0.07)	p = .012*
prob-V2/V3	0.05 (0.06)	p = .056	0.14 (0.05)	p < .001*
ventral	0.01 (0.05)	p = .475	0.03 (0.06)	p = .149
dorsal	0.05 (0.07)	p = .059	0.06 (0.05)	p = .005*

* p < 0.05.

Discussion

In this study, we trained a single participant a mapping that relates high pitch to large size and low pitch to small size (equivalent to mapping 2 in previous chapters). Over the run of 9 learning sessions the participant displayed significant learning progress, reaching the highest difficulty level on a number of blocks of the last three sessions, but did not achieve over-learning. Prior to training, the response patterns of reaction times in a speeded classification task and the

explicit matching of a few example stimulus parameters in an adjustment task pointed towards that the participant was associating high pitch with small size and low pitch with large size, the opposite of the to be learned mapping. After learning, both tasks showed remapping towards the learned mapping. Moreover, cross-modality decoding was significant in both primary auditory and visual cortices. Visual stimuli could be reliably decoded by a SVR model that was trained on BOLD-response patterns elicited by auditory stimuli, and vice versa. Decoding was also significant in the STG if the auditory response patterns were decoded, and in prob-V2-V3 and dorsal visual ROIs if visual response patterns were decoded, in either case there was also a strong trend for the opposite training scheme.

In the learning task, the participant showed rapid learning progress in the first three sessions and then a slower learning rate for the remaining 6 sessions. Only on the 7th of 9 sessions she reached the highest difficulty level. At the highest difficulty level the distracter stimulus was only one unit away from the correct mapping. Reaching that level means that the participant has almost mastered the one-to-one mapping. However, she remained there only for a couple of blocks per session and never reached a plateau of 100% accuracy. Unfortunately, we were not able to continue the learning experiment and find out if over-learning could be achieved at all. Overall, our results are consistent with previous learning studies: rapid learning of the overall structure followed by an increase in specificity over a longer period of time (Jeter et al., 2010). As far as we are aware of, this is the first study to have demonstrated successful learning of a nearly one-to-one mapping between auditory pitch and visual size in polar coordinates.

In the speeded classification task, the response patterns of reaction times suggested that learning caused a remapping of congruency from mapping 1 (high pitch/small size and low pitch/large size) to the learned mapping 2 (high pitch/large size and low pitch/small size). However, these results were not statistically significant. The task was kept short in order to avoid presenting participants with many incongruent stimuli and thereby demolish the learning effect. A learning effect acquired over a few sessions might not withstand influences from a mapping that a participant might have acquired throughout their life-time. However, given that we were looking for over-learning and hence expected a stable learning level, it should be considered to add more trials if this experiment is going to be repeated.

The adjustment task was incorporated in order to capture the participant's explicit mapping between the auditory and visual stimulus parameters before and after learning. Prior to training, no criteria for matching of the stimuli were provided. After training we expected it to reflect the learned mapping. Before training the participant intuitively matched the stimuli in a way that resembled mapping 1. After training the mapping reflected with a high accuracy the learned mapping 2. The matching results show a perfect fit at the extreme ends of the mapping and a few minor mistakes around the middle indicating that the lack of mapping accuracy resided in that area of the stimulus space. Taken together, the task confirms that learning has caused remapping of the participant's explicit pitch-size mapping and that the participant was able to reproduce the stimuli with a high accuracy.

The fMRI results after learning revealed varied results. Firstly, we demonstrated robust decoding of auditory stimuli from BOLD-response pattern it elicited in auditory ROIs and of visual stimuli from BOLD-response pattern it elicited in visual ROIs. Secondly, decoding of auditory stimuli from visual cortices and of visual stimuli from auditory cortices, after a SVR model was trained on BOLD-response patterns elicited by the same modality, was not successful. This is not entirely surprising given that in our previous study (Krugliak & Noppeney, in preparation), described in *Chapter 3*, there was high inter-subject variability in decoding of visual stimuli from auditory cortices and successful decoding of visual stimuli from auditory cortices was driven by consistent results across participants but it was significant at the single-subject level only in a few participants. Finally, we found positive results for cross-modality decoding. A SVR model was trained on BOLD-response patterns elicited by one modality and trained on BOLD-response patterns elicited by the other modality. Cross-modality decoding for both training schemes was significant in primary auditory and visual ROIs. Furthermore, significant decoding results were obtained in the STG if the auditory response patterns were decoded, and in prob-V2-V3 and dorsal visual ROIs if visual response patterns were decoded, the opposite training scheme produced a strong trend. Unfortunately, we do have neither pre-learning data nor data after learning the opposite mapping that could allow us to verify if these effects are in fact the result of learning. For future experiments, we recommend to reduce the number of conditions presented in the fMRI experiment in order to obtain more stable parameter estimates per condition. If indeed, our successful cross-modality

decoding results are caused by learning it would be consistent with the suggestion that cross-modality decoding can be induced by pairing stimuli from different modalities (Krugliak & Noppeney, in preparation).

Presenting the learning task as a computer game was a well received strategy. The participant commented that she was looking forward towards our experiment and felt highly motivated to level-up and stay at the highest difficulty level. Therefore, we recommend this approach for making perceptual learning experiments more engaging and reducing dropout rates as a result of fading motivation. For our experiment, preserving the circle size on the retina made it essential for participants to perform the task in the laboratory. For experiments without such constraints it might be worth considering to prepare apps or online games that participants can access from home in order to make long learning experiments more feasible and to reduce dropout rates.

