THE INTERPLAY BETWEEN ATTENTION AND MULTISENSORY INTEGRATION

Ambra Ferrari

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

September 2019

ABSTRACT

The world taxes our attentional resources with a constant influx of multisensory inputs. This raises the critical question of whether and how attention and multisensory integration interact to guide behaviour. Previous research has led to contrasting perspectives: while some investigations state that attention is a prerequisite for multisensory integration, others provide evidence of fast and automatic multisensory interactions which, instead, orient attention. The present thesis reconciles this artificial dichotomy by providing behavioural and neural evidence of a synergistic interplay between attention and multisensory integration at multiple levels of processing. Such flexible cooperation serves a common computational goal: to promote perceptual scene analysis adjusting for environmental conditions (competition for processing resources, sensory noise) and task demands (detection, discrimination). Specifically, here I show that multisensory integration captures attention in the presence of competing streams of information; moreover, attention modulates sensory uncertainty and determines selective read-out of internal task-relevant representations. Within a Bayesian framework, I further discuss how prior knowledge participates in this mutual interplay. Collectively, the emerging evidence of a tight functional interconnection between attention, multisensory integration and predictive processes provides a promising framework for characterising the development and flexible adjustment of effective behaviour in our complex and dynamic world.

ACKNOWLEDGEMENTS

None of the accomplishments of the past four years would have been possible without the help and support of many people.

Firstly, I would like to thank my supervisor prof. Uta Noppeney for her guidance throughout this journey. Your determination and dedication have pushed me out of my comfort zone and helped me grow both scientifically and personally. Moreover, I thank dr. Karin Petrini and dr. Wieske van Zoest for kindly agreeing to examine this thesis.

Thank you to the Computational Cognitive Neuroimaging lab for the mutual support and the lively scientific discussions. I have discovered many important aspects of science and life in our special gang of nerds and weirdoes.

A big thank you to my friends across England and Italy for popping the bubble of lab life and filling me with laughter and good food.

Finally, thank you to Alessandro and our families for their unconditional love. You are the key ingredient behind my smile.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

A1-2	Early auditory cortices	**MNI**	Montreal Neurological Institute
aIPS	Anterior intraparietal sulcus	**MR**	Magnetic resonance
ANOVA	Analysis of variance	**nW_{AV}**	Neural audio-visual weight index
BCI	Bayesian Causal Inference	**pIPS**	Posterior intraparietal sulcus
BIC	Bayesian information criterion	**PT**	Planum temporale
BMS	Bayesian model selection	**RMSE**	Root-mean-square error
BOLD	Blood-oxygenation-level-dependent	**ROI**	Region of interest
d'	d-prime	**RT**	Response time
EPI	Echo planar imaging	**SNR**	Signal-to-noise ratio
fMRI	Functional magnetic resonance imaging	**SPL**	Sound pressure level
FOV	Field of view	**SPM**	Statistical parametric mapping
FWE	Family-wise error	**STD**	Signal detection theory
FWHM	Full width at half maximum	**SVR**	Support vector regression
GLM	General linear model	**TE**	Echo time
HRF	Hemodynamic response function	**TR**	Repetition time
HRTF	Head-related transfer function	**V1-3**	Early visual cortices
MLE	Maximum likelihood estimation	**W_{AV}**	Audio-visual weight index

LIST OF MANUSCRIPTS AND ABSTRACTS

The following manuscripts form Chapters 3, 4, 5 and 6 of the present thesis:

Ferrari, A. & Noppeney, U. (in preparation). Attention modulates sensory reliability and impacts response selection during multisensory perceptual inference.

Ferrari, A. & Noppeney, U. (in preparation). Attention impacts multisensory perceptual inference via distinct computational principles along the cortical hierarchies.

Ferrari, A., Degano, G. & Noppeney, U. (in preparation). Cross-modal binding captures attention within a cocktail-party scenario.

Ferrari, A., Degano, G. & Noppeney, U. (in preparation). Cross-modal binding in auditory cortex recruits the attention network within a cocktail-party scenario.

The following published paper has been cited in the present thesis:

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. http://doi.org/10.1515/nf-2017-A066.

The following published conference abstracts relate to studies included in the present thesis:

Ferrari, A. & Noppeney, U. (2019). Attention influences how the brain integrates audiovisual signals into spatial representations. Organization for Human Brain Mapping, Rome, Italy.

Ferrari, A. & Noppeney, U. (2017). Endogenous modality-specific attention influences sensory reliability during multisensory integration. Ten Years of Mind/Brain Sciences at the University of Trento, CIMeC, Rovereto, Italy.

Ferrari, A. & Noppeney, U. (2017). The role of endogenous modality-specific attention in multisensory integration. International Multisensory Research Forum, Nashville, TN, USA.

Ferrari, A. & Noppeney, U. (2017). The role of endogenous modality-specific attention in multisensory integration. Festival of Neuroscience, British Neuroscience Association, Birmingham, UK

CHAPTER 1

GENERAL INTRODUCTION

Any number of impressions, from any number of sensory sources, falling simultaneously on a mind which has not yet experienced them separately, will fuse into a single undivided object for that mind. The law is that all things fuse that can fuse, and nothing separates except what must.

Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence.

–William James, The Principles of Psychology, 1890

The world is inherently multisensory (Soto-Faraco et al., 2019) and taxes our attentional resources with a constant influx of sensory inputs (Peelen & Kastner, 2014). Hence, it is conceivable that multisensory processes and attention interact in everyday situations. Indeed, many laboratory investigations have concentrated on their relationship and have produced an abundance of results (for reviews: Koelewijn et al., 2010; Macaluso et al., 2016; Talsma et al., 2010; Tang et al., 2016). However, they have also led to contrasting perspectives (Koelewijn et al., 2010): on the one hand, attention seems to be a prerequisite for the integration of inputs from the different senses; on the other hand, there is evidence of fast and automatic multisensory interactions which, instead, orient attentional resources. The present thesis aims to reconcile this artificial dichotomy by providing behavioural and neural evidence of a

synergistic interplay between attention and multisensory integration at multiple levels of processing. Such interplay serves a common computational goal: to guide adaptive behaviour in our complex world flexibly adjusting for environmental conditions and task demands.

The following chapter introduces the background literature that sustains and motivates the present empirical work. Firstly, I present key definitions and principles of multisensory integration and attention. Next, I outline the current debate about their relationship. Finally, I provide an overview of the thesis to highlight its scope and structure.

1.1 Definition and principles of multisensory integration

Multisensory integration refers to the simultaneous processing of information across different sensory modalities, which produces neural and behavioural effects distinct from the output of separate unisensory processing (Stein, 2012). At the neural level, multisensory integration is expressed in terms of response non-linearity such as superadditivity, where the response to multisensory stimuli is greater than the sum of responses to the respective unisensory components (Stein & Stanford, 2008). This implies the presence of neural populations that are sensitive to multisensory interactions, instead of mere convergence of activations from independent unisensory neural populations (James & Stevenson, 2012; Noppeney, 2012). Seminal electrophysiological recordings in cat superior colliculus have exploited superadditivity to characterise the fundamental principles of interaction across the senses. The *temporal rule* states that the strength of multisensory integration is directly proportional to the degree of temporal proximity of unisensory signals (i.e. maximal integration for synchronous stimuli; Meredith et al., 1987). The *spatial rule* postulates that the strength of multisensory integration is directly proportional to the degree of spatial proximity of unisensory signals (i.e. maximal integration for co-located stimuli; Meredith & Stein, 1986). The *principle of*

inverse effectiveness states that multisensory integration is maximal when the individual unisensory components can only weakly generate the corresponding response (Meredith & Stein, 1983). At the behavioural level, the integration of multisensory inputs determines the enhancement of perceptual salience, which is critical for stimuli detection (Stein & Stanford, 2008). Crucially, the same principles outlined above also apply to behavioural performance: when near-threshold unisensory inputs are presented close in space and time, they increase detection accuracy and perceived signal strength. For instance, visual detection (Frassinetti et al., 2002a; Gleiss & Kayser, 2013; Noesselt et al., 2008, 2010) is enhanced by simultaneous auditory inputs; vice-versa, auditory loudness (Odgaard et al., 2004) and detection of speech and non-speech sounds (Eramudugolla et al., 2011; Lovelace et al., 2003) are boosted by simultaneous visual stimuli. Similarly, auditory detection and perceived auditory loudness are enhanced by congruent tactile stimulation (Gillmeister & Eimer, 2007).

When it comes to constructing task-specific representations that guide categorisation and discrimination responses, the *modality-appropriateness hypothesis* (Welch & Warren, 1980) captures the important principle that different sensory modalities are given different weights depending on the stimulus characteristic under evaluation. In particular, the more precise a sensory modality is in a specific task domain, the more it dominates the construction of the perceptual representation. For example, vision is normally the most precise sense in the spatial domain (Freides, 1974) and thus it dominates perception in spatial tasks. This is exemplified by the *spatial ventriloquist effect* (Jack & Thurlow, 1973; for review: Chen & Vroomen, 2013), where the perceived location of an auditory signal is biased toward the position of a synchronous yet spatially discordant visual signal.

The computational approach of maximum likelihood estimation (MLE) has formally explained how redundant information across the senses (e.g. information about spatial

location) are combined into unified perceptual representations: observers average unisensory components weighting each of them in accordance with their relative reliabilities (i.e. *reliability-weighted integration*; Alais & Burr, 2004; Ernst & Banks, 2002; Ernst & Bülthoff, 2004). The less variable or noisy (i.e. the more reliable) a unisensory input is, the greater is its weight and the more it drives the final multisensory output, in line with the modality-appropriateness hypothesis. Moreover, if both unisensory inputs are very noisy, the variance of the optimally integrated percept is lower than both the unisensory variances, in line with the inverse effectiveness principle.

Critically, the MLE model postulates mandatory integration of unisensory inputs into a unified percept even for slight (unnoticed) cross-modal discrepancies (Helbig & Ernst, 2007; Vatakis & Spence, 2007), based on the assumption of a common cause (i.e. *unity assumption*, Welch & Warren, 1980; for review: Chen & Spence, 2017). However, in real-life conditions we constantly receive inputs from different sources. In order to form a coherent representation of the surrounding environment, it is then fundamental to integrate inputs from a common source and segregate those from separate sources. An extension of the MLE model, known as Bayesian Causal Inference (BCI), formally explains how observers solve the cross-modal binding (or causal inference) problem taking into consideration both causal uncertainty and sensory noise (Körding et al., 2007). By allowing flexible arbitration between integration and segregation, the BCI model has proven the best at describing behavioural performance in multisensory contexts (Körding et al., 2007; Shams & Beierholm, 2010). Following a Bayesian perspective, the BCI model introduces priors capturing observers' a priori binding tendency (i.e. strength of unity assumption) and perceptual biases (e.g. *central bias* for spatial

localisation, Odegaard et al., 2015)[1]. Importantly, the BCI model accounts for the temporal and spatial rules of integration, by showing that decreased temporal and spatial consistencies across the senses lower the probability of perceptual integration. This translates in decreased cross-modal illusions, such as the spatial ventriloquist effect, for increasingly incongruent stimuli (Rohe et al., 2019; Rohe & Noppeney, 2015a, 2015b, 2016). In order to arbitrate between integration and segregation and consequently form a perceptual representation (i.e. *multisensory perceptual inference*), observers exploit a vast set of cues that includes not only temporal synchrony (Lee & Noppeney, 2011a; Lewis & Noppeney, 2010; Magnotti et al., 2013; Maier et al., 2011; Munhall et al., 1996; Noesselt et al., 2007; Parise & Ernst, 2016; Parise et al., 2012; van Wassenhove et al., 2007) and spatial congruence (Lewald & Guski, 2003; Slutsky & Recanzone, 2001; Spence, 2013), but also semantic correspondences (Adam & Noppeney, 2010; Bishop & Miller, 2011; Kanaya & Yokosawa, 2011; Lee & Noppeney, 2011b; Noppeney et al., 2010) and synesthetic correspondences (Parise & Spence, 2009; Sadaghiani et al., 2009; Spence, 2011). The (in)congruence of such cues is supposed to manipulate the strength of the prior binding tendency (Chen & Vroomen, 2013; Chen & Spence, 2017; Parise, 2015; Spence, 2011). Key factors are the number of redundant congruent properties (e.g. spatial co-location) and the learned associations between non-redundant congruent properties (e.g. small size and high pitch; Deroy & Spence, 2016). Within this perspective, *cross-modal binding* (i.e. the automatic grouping of coherent cross-modal features into a unified object or event, Bizley et al., 2016) belongs to multisensory perceptual inference as it represents one of its possible outcomes. Specifically, cross-modal binding corresponds to the integration of multisensory inputs (via selection of one common cause) based on prior binding knowledge and congruent perceptual information. It has been