Taken together, our findings point into the direction that we expected, namely that learning a one-to-one mapping of pitch and size is possible and that the effects were reflected in behavioral and neural measures. However, our results were not robust: (i) over-learning was not reached, (ii) the results of the speeded classification task were not significant, (iii) no comparison data was available to verify that successful cross-modality decoding was indeed a result of learning. All, in all our results indicate that the learning should have been continued until over-learning was reached - until the participant maintained level 1 over multiple sessions and the learned mapping was robustly detectable with independent behavioral measures.

CHAPTER 6: GENERAL DISCUSSION

The work presented in this thesis aimed at advancing the understanding of the interplay between the auditory and visual modalities, specifically between auditory pitch and visual size in polar coordinates, at the behavioural and neural level. In *Chapter 2*, we explored the metaphoric pitch-size mapping in the speeded classification task. In *Chapter 3*, we investigated the relationship between the neural representations of pitch and size within and across modalities. Finally, we attempted to induce an artificial pitch-size mapping at the behavioural (*Chapter 3 and Chapter 4*) and neural level (*Chapter 4*). In this chapter, I summarize the findings of the empirical chapters, describe how they contribute to explaining multisensory processing and provide directions for future research.

Overview of findings

Chapter 2: Synaesthetic interactions across vision and audition

The brain utilizes a variety of cues in order to decide which sensory information to bind and which to segregate. Besides the traditional amodal cues like temporal, spatial and semantic congruency also metaphoric mappings between seemingly arbitrary stimulus features have been shown to influence sensory integration. We investigated the metaphoric mapping between auditory pitch and visual size in a speeded classification task.

Our findings were converging in that participants responded faster if small size was presented simultaneously with low pitch and large size with high pitch. These results were surprising because they were consistent with the finding that rising

pitch was associated with growing size (Eitan, Schupak, Gotler & Marks, 2014) but revealed the opposite mapping compared with most studies that presented static visual discs or circles (Marks, Hammeal, Bornstein & Smith, 1987; Mondloch & Maurer, 2004; Gallace & Spence, 2006; Parise & Spence, 2009; Evans & Treisman, 2010; Bien, ten Oever, Goeber & Sack, 2012; Parise & Spence, 2012; Eitan et al., 2014; Tonelli, Cuturi & Gori, 2017; Brunetti, Indraccolo, Del Gatto, Spence & Santangelo, 2018; Ueda, Mizuguchi, Yakushijin & Ishiguchi, 2018). A series of control experiments and conditions confirmed that our findings could not simply be attributed to confounding effects of luminance of the visual stimuli or differences in perceptual loudness between the auditory stimuli. Further, to exclude that the relatively short stimulus onset asynchrony together with the instructions to fixate the centre of the visual stimuli might have created a dynamic context, we repeated the experiment with a longer stimulus asynchrony and removed the fixation cross together with all fixation instructions. Subsequently, we explored if the choice of specific stimulus parameters could explain the resulting response profiles. Still, the response profiles remained consistent even in conditions that manipulated the similarity and the absolute relationship between the stimulus parameters. We observed a reduced congruency between similar pitch and size parameters compared to more distinct parameters. This is consistent with findings that interference is stronger if the difference between congruent and incongruent stimuli is larger (Rothen & Meier, 2014). Interestingly, the mapping remained the same irrespective if a pitch/size parameter was presented as low/small option in one condition and as high/large option in a

different condition, indicating that it is the relative relationship and not the absolute values of pitch and size that are driving this phenomenon, at least for the stimulus parameters chosen in this study. This finding was recently confirmed in a different study (Brunetti et al., 2018).

Chapter 3: The neural basis of the relationship between auditory pitch and visual size in polar coordinates

Cross-modal influences in the neocortex occur already at the primary cortical level. The content of stimuli presented in one sensory modality, like muted videos depicting actions or the spatial location of objects, can be decoded from BOLD-response patterns in other sensory areas (Meyer et al., 2010; Man, Kaplan, Damasio & Meyer, 2012; Liang, Mouraux & Iannetti, 2013; de Haas, Schwarzkopf, Urner & Rees, 2013; Vetter, Smith & Muckli, 2014; Petro, Paton & Muckli, 2017). According to our knowledge, so far only stimuli that share amodal properties e.g. are spatially or semantically related have been decoded. Here we utilized support vector regression (SVR) to address the question if auditory and visual sensory cortices carry information about the modality-specific stimulus features auditory pitch and visual size in polar coordinates not only about the preferred but also about the non-preferred sensory modality.

After presenting pure tones of various frequencies and circles of various sizes (centred around fixation) unimodally in separate runs, we were able to successfully decode not only auditory pitch from auditory cortices and visual size from visual cortices but also visual size from auditory cortices. Significant decoding of visual

stimuli in auditory cortices but not vice versa is consistent with previous studies (Meyer et al., 2010; Man et al., 2012). Further findings conforming with previous literature were that decoding results in the non-corresponding modality were substantially weaker and less stable than in the preferred modality (Meyer et al., 2010; Man et al., 2012; de Haas et al., 2013; Liang et al., 2013; Vetter et al., 2014; Petro et al., 2017), cross-modality decoding was not successful at all. So far cross-modality decoding has only been shown for stimuli that were either paired prior to being presented unimodally or had naturally a strong semantic relationship (Meyer et al., 2010; Man et al., 2012).