[1] See Chapter 2 for an extensive description of the BCI model, which will be employed in Chapters 3 and 4.

proposed that in complex everyday situations, the automatic binding of naturalistic cross-modal signals (e.g. speech) is supported by their temporal coherence (Atilgan et al., 2018; Bizley et al., 2016; Maddox et al., 2015; Noppeney & Lee, 2018; Shamma et al., 2011). A classic example consists in the integration of the voice and moving lips of a speaker, which aids selective listening in multi-talker conditions (i.e. *cocktail-party scenarios*; Cherry, 1953).

1.2 Definition and principles of attention

The natural environment provides us with a constant influx of multiple inputs, which generate competition for our limited processing capacities. Attention is the adaptive cognitive function that resolves such competition creating a processing bias in favour of a subset of the available information (Desimone & Duncan, 1995). While the general computational goal of attention is clear, its characterisation is complex and multifaceted and speaks against the presence of a unitary process (Chun et al., 2011; Nobre & Kastner, 2014; Serences & Kastner, 2014). More specifically, the term attention embraces a collection of mechanisms responsible for the selection and modulation of sensory or representational information (Chun et al., 2011). Selection refers to the process of biasing competition among concurrent information; modulation refers to changes of the selected representation, such that attended relative to unattended representations show enhanced signal-to-noise ratio (SNR). In this way, attention determines the formation of *priority maps*, which guide behaviour based not only on perceptual salience but also on behavioural relevance (Bisley & Goldberg, 2010; Serences & Yantis, 2006). Electrophysiological and neuroimaging investigations have characterised distinct neural mechanisms through which attention forms priority maps: signal enhancement boosts the strength of attended representations; external noise suppression inhibits distractors; internal noise reduction boosts the reliability of attended representations (Reynolds & Heeger,

2009; Serences & Kastner, 2014). Furthermore, attention can modify how stored representations inform perceptual decisions independent of the above sensory modulations. This is accomplished via selective read-out of information from pools of neurons that are optimally tuned to discriminate the attended feature (Pestilli et al., 2011; Serences & Kastner, 2014; Sprague et al., 2018). In other words, selective read-out mechanisms do not change the quality (i.e. SNR) of sensory representations; instead they control how these representations are selected or disregarded. This is accomplished by increasing the weight of neural signals associated with attended stimuli and efficiently shunting interference from sensory neurons that encode irrelevant information. Abundant evidence shows that the neural mechanisms of attention act upon representations in sensory cortex (Buschman & Kastner, 2015; Serences & Kastner, 2014) and are controlled by a distributed network encompassing superior and inferior frontal regions, posterior parietal and temporo-parietal regions, thalamic and midbrain regions (Corbetta et al., 2008; Corbetta & Shulman, 2002; Santangelo et al., 2009; Santangelo, 2018; Serences et al., 2004; Shomstein & Yantis, 2006). As a result, behavioural performance is improved for attended relative to unattended information in terms of faster reaction and response times (e.g. Coull & Nobre, 1998; Donohue et al., 2015; Spence et al., 2001), enhanced target detection (e.g. Theeuwes & Chen, 2005; Theeuwes et al., 2004) and improved perceptual discrimination (Anton-Erxleben & Carrasco, 2013; Carrasco, 2011). Crucially, the ability to selectively attend to specific information while ignoring concurrent distractors interacts with task difficulty, as explained by load theory (Lavie, 2005, 2010): engaging in a difficult task absorbs our limited processing capacities and decreases distractors' influence; vice-versa, engaging in an easy task frees residual processing resources that can spill over to distractors and thus enhance their influence on performance.

Further underscoring its complex and multifaceted nature, attention shows multiple characteristics that can be flexibly combined according to environmental conditions and the task at hand (Chun et al., 2011). Attention can concentrate on one specific item or multiple items at a time (i.e. selective or divided attention; Pashler, 1998); it is possible to hold attention on a specific element, disengage and orient attention on a different element, or alternate attention among them (Corbetta et al., 2008; Posner et al., 1980); moreover, attention can move with eye movements or eye fixation (i.e. overt or covert attention; Juan et al., 2004). A proposed taxonomy of attention that is embraced by the present thesis highlights the distinction between internal and external attention, namely the selection and modulation of sensory information on the one hand and internally generated representations on the other hand (Chun et al., 2011). Depending on the medium, external attention can be further classified into modality-specific (i.e. attend to stimuli in one sensory modality, Spence et al., 2001), spatial (i.e. attend to stimuli in a specific portion of space, Carrasco, 2011), temporal (i.e. attend to specific time points and structures that unfold in time, Coull & Nobre, 1998; Nobre & Van Ede, 2018), feature-based (i.e. attend to one feature across objects, Carrasco, 2011) or object-based (i.e. attend to all the features within the same object, Chen, 2012; Marinato & Baldauf, 2019). Importantly, attention can spread across sensory modalities, e.g. when multisensory inputs share the same attended location (cross-modal spread of spatial attention, Driver & Spence, 1998) or pertain to the same attended object or event (cross-modal spread of object-based attention, Busse et al., 2005). Internal attention, instead, operates over representations stored in memory and is responsible for response selection during task execution (Chun et al., 2011). Hence, internal attention is inherently associated with selective read-out (Pestilli et al., 2011; Serences & Kastner, 2014), which represents the efficient selection of internal representations from pools of neurons that encode task-relevant

information. On the other hand, only external attention can be by definition exogenous and stimulus-driven, namely it can be automatically captured by specific environmental features (Chun et al., 2011). For instance, objects (i.e. perceptual units resulting from Gestalt organisation) capture attention (Humphreys & Riddoch, 2003; Kimchi et al., 2007; Yeshurun et al., 2009). Instead, both external and internal attention can be endogenous and goal-directed, namely they can be intentionally controlled by the individual to perform the task at hand (Chun et al., 2011).

Importantly, the present thesis deliberately avoids a sharp dichotomy between top-down and bottom-up definitions, whose plausibility has recently been re-discussed (Awh et al., 2012; Macaluso & Doricchi, 2013; Theeuwes, 2018). This is motivated by the tight functional connection between attention and *selection history* biases (Theeuwes, 2018), as exemplified by priming of pop-out (e.g. Theeuwes & van der Burg, 2011), statistical learning (e.g. Zuanazzi & Noppeney, 2018) and value-driven attentional capture (e.g. Anderson et al., 2011). More generally, the present work sustains no sharp dichotomy between top-down and bottom-up cognitive processes by embracing a Bayesian perspective on multisensory integration (Körding et al., 2007; Shams & Beierholm, 2010) and the closely related predictive coding framework (Friston, 2010; Feldman & Friston, 2010). Accordingly, the terms *endogenous attention* and *attentional capture* will be used to characterise the degree of intentionality of the attentional process at hand: while the former will indicate intentional control of attention based on task instructions, the latter will refer to non-intentional orienting of attention following salient environmental features.

1.3 The interplay between attention and multisensory integration

The extent to which attention and multisensory integration interact at the behavioural and neural levels has direct implications for the conceptualisation of multisensory processes themselves. Across several years of research, three alternative frameworks have been proposed (Koelewijn et al., 2010). The late integration framework (Figure 1.1A) states that selective attention affects the analysis of unisensory inputs and subsequently supports their integration into a unified percept, similar to the Feature Integration Theory (Treisman & Gelade, 1980) in unisensory contexts. In other words, the late integration framework postulates that attention is a necessary prerequisite for multisensory integration. Such perspective derives from the traditional view that early sensory areas are functionally specialised for the processing of unisensory stimuli in the respective sensory modality, whereas multisensory integration is deferred to higher-order association areas (Calvert & Thesen, 2004). In accordance, early multisensory studies with primates (Avillac et al., 2007; Barraclough et al., 2005; Bruce et al., 1981; Desimone & Ungerleider, 1986; Linden et al., 1999; Watanabe, 1992) and humans (Beauchamp et al., 2004; Bremmer et al., 2001; Calvert et al., 2000; Macaluso et al., 2003) showed the recruitment of temporo-parietal or prefrontal cortices, which are also implicated in attentional control (Corbetta et al., 2008; Corbetta & Shulman, 2002; Santangelo et al., 2009; Santangelo, 2018; Serences et al., 2004; Shomstein & Yantis, 2006). Even recently, after the emergence of pervasive multisensory interactions across the neo-cortex (for review: Ghazanfar & Schroeder, 2006), neuroimaging studies keep showing the key role of attention in shaping neural responses to multisensory stimuli (Fairhall & Macaluso, 2009; Morís Fernández et al., 2015; Rohe & Noppeney, 2016, 2018; Talsma et al., 2007; Talsma & Woldorff, 2005). Consistently, abundant behavioural evidence sustains

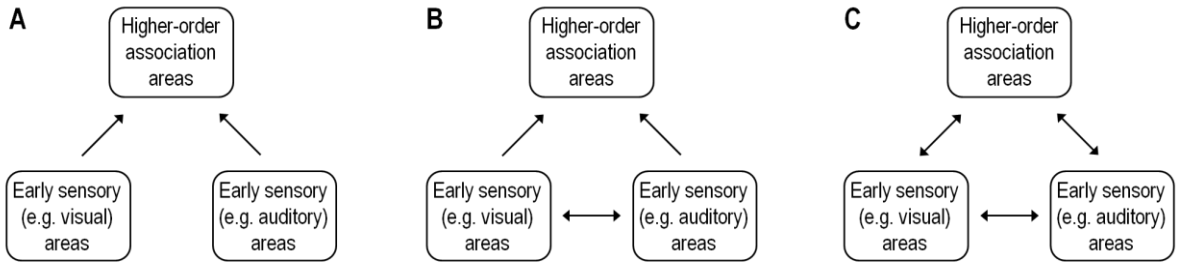

Figure 1.1: Multisensory integration frameworks

Schematic representation of A) late integration framework, B) early integration framework and C) parallel integration framework. Adapted from Koelewijn et al., 2010.

the influence of attention on multisensory perceptual inference, as exemplified by studies that manipulated selective modality-specific attention (Odegaard et al., 2016; Vercillo & Gori, 2015), selective spatial attention (Donohue et al., 2015; Van der Stoep et al., 2015) and attentional load via dual-task conditions (Alsius et al., 2005, 2007; Michail & Keil, 2018). In particular, it is conceivable that attention boosts the SNR of attended versus unattended sensory information and thereby impacts reliability-weighted integration (Odegaard et al., 2016; Rohe & Noppeney, 2018; Vercillo & Gori, 2015). Moreover, task relevance determines the selective read-out of internal multisensory representations to instruct final localisation (Rohe & Noppeney, 2016) and categorisation (Cao et al., 2019) responses in accordance with Bayesian Causal Inference (see Section 2.2.1). Finally, it has been demonstrated that even prior experience (Gau & Noppeney, 2016; Nahorna et al., 2012) and training (Ernst, 2007; Jicol et al., 2018; Lee & Noppeney, 2011a; Love et al., 2012; Petrini et al., 2009; Petrini et al., 2011) modulate the strength of integration, further sustaining the view that multisensory interactions are pervaded by top-down cognitive processes.