Chapter 4: Seeing pitch and hearing size

Natural environmental statistics play a key role in shaping our perception. Frequent co-occurrence of specific stimulus features has been proposed to result in metaphoric mappings between seemingly arbitrary stimulus features, like pitch-brightness or pitch-elevation (Spence, 2011). We attempted to artificially induce a one-to-one mapping between auditory pitch and visual size in polar coordinates in an active audiovisual perceptual learning experiment.

On average, participants displayed a significant learning effect but have not learned the pitch-size mapping sufficiently either to reproduce the one-to-one relationship between the auditory and visual stimulus parameters nor to evoke a congruency effect in a speeded classification task. We identified several methodological shortcomings in this experiment. During learning participants were presented with a higher proportion of incongruent than congruent stimuli, making

the learning task unnecessary difficult and inefficient. Moreover, in the final blocks, we expected to test the stability of learning and removed any source of feedback. Unfortunately, we thereby eliminated the presentation of congruent stimuli altogether. These last blocks seem to have caused confusion, resulting in a drop of performance. We cannot exclude that this issue obliterated the learning effect and prevented capturing it in the subsequent speeded classification task. Further suggestions for improvements of the task included extending the experiment to multiple sessions in order to allow participants sufficient time for learning the specific mapping and to make the task more engaging.

Chapter 5: An attempt to induce a synaesthesia-inspired pitch-size mapping

After the first attempt to artificially induce a one-to-one pitch-size mapping, as reported in *Chapter 4*, we revised the learning task and the testing procedure. In this chapter, we described a case study that combined learning with a revised version of the functional Magnetic Resonance Imaging (fMRI) experiment as described in *Chapter 3*. Before undergoing the fMRI experiment, one participant was extensively trained in a specifically designed computer game to map high pitch on small size and low pitch on large size in a linear fashion.

The participant displayed significant learning progress and achieved at times almost a one-to-one mapping between the given pitch and size parameters. An adjustment task that tested the participant's explicit mapping and the response patterns of reaction times in a speeded classification task revealed that learning caused remapping of the participant's original size-pitch mapping towards the

learned mapping. Furthermore, after learning not only decoding within the preferred sensory modalities was significant but also cross-modality decoding. A SVR model was trained on BOLD-responses elicited by auditory stimuli and trained on BOLD-response patterns elicited by visual stimuli, and vice versa. Both training schemes yielded significant results in the primary auditory and primary visual ROIs. Furthermore, in STG, prob-V2/V3, and the dorsal ROI, decoding was significant if the preferred modality was in the test set. The opposite training scheme produces strong trends in prob-V2/V3 and the dorsal ROI. Unfortunately, external circumstances interrupted the experiment before we could train the participant on the anti-correlated mapping. Consequently, we do not have fMRI data either from a pre-learning baseline or after learning the anti-correlated mapping that would allow us to compare the response patterns before and after learning, or after learning two anti-correlated mappings, and, therefore, are not able to confirm that the cross-modality decoding results are indeed related to learning.

Contributions and future directions

In *Chapter 2*, we investigated if visual size presented as circles centered around fixation i.e. size in polar coordinates presented together with auditory pitch in the speeded classification task could replicate the general finding in static context in which high pitch is associated with small size and low pitch with large size (Gallace & Spence, 2006; Evans & Treisman, 2010; Eitan et al., 2014) or dynamic context in which the opposite pattern is predicted (Eitan et al., 2014). Our results

showed consistently a response pattern in accordance with the dynamic context e.g. high pitch was associated with large size and low pitch with small size. This led us to conclude that our stimuli seem to be more likely interpreted not as objects of different sizes but rather as the same object at a different distance from the observer. This is not entirely surprising since the main motivation for using visual size in polar coordinates was to exploit the fact that such stimuli also can be used to map the visual retinotopic dimension of eccentricity in visual areas. Importantly, to our knowledge, this is the first study that systematically investigated how manipulations of size and pitch parameters affect the pitch-size mapping in the speeded classification task. Surprisingly, none of the manipulations reversed the response pattern, highlighting the importance of context over absolute parameters. It would be interesting to repeat our experiment with an additional condition of explicitly dynamic context in order to directly contrast the consequences of our stimulus choice both in a static and a dynamic context. Furthermore, we demonstrated that the mapping is more likely to be relative than absolute. This finding was confirmed in a recent study (Brunetti et al., 2018). Taken together, our results suggest that the pitch-size mapping is not robust, it reflects the relative rather than the absolute relationship between pitch and size, and the direction of this mapping can be affected by context. Therefore, we recommend future studies to pay special attention to how instructions are given in order to ensure that the expected context is studied.

In *Chapters 3 and 5*, we explored the relationship between auditory pitch and visual size in polar coordinates at the level of neural representations in fMRI. We

used SVR to decode pitch and size from BOLD-response patterns in auditory and visual cortices. First of all, we demonstrated highly reliable decoding of very similar stimuli: auditory pitch in auditory cortices and visual size in polar coordinates in visual cortices. This is the first study that presents successful decoding of highly similar, topographically specific auditory and visual stimuli in regions of interests in the auditory and visual cortices using SVR. For the following analysis of our data, these results mean that our stimuli were eliciting response patterns that were sufficiently different from each other to allow optimal decoding. Next, in *Chapter 3*, we showed that it is possible to decode visual size in polar coordinates from primary auditory cortices, and we also reported a trend for STG. This means that auditory areas contain information about visual eccentricity that is stable enough to produce response patterns that allow reliable decoding. This is the first study to successfully decode visual eccentricity from auditory cortices of normally hearing participants and the first study to decode highly similar stimuli using SVR (see for decoding of eccentricity in auditory cortices of congenitally deaf: Almeida et al., 2015). However, we were not able to identify a clear mapping underlying these patterns. Our stimuli were designed for taking advantage of the sensitivity of multivariate pattern analysis but by doing that we compromised the representability of our results with univariate methods. Further multivariate analyses of our data, like multi-class support vector classification (SVC) and representational dissimilarity analyses (RSA) might provide insights into the relationship between individual stimulus parameters. These analyses were still in progress at the time of the submission deadline of this thesis and could therefore

not be included. For future studies into the relationship between local topographic maps and the spatial representation of auditory pitch and visual size in polar coordinates, I recommend using fewer, more salient stimulus exemplars in order to evoke more robust BOLD-response patterns than those that we obtained in our experiments. An example of a more efficient design could look as follows: three categories: small size/low pitch, medium size/pitch and large size/high pitch, each with three similar stimulus examples per sensory modality. Such a design would still be covering a wide range of the pitch and size stimulus space and at the same time allow studying highly similar stimulus exemplars within each category. It would be interesting to follow-up: (i) if pitch and size indeed produce topographically specific activations in non-corresponding cortices, for example like the auditory pitch representations that Watkins and colleagues (2013) reported in V5 of congenitally blind participants, (ii) trace the spatial location of the voxels that contributed most to reliable decoding, as demonstrated elegantly by Liang and colleagues (2013).

Interestingly, we were not able to reliably decode auditory stimuli from visual cortices. This asymmetric decoding pattern is largely consistent with the literature that decoded spatially or semantically related auditory and visual stimuli from non-corresponding cortices (Meyer et al., 2010; Man et al., 2012). Interestingly, in the study that did demonstrate significant decoding of visual stimuli in auditory cortices, participants were blindfolded (Vetter et al., 2014). Taken together with reports of activations and even topographic representations of auditory stimuli in visual cortices of congenitally blind participants (Watkins et al., 2013), it is

tempting to speculate that response patterns that are elicited by auditory stimuli in visual cortices are simply overwritten by the dominance of visual information, except under condition of visual deprivation. This seems not to be the case; the experiment by Vetter and colleagues was replicated under an eyes-open condition in which a black screen was presented (Petro et al., 2017). However, the classification results were lower than with blindfolding. The authors speculated that decoding results might not reach significance in the presence of more demanding visual stimulation. It is for future studies to investigate and to control for the balance of the saliency of auditory influences of visual cortices with demands on visual processing, in order to further unravel the auditory influences in visual cortices using multivariate pattern analysis.

Furthermore, cross-modality decoding of unrelated pitch and visual size in polar coordinates was not successful either (*Chapter 3*). This is not entirely surprising because so far cross-modality decoding has only been shown for stimuli that were either paired prior to being presented unimodally or had naturally a strong semantic relationship (Meyer et al., 2010; Man et al., 2012). Taken together with the finding that the effect of congruency between audiovisual stimuli seems to primarily affect the reliability of the BOLD-response pattern elicited by a stimulus but not the similarity between patterns (De Haas et al., 2013), this fits neatly into the framework that the role of cross-modality interactions at early cortical levels contributes mostly to the manipulation of saliency and detection of stimuli (e.g. Noppeney, Jones, Rohe & Ferrari, 2018). In *Chapter 5*, cross-modality decoding was successful after one participant was extensively trained on a one-to-one

mapping of pitch and visual size in polar coordinates. Auditory pitch could be decoded from auditory and visual primary cortices after the SVR model was trained on BOLD-response patterns elicited by visual stimuli. Visual size in polar coordinates could be decoded from visual cortices and the primary auditory cortices. Unfortunately, the experiment was interrupted and we were not able to obtain a data set of the pre-learning baseline or after training the participant on an anti-correlated mapping to verify if these findings indeed reflected learning. However, these results are interesting in that they suggest that we might have successfully induced a cross-modal mapping between the neural representations of our auditory pitch and visual size in polar coordinates. It remains for future studies to verify if and how cross-modality decoding can be influenced through learning procedures and if consequently topographically specific co-activations can be revealed in auditory and visual cortices.

In *Chapters 4 and 5*, we took advantage of the fact that the pitch-size mapping appeared to be not quite robust and trained participants in an active audiovisual learning paradigm to reproduce one of two artificially constructed linear one-to-one mappings between auditory pitch and visual size in polar coordinates. First of all, we have demonstrated that participants were able to learn a highly specific mapping between auditory pitch and visual size in polar coordinates. Interestingly, participants (*Chapter 4*) learned both kinds of mappings equally well, irrespective if that mapping related high pitch/small size and low pitch/large size or high pitch/large size and low pitch/small size. This supports our general finding that the pitch-size mapping is not particularly robust (Krugliak & Noppeney, 2016; Krugliak

& Noppeney, in preparation). It proved difficult to induce congruency effects reflecting learning in independent behavioural measures like the speeded classification task. In *Chapter 4*, the lack of significant congruency effects can be explained by methodological shortcomings of the learning task and the experimental procedure. However, in *Chapter 5*, we revised the learning procedure. We noted a clear remapping in an explicit adjustment task but still did not find a significant congruency effect in the speeded classification task, even though after about 9 hours (9 days) of learning the response patterns of reaction times and accuracy already reflected congruency with the learned mapping. Previous studies that attempted to induce synaesthesia used even longer training regimes than those that we used and finally succeeded to capture effects congruent with the learning task in synaesthetic Stroop and contextual priming tasks (Howells, 1944; Rothen, Wantz & Meier, 2011; Colizoli, Murre & Rouw, 2012; Bor, Rothen, Schwartzman, Clayton & Seth, 2014). The learning procedure in *Chapter 5* was interrupted before over-learning was reached. Therefore, we cannot tell at this point if further learning would have led to the aspired effects. It would be interesting to see this experiment completed in the future as a within-subject design which trains participants on both mappings and compares the behavioural and neural patterns after learning.