Conversely, the early integration framework (Figure 1.1B) states that sensory inputs interact at an early sensory level and subsequently support attentional orienting via the

recruitment of association areas. In other words, the early integration framework postulates that multisensory integration is a pre-attentive mechanism that can in turn impact attention. Accordingly, it appears that multisensory congruence generates a salience-driven selection bias (Desimone & Duncan, 1995) when perceivers need to detect information under high attentional competition. In particular, spatial cueing studies have shown that under dual task conditions only bisensory task-irrelevant cues capture spatial attention, whereas in no-load conditions both unisensory and bisensory task-irrelevant cues are equally effective (Ho et al., 2009; Santangelo et al., 2008; Santangelo & Spence, 2007). Similarly, task-irrelevant auditory onsets boost the ability of visual onsets to produce a "pop-out" effect for visual targets in the midst of concurrent distractors, as indexed by increased efficiency during visual search tasks (Matusz & Eimer, 2011; Van der Burg et al., 2008). Moreover, visual detection in the unattended hemi-space of unilateral spatial neglect patients is enhanced by simultaneous and co-located task-irrelevant sounds (Frassinetti et al., 2002b). Finally, lip-reading enhances selective listening in cluttered environments (Cherry, 1953; see also Bernstein et al., 2004; Grant & Seitz, 2002; Sumby & Pollack, 1954), both in the presence of high sensory noise (Crosse et al., 2016; Ross et al., 2007) and multiple competing speakers (Helfer & Freyman, 2005; Zion Golumbic et al., 2013). Hence, it is conceivable that cross-modal binding (in this context, grouping of temporally coherent voice and lip movements) biases the allocation of selective attention to promote scene analysis in complex environments (Maddox et al., 2015; Zion Golumbic et al., 2013). Further enhancing the controversy with the late integration framework, there is also evidence that spatial attention does not impact audio-visual integration during spatial localisation (Bertelson et al., 2000; Vroomen et al., 2001) and modality-specific attention does not impact visuo-tactile integration during size discrimination (Helbig & Ernst, 2008). At the neural level, the early integration framework is

supported by ample evidence of cross-modal interactions between early sensory areas (for reviews: Driver & Noesselt, 2008; Foxe & Schroeder, 2005; Ghazanfar & Schroeder, 2006; Kayser & Logothetis, 2007; Schroeder & Foxe, 2005). Following the seminal work on the recruitment of cat superior colliculus for multisensory integration (Meredith & Stein, 1983), several studies with humans (Besle et al., 2008; Foxe et al., 2000, 2002; Hoefer et al., 2013; Lewis & Noppeney, 2010; Liang et al., 2013; Martuzzi et al., 2007; Molholm et al., 2002, 2004; Noesselt et al., 2007; Schürmann et al., 2006; Werner & Noppeney, 2010), primates (Fu et al., 2003; Kayser et al., 2005, 2008, 2010; Lakatos et al., 2007; Schroeder & Foxe, 2002; Schroeder et al., 2001) and rodents (Bizley & King, 2009; Bizley et al., 2007) have proved driving or modulatory effects of cross-modal stimuli at the bottom of the sensory processing hierarchy. Moreover, these effects are supported by direct thalamo-cortical and cortico-cortical anatomical connections (Musacchia & Schroeder, 2009; Schroeder et al., 2003) and reciprocal stimulus-driven entrainment (Kayser et al., 2008, 2010; Lakatos et al., 2007; Senkowski et al., 2008). Importantly, there is evidence that such early multisensory interactions support cross-modal binding (Atilgan et al., 2018), which in turn enhances selective attention during competition for processing resources (Maddox et al., 2015), as mentioned above.

To reconcile the apparent dichotomy between late and early integration frameworks, the parallel integration framework (Figure 1.1C) proposes that multisensory integration takes place at different processing stages in a task-dependent manner (Calvert & Thesen, 2004; Noppeney et al., 2018). On the one hand, multisensory interactions in early sensory areas boost perceptual salience for stimuli detection; on the other hand, multisensory interactions in higher-order association areas are responsible for the generation of more complex task-specific representations (Lewis & Noppeney, 2010; Werner & Noppeney, 2010). This is also

in line with recent neuroimaging evidence that Bayesian Causal Inference is accomplished at the top of the dorsal cortical hierarchy (i..e. anterior intraparietal sulcus) for spatial localisation (Rohe & Noppeney, 2015a, 2016). In other words, the parallel integration framework underscores the need to move beyond identifying multisensory interactions, since they are ubiquitous across the (sub)cortical hierarchy (Driver & Noesselt, 2008; Foxe & Schroeder, 2005; Ghazanfar & Schroeder, 2006; Kayser & Logothetis, 2007; Schroeder & Foxe, 2005). Instead, it is necessary to characterise the neural properties, computational principles and behavioural relevance of these interactions (Noppeney et al., 2018). The parallel integration framework does not explicitly describe the role of attention; however, it may offer a fruitful way to also reconcile the multifaceted findings regarding the interplay between attention and multisensory processes (Koelewijn et al., 2010). The present thesis contributes to the ongoing debate (Koelewijn et al., 2010; Macaluso et al., 2016; Talsma et al., 2010; Tang et al., 2016) by providing behavioural and neural evidence of a parallel integration framework whereby attention and multisensory integration synergistically interact at multiple levels of processing to guide effective behaviour.

1.4 Overview of the present thesis

The current work embraces the Bayesian Causal Inference framework (Körding et al., 2007; Shams & Beierholm, 2010) and targets the interplay between attention and multisensory integration from two complementary perspectives.

On the one hand, I investigate whether selective attention impacts multisensory perceptual inference. In particular, external attention may modulate the reliability of sensory information and thereby impact reliability-weighted integration. Furthermore, internal attention may determine the selection of internal task-relevant representations and thus bias

final responses. With these hypotheses in mind, Chapter 3 combines psychophysics and computational modelling to address whether and how endogenous modality-specific attention influences the spatial ventriloquist effect. Moreover, Chapter 4 employs functional magnetic resonance imaging (fMRI) to characterise the respective neural mechanisms.

On the other hand, I investigate whether multisensory objects (determined via cross-modal temporal coherence) capture attention during competition for attentional resources (as do unisensory objects, Humphreys & Riddoch, 2003; Kimchi et al., 2007; Yeshurun et al., 2009). To this end, Chapter 5 employs psychophysics to evaluate whether cross-modal binding changes target detectability within a cocktail-party scenario, thus indexing object-based attentional capture. Furthermore, Chapter 6 employs fMRI to address the neural mechanisms of cross-modal binding with or without competition for attentional resources (i.e. within or outside a cocktail-party scenario).