Our learning experiments demonstrate the complexity of attempting to replicate learning processes that under natural circumstances take years to develop. Remapping of sensory processes during sensory deprivation occurs relatively rapidly but it proves more difficult to achieve profound sensory remapping effects

in healthy participants (e.g. Proulx, Brown, Pasqualotto & Meijer, 2014). Considering that decoding of auditory stimuli from visual cortices produced higher accuracies when participants were blindfolded (Vetter et al., 2014) compared to when they were viewing a blank screen (Petro et al., 2017), does this mean that to study cross-modal influences with fMRI it is necessary to significantly restrict stimulation of the preferred modality in order to reveal influences of non-preferred sensory modalities? The data obtained from fMRI, with its low temporal resolution and voxels that cover large populations of neurons, reflects a summary of a multitude of neural and metabolic processes. This complicates detection of subthreshold influences and processes that are known to be mediated by relatively sparse projections (Falchier, Clavagnier, Barone & Kennedy, 2002; Rockland & Ojima, 2003). Therefore, manipulating the sensory stimulation in order to reveal cross-modal influences in non-corresponding cortices might provide a solution. Importantly, this is a different approach than adding noise to reduce the informativeness of a sensory modality in order to increase the contribution of multisensory enhancement (e.g. Klemen & Chambers, 2012; Noppeney et al., 2018). The first approach allows looking at cross-modal influences while the preferred sensory modality is ‘muted’, the latter approach, on the other hand, reveals the interplay of different sensory modalities in improving extraction of information from an unreliable source.

The question about how to match stimuli in different sensory modalities in studies of multisensory processing brings us to a general problem in multisensory research: how to design tasks and select stimuli in a way that optimises the

comparison between different sensory modalities. In our learning experiments, we assumed equal discrimination abilities between visual size and auditory pitch. However, while most observers are very accurate in the visual modality, only very few have perfect pitch. It would be interesting to repeat our pilot experiment with participants who have absolute pitch.

Conclusions

The experiments presented in this thesis investigated and manipulated the interplay between auditory pitch and visual size in polar coordinates. We have shown that at the behavioural level the pitch-size mapping is relative, depending on the context in which it is presented and how it is interpreted by the observer. At the neural level, we have revealed that auditory cortices contain information about visual size in polar coordinates (eccentricity) and that learning a specific mapping between auditory pitch and visual size in polar coordinates could have potentially contributed to the generalization of auditory pitch into representations of visual size on polar coordinates and vice versa, both in primary auditory and primary visual cortices. Furthermore, we have demonstrated that approximate learning of a specific pitch-size mapping can be achieved within a 90-minutes session and achieving a replication of an almost one-to-one mapping requires a longer learning period.

References

Adam, R., & Noppeney, U. (2010). Prior auditory information shapes visual category-selectivity in ventral occipito-temporal cortex. *NeuroImage*, *52*(4), 1592–1602.

Allman, B. L. (2009). Not just for bimodal neurons anymore: The contribution of unimodal neurons to cortical multisensory processing. *Brain Topography*, *21*(3–4), 157–167.

Allman, B. L., Bittencourt-Navarrete, R. E., Keniston, L. P., Medina, A. E., Wang, M. Y., & Meredith, M. A. (2008). Do cross-modal projections always result in multisensory integration? *Cerebral Cortex*, *18*(9), 2066–2076.

Allman, B. L., & Meredith, M. A. (2007). Multisensory Processing in "Unimodal" Neurons: Cross-Modal Subthreshold Auditory Effects in Cat Extrastriate Visual Cortex. *Journal of Neurophysiology*, *98*(1), 545–549.

Almeida, J., He, D., Chen, Q., Mahon, B. Z., Zhang, F., Gonçalves, Ó. F., Bi, Y. (2015). Decoding Visual Location From Neural Patterns in the Auditory Cortex of the Congenitally Deaf. *Psychological Science*, *26*(11), 1771–1782.

Antovic, M. (2009). Musical Metaphors in Serbian and Romani Children: An Empirical Study. *Metaphor and Symbol*, *24*(3), 184–202.

Arnold, G., & Auvray, M. (2018). Tactile recognition of visual stimuli: Specificity versus generalization of perceptual learning. *Vision Research, 152,* 40-50.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851.

Ashley, R. (2004). Musical pitch space across modalities: spatial and other mappings through language and culture. *Proceedings of the 8th International Conference on Music Perception & Cognition*, 64–71.

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. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *26*(47), 12260–12265.

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: role of stimulus difference. *Perception & Psychophysics*, *57*(8), 1151–1162.

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

Bien, N., ten Oever, S., Goebel, R., & Sack, A. T. (2012). The sound of size: crossmodal binding in pitch-size synesthesia: a combined TMS, EEG and psychophysics study. *NeuroImage*, *59*(1), 663–672.

Bor, D., Rothen, N., Schwartzman, D. J., Clayton, S., & Seth, A. K. (2014). Adults Can Be Trained to Acquire Synesthetic Experiences. *Scientific Reports*, *4*, 7089.

Brang, D., Rouw, R., Ramachandran, V. S., & Coulson, S. (2011). Similarly shaped letters evoke similar colors in grapheme-color synesthesia. *Neuropsychologia*, *49*(5), 1355–1358.