Importantly, the present empirical work exploits different types of sensory pairings (audio-visual stimuli in Chapters 3-4; audio-tactile stimuli in Chapters 5-6). This choice reflects the deliberate attempt to characterise the interplay between attention and multisensory integration from a general computational point of view, no matter the specific sensory combinations that are adopted. Furthermore, the choice of different sensory scenes and tasks (spatial ventriloquism in Chapters 3-4; target detection and cocktail-party scenario in Chapters 5-6) reflects the attempt to fill the gap of knowledge in the current literature building on existing empirical work, as discussed in further details in each empirical chapter. Nevertheless, no matter these marginal differences, the four empirical chapters collectively contribute to the same debate: they provide complementary evidence for the mutual interplay between attention and multisensory integration at the behavioural and neural level. Chapter 7 connects the present empirical findings, integrates them with background literature into a

cohesive model and thereby suggests future directions of research. Importantly, before delving into the empirical work, Chapter 2 provides a general overview of the principal research techniques employed in the present thesis.

CHAPTER 2

METHODOLOGIES

The following chapter provides an overview of the primary research techniques employed in the present thesis. Firstly, I outline the approaches used for the analysis of behavioural responses in spatial localisation tasks (see Chapters 3 and 4) and target detection tasks (see Chapter 5). Next, I describe the Bayesian Causal Inference (BCI) approach to the study of multisensory integration that is employed in Chapters 3 and 4. Finally, I summarise the functional neuroimaging techniques employed in Chapters 4 and 6.

2.1 Behavioural analyses

2.1.1 Audio-visual spatial localisation

The characterisation of spatial localisation performance represents a key behavioural measure for Chapters 3 and 4 of the present thesis, where participants reported the perceived location of auditory or visual stimuli sampled from various azimuthal positions.

In the case of non-ambiguous spatial locations (i.e. unisensory stimuli or audio-visual stimuli presented in the same location), the root-mean-square error (RMSE) between response y and true location x can be computed for each participant as follows:

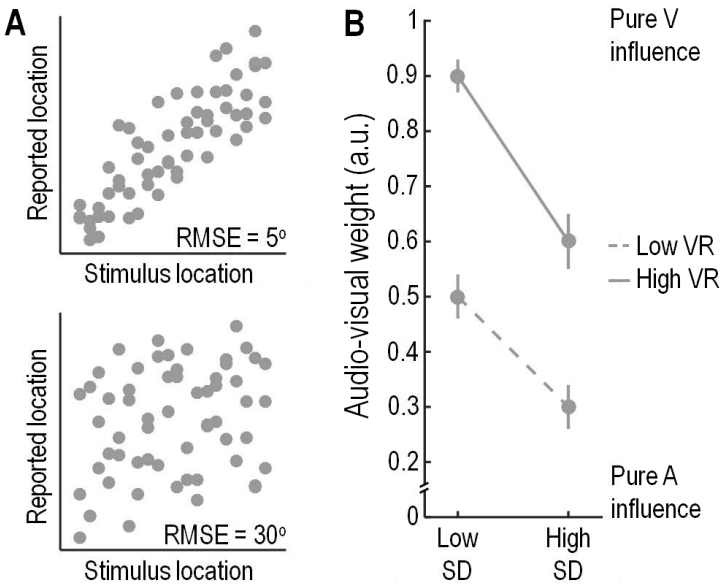

Figure 2.1: Behavioural analysis of audio-visual spatial localisation

A) Root-mean-square error (RMSE) between reported location y and true stimulus location x: lower RMSE reflects higher localisation accuracy. Dots represent individual reported locations from a hypothetical observer. B) Audio-visual weight index: (Reported location – A location) ÷ (V location – A location) as a function of audio-visual spatial disparity (SD) and visual reliability (VR). Adapted from Rohe & Noppeney, 2015b: participants always reported the auditory location; SD was manipulated in degrees of visual angle along the azimuth; VR was manipulated via the spread of the Gaussian cloud of dots which constituted the visual stimulus (more spread = less spatial reliability). A: auditory; V: visual.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}}$$

where i = single trial, n = total number of trials. RMSE measures in degrees of visual angle the average magnitude of response error, or complementarily the overall localisation accuracy (e.g. Jones et al., 2019), as shown in Figure 2.1A.

When auditory and visual stimuli are sampled from different locations along the azimuth and participants selectively respond to one sensory modality, a measure called audio-visual weight index W_{AV} (Figure 2.1B) can be computed to quantify the relative influence of the auditory and visual signals on the perceived location (Petrini et al., 2015; Rohe &

Noppeney, 2015b). Specifically, W_{AV} computes the difference between the reported location and the true auditory location, normalised by the distance between true visual and auditory locations:

$$W_{AV} = \frac{Reported\ location - Auditory\ location}{Visual\ location - Auditory\ location}$$

As a result, W_{AV} is a ratio index that varies between 0 and 1. A W_{AV} of 0 reflects no influence of the visual signal location on the localisation response (i.e., full influence of the auditory signal location); a W_{AV} of 1 reflects full influence of the visual signal location on the localisation response (i.e., no influence of the auditory signal location). In other words, the W_{AV} index represents a quantitative measure of spatial ventriloquism.

2.1.2 Signal detection theory

The characterisation of detection performance under conditions of perceptual uncertainty and attention competition represents a key behavioural measure for Chapter 5 of the present thesis. Signal detection theory (STD) provides a general framework for the quantitative analysis of performance under conditions of uncertainty, namely when target signals are presented in the midst of noise (Wickens, 2002). In the example of a target detection task, where participants answer "yes" or "no" to the question "did you perceive a target?", trials may be sampled from one of two distributions (Figure 2.2): either from a signal distribution (i.e. the trial contains a target) or a noise distribution (i.e. the trial does not contain a target)[1].

Given two types of trials (i.e. signal or noise) and two types of responses (i.e. yes or no), participants' answers can be partitioned into four categories (Table 2.1): yes response when signal is present (Hit); no response when signal is present (Miss); yes response when signal is absent (False alarm); no response when signal is absent (Correct rejection).

[1] Following the *equal-variance Gaussian model*, signal and noise distributions are Gaussians with variance = 1.

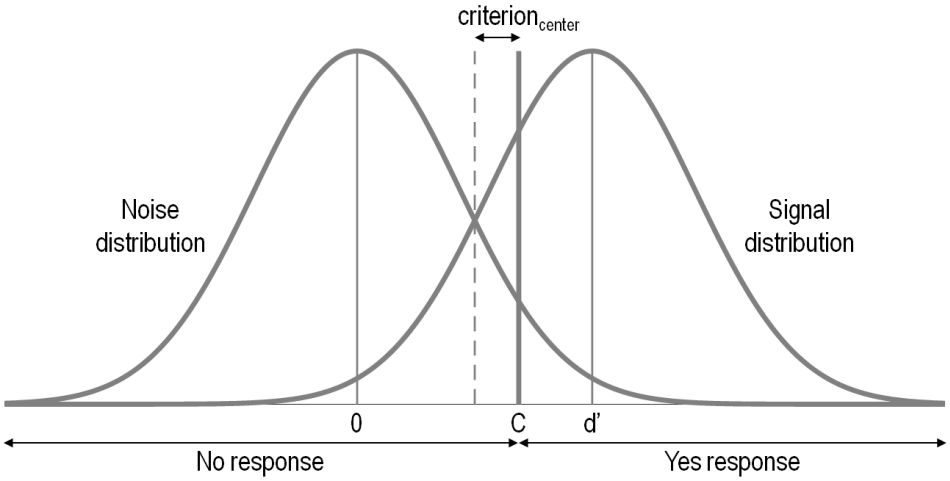

Figure 2.2: Schematic representation of Signal Detection Theory (STD)

On the left, noise distribution; on the right, signal distribution. Perceptual sensitivity (d') is the degree of separation between signal and noise distributions; response bias (criterion$_{center}$) is the distance between the criterion C and the point halfway between signal and noise distributions. Adapted from Wickens. 2002.

Importantly, hits and correct rejections represent correct responses, whereas misses and false alarms represent errors.

	YES RESPONSE	**NO RESPONSE**
SIGNAL TRIAL	Hit	Miss
NOISE TRIAL	False Alarm	Correct rejection

Table 2.1: Types of responses in Signal Detection Theory

In signal detection theory, answers can be partitioned into four categories: hits, misses, false alarms, correct rejections. Signal trials contain target; noise trials do not contain target; "yes" and "no" responses are relative to the question "did you perceive a target?".

Consequently, it is possible to calculate the following critical measures:

$$hit\ rate\ (h) = \frac{number\ of\ hits}{number\ of\ signal\ trials}$$

$$false\ alarm\ rate\ (f) = \frac{number\ of\ false\ alarms}{number\ of\ noise\ trials}$$

and the subsequent complementary measures:

$$miss\ rate = 1 - h$$

$$correct\ rejection\ rate = 1 - f$$

The signal detection estimation process translates h and f into quantities that separate perceptual and decisional processes. Perceptual sensitivity (also called d') represents the ability to perceptually discriminate signal from noise. It corresponds to the degree of separation between signal and noise distributions (Figure 2.2) and is calculated as the difference between z-normalised h and f (Wickens, 2002):

$$d' = Z(h) - Z(f)$$

The larger d' the more capable is the observer to perceptually discriminate between signal and noise. Instead, response bias (also called $criterion_{center}$) reflects the decisional strategy applied by the observer to produce a response, i.e. the propensity to respond yes or no. It corresponds to the distance between the criterion (i.e. the boundary that separates yes and no responses) and the point halfway between signal and noise distributions (Figure 2.2), and is calculated as follows (Wickens, 2002):

$$criterion_{center} = -\frac{Z(h) + Z(f)}{2}$$

A $criterion_{center}$ equal to 0 represents no preference for yes or no responses; a negative $criterion_{center}$ reflects a liberal decision strategy, i.e. the tendency to respond yes; a positive $criterion_{center}$ represents a conservative decision strategy, i.e. the tendency to respond no.

2.2 Computational modelling

2.2.1 Bayesian Causal Inference model

Mathematical modelling was used in Chapters 3 and 4 of the present thesis to characterise the computational principles underlying multisensory perceptual inference. Specifically, we employed the Bayesian Causal Inference (BCI) model (Körding et al., 2007), which posits that during perceptual inference the brain inverts a probabilistic generative model of the sensory inputs (Figure 2.3), such that the cause of stimulation is inferred based on prior knowledge and available sensory evidence. In line with the focus of the present thesis, I will concentrate on the application of this model to the characterisation of audio-visual localisation in a spatial ventriloquist paradigm, where synchronous auditory and visual stimuli are sampled from the same or different azimuthal positions and participants report their perceived auditory or visual spatial position.

For each experimental trial, auditory and visual spatial representations are modelled as Gaussian likelihood distributions centred on the true auditory (respectively, visual) spatial location x_A (respectively, x_V) and with a given standard deviation σ_A (respectively, σ_V). Thus, σ_A and σ_V define how noisy (or complementarily, how reliable) each sensory modality is in the spatial domain. The model also accommodates the prior tendency to locate stimuli in the centre of the field of view (i.e. central bias, Odegaard et al., 2015), which is modelled as a spatial prior distribution $N(\mu_P, \sigma_P^2)$ with $\mu_P = 0$. Thus, σ_P models the strength of the central bias (i.e. the bigger σ_P, the more noisy or weaker the central bias). The likelihoods and spatial prior are used to compute the posterior probability of the stimuli location given a particular causal structure, namely one common cause (i.e. forced fusion, see C = 1 in Figure 2.3) or two

separate causes (i.e. full segregation, see C = 2 in Figure 2.3)[2]. Under the assumption of a common cause, the posterior probability of the audio-visual source s is given by Bayes' rule:

$$p(s|x_A, x_V; C = 1) = \frac{p(x_A|s)p(x_V|s)p(s)}{p(x_A x_V)}$$

Given Gaussian distributions, the best perceptual estimate of the audio-visual location is a reliability-weighted average of auditory and visual spatial representations and spatial prior:

$$\hat{S}_{A,C=1} = \hat{S}_{V,C=1} = \frac{\frac{x_A}{\sigma_A^2} + \frac{x_V}{\sigma_V^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2} + \frac{1}{\sigma_P^2}}$$