Brang, D., Towle, V. L., Suzuki, S., Hillyard, S. A., Di Tusa, S., Dai, Z., Grabowecky, M. (2015). Peripheral sounds rapidly activate visual cortex:

evidence from electrocorticography. *Journal of Neurophysiology*, *114*(5), 3023–3028.
Brunetti, R., Indraccolo, A., Del Gatto, C., Spence, C., & Santangelo, V. (2018). Are crossmodal correspondences relative or absolute? Sequential effects on speeded classification. *Attention, Perception, and Psychophysics*, *80*(2), 527–534.
Butler, A., & James, K. (2012). Active Learning of Novel Sound-producing Objects: Motor Reactivation and Enhancement of Visuo-motor Connectivity. *Journal of Cognitive Neuroscience, 25*(2), 203-218.
Da Costa, S., Van der Zwaag, W., Marques, J. P., Frackowiak, R. S. J., Clarke, S., & Saenz, M. (2011). Human primary auditory cortex follows the shape of Heschl's gyrus. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *31*(40), 14067–14075.
Cappe, C., Rouiller, E. M., & Barone, P. (2009). Multisensory anatomical pathways. *Hearing Research*, *258*(1–2), 28–36.
Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27.
Cohen Kadosh, R., Sagiv, N., Linden, D. E. J., Robertson, L. C., Elinger, G., and Henik, A. (2005). When blue is larger than red: colors influence numerical cognition in synesthesia. *J. Cogn. Neurosci, 17*, 1766–1773.
Colizoli, O., Murre, J. M. J., and Rouw, R. (2012). Pseudo-synesthesia through reading books with colored letters. *PLoS ONE, 7*(6), 1-10.
De Haas, B., Schwarzkopf, D. S., Urner, M., & Rees, G. (2013). Auditory modulation of visual stimulus encoding in human retinotopic cortex. *NeuroImage*, *70*, 258–267.
Deroy, O., Spence, C., 2013. Why we are not all synesthetes (not even weakly so). *Psychon. Bull. Rev. 20,* 643–664.
Donohue, S. E., Roberts, K. C., Grent-'t-Jong, T., & Woldorff, M. G. (2011). The cross-modal spread of attention reveals differential constraints for the temporal and spatial linking of visual and auditory stimulus events. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(22), 7982–7990.
Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on "sensory- sensory-specific" brain regions, neural responses, and judgments. *Neuron*, *57*(1), 11–23.
Eitan, Z., Schupak, A., Gotler, A., & Marks, L. E. (2014). Lower pitch is larger, yet falling pitches shrink. *Experimental Psychology*, *61*(4), 273–284.
Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch, *Journal of Vision, 7,* 1–14.
Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features, *Journal of Vision, 10*, 1–12.
Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *22*(13), 5749–5759.
Falchier, A., Schroeder, C. E., Hackett, T. A., Lakatos, P., Nascimento-Silva, S., Ulbert, I., Smiley, J. F. (2010). Projection from visual areas V2 and prostriata to caudal auditory cortex in the monkey. *Cerebral Cortex*, *20*(7), 1529–1538.

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox., P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cerebral Cortex*, *26*(8), 3508–3526.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierachical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47.

Fernández-Prieto, I., Navarra, J., & Pons, F. (2015). How big is this sound? Crossmodal association between pitch and size in infants. *Infant Behavior and Development*, *38*, 77–81.

Formisano, E., Kim, D. S., Di Salle, F., Van de Moortele, P. F., Ugurbil, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron*, *40*(4), 859–869.

Friston, K. J., Holmes, A. P., Worsley, K. J., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, *2*, 189–210.

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

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, *10*(6), 278–85.

Grinband, J., Hirsch, J., & Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron*, *49*(5), 757–763.

Haryu, E., & Kajikawa, S. (2012). Are higher-frequency sounds brighter in color and smaller in size? Auditory-visual correspondences in 10-month-old infants. *Infant Behavior and Development*, *35*(4), 727–732.

Hebart, M. N., Görgen, K., Haynes, J.-D., & Dubois, J. (2015). The Decoding Toolbox 121 (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*(January), 1–18.

Howells, T. H. (1944). The experimental development of color-tone synesthesia. J. Exp. Psychol. *34*, 87–103.

Hussain, Z., Bennett, P. J., & Sekuler, A. B. (2012). Versatile perceptual learning of textures after variable exposures. *Vision Research*, *61*, 89–94.

Ibrahim, L. A., Mesik, L., Ji, X. Y., Fang, Q., Li, H. F., Li, Y. T., Zingg, B., Zhang, L. I., & Tao, H. W. (2016). Cross-modality sharpening of visual cortical processing through layer-1-mediated inhibition and disinhibition. *Neuron, 89 (5),* 1031–1045.

Iurilli, G., Ghezzi, D., Olcese, U., Lassi, G., Nazzaro, C., Tonini, R., Tucci, V., Bonfenati, F., & Medini, P. (2012). Sound-driven synaptic inhibition in primary visual cortex. *Neuron, 73 (4),* 814–828.

Jeter, P. E., Dosher, B. A., Liu, S. H., & Lu, Z. L. (2010). Specificity of perceptual learning increases with increased training. *Vision Research*, *50*(19), 1928–1940.

Karunanayaka, P. R., Wilson, D. A., Vasavada, M., Wang, J., Martinez, B., Tobia, M. J., Yang, Q. X. (2015). Rapidly acquired multisensory association in the olfactory cortex. *Brain and Behavior*, *5*(11), 1-14.

Kayser, C., Logothetis, N. K., & Panzeri, S. (2010). Visual Enhancement of the Information Representation in Auditory Cortex. *Curr. Biol. 20 (1),* 19–24.

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex, 18 (7),* 1560–1574.
Kayser, C., Petkov, C. I., & Logothetis, N. K. (2009). Multisensory interactions in primate auditory cortex: fMRI and electrophysiology. *Hearing Research*, *258(2009),* 80–88.
Kelly, E. L. (1934). An experimental attempt to produce artificial chromaesthesia by the technique of the conditioned response. *J. Exp. Psychol. 17,* 315–341.
Kim, K. H., & Iwamiya, S. I. (2008). Formal Congruency between Telop Patterns and Sound Effects. *Music Perception*, *25*(5), 429-448.
Klemen, J., & Chambers, C. D. (2012). Current perspectives and methods in studying neural mechanisms of multisensory interactions. *Neuroscience and Biobehavioral Reviews*, *36*(1), 111–133.
Krugliak, A., & Noppeney, U. (2016). Synaesthetic interactions across vision and audition. *Neuropsychologia*, *88*, 65–73.
Kusnir, F., and Thut, G. (2012). Formation of automatic letter–colour associations in non-synaesthetes through likelihood manipulation of letter–colour pairings. *Neuropsychologia, 50,* 3641–3652.
Laurienti, P. J., Burdette, J. H., Wallace, M. T., Yen, Y.-F., Field, A. S., & Stein, B. E. (2002). Deactivation of sensory-specific cortex by cross-modal stimuli. *Journal of Cognitive Neuroscience*, *14*, 420–429.
Laurienti, P. J., Kraft, R. a., Maldjian, J. a., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, *158*, 405–414.
Lee, H., & Noppeney, U. (2011). PNAS Plus: Long-term music training tunes how the brain temporally binds signals from multiple senses. *Proceedings of the National Academy of Sciences*, *108*(51), 1441–1450.
Lee, H., & Noppeney, U. (2014). Music expertise shapes audiovisual temporal integration windows for speech, sinewave speech, and music. *Frontiers in Psychology*, *5*(August), 1–9.
Leitão, J., Thielscher, A., Werner, S., Pohmann, R., & Noppeney, U. (2013). Effects of parietal TMS on visual and auditory processing at the primary cortical level-a concurrent TMS-fMRI study. *Cerebral Cortex*, *23*(4), 873–884.
Liang, M., Mouraux, A., Hu, L., & Iannetti, G. D. S. (2013). Primary sensory cortices contain distinguishable spatial patterns of activity for each sense. *Nature Communications*, *4*, 1979.
Lewis, R., & Noppeney, U. (2010). Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(37), 12329–12339.
Macaluso, E., & Driver, J. (2005). Multisensory spatial interactions: A window onto functional integration in the human brain. *Trends in Neurosciences*, *28*(5), 264–271.
Man, K., Kaplan, J. T., Damasio, A., & Meyer, K. (2012). Sight and sound converge to form modality-invariant representations in temporoparietal cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *32*(47), 16629–36.

Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 384–394.
Marks, L. E., Hammeal, R. J., Bornstein, M. H., & Smith, L. B. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development*, *52*(1, Whole No. 215), 1-102.
Marks, L. E. (2004). Cross-modal interactions in speeded classification. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 85–105). Cambridge: MIT Press.
Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, *28*(7), 903–923.
Martuzzi, R., Murray, M. M., Michel, C. M., Thiran, J. P., Maeder, P. P., Clarke, S., & Meuli, R. A. (2007). Multisensory interactions within human primary cortices revealed by BOLD dynamics. *Cerebral Cortex, 17 (7),* 1672–1679.
McIntosh, A. R. M. C., Cabeza, R. E., & Lobaugh, N. J. (1998). Analysis of Neural Interactions Explains the Activation of Occipital Cortex by an Auditory Stimulus. *J Neurophysiol, (80),* 2790-2796.
Meier, B., and Rothen, N. (2009). Training grapheme-colour associations produces a synaesthetic Stroop effect, but not a conditioned synaesthetic response. *Neuropsychologia, 47,* 1208–1211.
Melara, R.D., & O'Brien, T.P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General, 116,* 323–336.
Meyer, M., Baumann, S., Marchina, S., & Jancke, L. (2007). Hemodynamic responses in human multisensory and auditory association cortex to purely visual stimulation. *BMC neuroscience*, *8*, 14.
Meyer, K., Kaplan, J. T., Essex, R., Webber, C., Damasio, H., & Damasio, A. (2010). Predicting visual stimuli on the basis of activity in auditory cortices. *Nature Neuroscience*, *13*(6), 667–668.
Michel, M., & Jacobs, R. (2007). Parameter learning but not structure learning: A Bayesian network model of constraints on early perceptual learning. *Journal of Vision*, *7*, 1–18
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.
Niccolai, V., Wascher, E., & Stoerig, P. (2012). Distinct neural processes in grapheme–colour synaesthetes and semantic controls*. Eur. J. Neurosci. 36*, 3593–3601.
Noppeney, U., Jones, S. A., Rohe, T., & Ferrari, A. (2018). See what you hear-How the brain forms representations across the senses. *Neuroforum*, *24*(4), 237–246.
Nunn, J. A., Gregory, L. J., Brammer, M., Williams, S. C. R., Parslow, D. M., Morgan, M. J., et al. (2002). Functional magnetic resonance imaging of synesthesia: activation of V4/V8 by spoken words. *Nat. Neurosci. 5,* 371–375.
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. (2012). Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Experimental Brain Research*, *220*(3-4), 319–333.
Parise, C. V, Knorre, K., & Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 6104–6108.
Patching, G. R., & Quinlan, P. T. (2002). Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology. Human Perception and Performance*, *28*(4), 755–775.
Petro, L. S., Paton, A. T., & Muckli, L. (2017). Contextual modulation of primary visual cortex by auditory signals. *Philos.Trans. R. Soc. B*. 372. 20160104.
Proulx, M. J., Brown, D J., Pasqualotto, A., & Meijer, P. (2014) Multisensory perceptual learning and sensory substitution. *Neuroscience & Biobehavioral Reviews*, *41*, 16-25.
Rich, a. N., Bradshaw, J. L., & Mattingley, J. B. (2005). A systematic, large-scale study of synaesthesia: Implications for the role of early experience in lexical-colour associations. *Cognition*, *98*(1), 53–84.
Rockland, K. S., & Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *Int. J. Psycho- physiol. 50 (1–2),* 19–26.
Rothen, N., Nikoli´c, D.,Jürgens,U.M., Mroczko-Wa˛sowicz, A., Cock, J., & Meier, B. (2013). Psychophysiological evidence for the genuineness of swimming-style colour synaesthesia. *Conscious. Cogn. 22,* 35–46.
Rothen, N., & Meier, B. (2014). Acquiring synaesthesia: insights from training studies. *Frontiers in Human Neuroscience*, *8*(March), 109.
Rothen, N., Wantz, A.-L., & Meier, B. (2011). Training synaesthesia*. Perception 40,* 1248–1250.
Sadaghiani, S., Maier, J. X., & Noppeney, U. (2009). Natural, metaphoric, and linguistic auditory direction signals have distinct influences on visual motion processing. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(20), 6490–6499.
Schrouff, J., Rosa, M. J., Rondina, J. M., Marquand, A. F., Chu, C., Ashburner, J., & Mourão-Miranda, J. (2013). PRoNTo: pattern recognition for neuroimaging toolbox. *Neuroinformatics*, *11*(3), 319–337.
Simner, J., Harrold, J., Creed, H., Monro, L., & Foulkes, L. (2009). Early detection of markers for synaesthesia in childhood populations. *Brain, 132,* 57–64.
Spence, C. (2011). Crossmodal correspondences: a tutorial review. *Attention, Perception & Psychophysics*, *73*(4), 971–995.
Spence, C., & Deroy, O. (2013). How automatic are crossmodal correspondences? *Consciousness and Cognition*, *22*(1), 245–260.
Stokes, M., Thompson, R., Cusack, R., & Duncan, J. (2009). Top-Down Activation of Shape-Specific Population Codes in Visual Cortex during Mental Imagery. *J Neurosci. 29*(5), 1565–1572.
Suzuki, Y., & Takeshima, H. (2004). Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America*, *116*(2), 918–933.
Takeshima, Y., & Gyoba, J. (2012). High-intensity sound increases the size of visually perceived objects. *Attention, Perception & Psychophysics*, 501–507.
Tanabe, H. C., Honda, M., & Sadato, N. (2005). Functionally segregated neural substrates for arbitrary audiovisual paired-association learning. *The Journal of*

Neuroscience : the official journal of the Society for Neuroscience, *25*(27), 6409–6418.
Teramoto, W., Hidaka, S., & Sugita, Y. (2010). Sounds move a static visual object. *PloS one*, *5*(8), 1-5.
Tonelli, A., Cuturi, L. F., & Gori, M. (2017). The influence of auditory information on visual size adaptation. *Frontiers in Neuroscience*, *11*(OCT), 1–8.
Ueda, S., Mizuguchi, A., Yakushijin, R., & Ishiguchi, A. (2018). Effects of the Simultaneous Presentation of Corresponding Auditory and Visual Stimuli on Size Variance Perception. *I-Perception*, *9*(6), 204166951881570.
Van Atteveldt, N., Formisano, E., Goebel, R., & Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron*, *43*, 271–282.
Van Atteveldt, N. M., Formisano, E., Blomert, L., & Goebel, R. (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cerebral Cortex*, *17*(April), 962–974.
Vetter, P., Smith, F. W., & Muckli, L. (2014). Decoding Sound and Imagery Content in Early Visual Cortex. *Current Biology*, *24*(11), 1256–1262.
Von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Ives, D. T., & Griffiths, T. D. (2007). Neural representation of auditory size in the human voice and in sounds from other resonant sources. *Current Biology : CB*, *17*(13), 1123–1128.
Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception & Psychophsics*, *72*(4), 871-884.
Wallace, M. T. Wilkinson, L. K., & Stein, B. E. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. *Journal of Neurophysiology, 76(2),* 1246-1266.
Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research*, *158*, 252–258.
Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188–200.
Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron*, *56*(2), 366–383.
Wang, L., Mruczek, R. E. B., Arcaro, M. J., & Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, *25*(10), 3911–3931.
Watkins, K. E., Shakespeare, T. J., O'Donoghue, M. C., Alexander, I., Ragge, N., Cowey, A., & Bridge, H. (2013). Early Auditory Processing in Area V5/MT+ of the Congenitally Blind Brain. *Journal of Neuroscience*, *33*(46), 18242–18246.
Zangenehpour, S., & Zatorre, R. J. (2010). Crossmodal recruitment of primary visual cortex following brief exposure to bimodal audiovisual stimuli. *Neuropsychologia*, *48*(2), 591–600.