Thus, the more reliable a source of information is (i.e. the smaller the corresponding σ), the more it drives the final audio-visual integrated estimate. It is worth noting that this portion of the BCI model descends from the Maximum Likelihood Estimation (MLE) model for multisensory integration (Alais & Burr, 2004; Ernst & Banks, 2002; Ernst & Bülthoff, 2004), with the addition of the spatial prior. Therefore, the BCI model represents an extension of the MLE model, wherein there is no mandatory integration but instead observers flexibly transition between integration and segregation depending on which causal structure is more likely in the current environmental context. Under the assumption of two separate causes, the posterior probability of the auditory (and respectively, visual) location s is given by:

$$p(s_A|x_A; C = 2) = \frac{p(x_A|s_A)p(s)}{p(x_A)}, p(s_V|x_V; C = 2) = \frac{p(x_V|s_V)p(s)}{p(x_V)}$$

It follows that the best perceptual estimate of the auditory (and respectively, visual) location given two separate sources is described by:

$$\hat{S}_{A,C=2} = \frac{\frac{x_A}{\sigma_A^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_P^2}}, \hat{S}_{V,C=2} = \frac{\frac{x_V}{\sigma_V^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_V^2} + \frac{1}{\sigma_P^2}}$$

[2] At the computational level, forced fusion equates to cross-modal binding (Bizley et al., 2016).

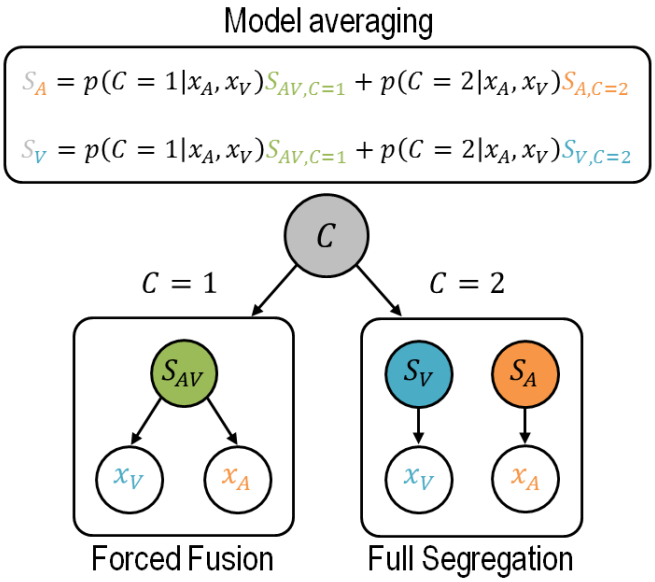

Figure 2.3: Probabilistic generative model of Bayesian Causal Inference (BCI)

The latent variable C (causal inference) determines, via the model averaging decision function, which sub-model generates the data: if C = 1 (left), one common cause is responsible for both visual and auditory inputs (forced fusion); if C = 2 (right), two independent causes are separately responsible for visual and auditory inputs respectively (full segregation). The BCI model is inverted during perceptual inference to determine the cause(s) of stimulation based on prior knowledge and available sensory evidence. Adapted from Körding et al., 2007.

Also the underlying causal structure is inferred based on prior knowledge and available sensory evidence and its posterior probability is given by:

$$p(C|x_A, x_V) = \frac{p(x_A, x_V|C)p(C)}{p(x_A, x_V)}$$

In particular, the posterior probability of a common cause is described by:

$$p(C = 1|x_A, x_V) = \frac{p(x_A, x_V|C = 1)p_{common}}{p(x_A, x_V|C = 1)p_{common} + p(x_A, x_V|C = 2)(1 - p_{common})}$$

where p_{common} represented the prior probability of a common cause (i.e. prior binding tendency), which ranges between 0 and 1. Importantly, a p_{common} equal to 0 represents the a priori tendency to fully segregate auditory and visual inputs, whereas a p_{common} equal to 1 represents the a priori tendency to fully integrate them into a unified percept. Accordingly,

imposing a fixed p_{common} equal to 0 reduces the BCI model to the so-called Full Segregation model; imposing a fixed p_{common} equal to 1 reduces the BCI model to the so-called Forced Fusion model. For completeness, the posterior probability of two separate sources is given by:

$$p(C = 2|x_A, x_V) = 1 - p(C = 1|x_A, x_V)$$

Critically, the final estimates of the auditory and visual locations ($\hat{S}_A$ and $\hat{S}_V$ respectively) account for the fact that the observer does not know the underlying causal structure. The model performs this final computation via a decision function, i.e. a strategy that combines the perceptual estimates under forced fusion and full segregation in the face of causal uncertainty. This allows to gracefully transition between integration and segregation based on how likely each causal structure is. It has been proven that the best decision function at simulating participants' behaviour is *model averaging* (Figure 2.3; Rohe & Noppeney, 2015b). Here, the integrated spatial estimate $\hat{S}_{AV,C=1}$ is combined with the segregated spatial estimate in the sensory modality that needs to be reported ($\hat{S}_{A,C=2}$ for auditory report; $\hat{S}_{V,C=2}$ for visual reports) and each spatial estimate is weighted in proportion to the posterior probability of the respective causal structure, as described by

$$\hat{S}_A = p(C = 1|x_A, x_V)\hat{S}_{AV,C=1} + p(C = 2|x_A, x_V)\hat{S}_{A,C=2}$$

$$\hat{S}_V = p(C = 1|x_A, x_V)\hat{S}_{AV,C=1} + p(C = 2|x_A, x_V)\hat{S}_{V,C=2}$$

Consequently, the model explicitly accounts for the effect of task relevance (i.e. which sensory modality needs to be reported) by utilising either of the final BCI estimates: $\hat{S}_A$ or $\hat{S}_V$.

2.2.2 Model fitting

In summary, the BCI generative model comprises the following set of free parameters: the common-source prior p_{common}, the spatial prior standard deviation σ_P, the auditory standard deviation σ_A and the visual standard deviation σ_V. The BCI model fitting procedure utilises a

two-stage process to optimise the model's free parameters to each participant's responses, via maximisation of the respective log-likelihood (i.e. a probabilistic measure that quantifies how well the model describes the current data; Körding et al., 2007). First, a grid search is run through an initial parameter space, which is initialised with a likely range of values for each parameter of the model. Via an iterative process (here 10,000 iterations), the BCI model produces predicted distributions of the stimuli locations x_A and x_V for each condition and set of parameters and the respective log-likelihoods are summed over conditions. Second, the combination of parameters with the highest log-likelihood is used as a starting point in an optimisation process (as implemented in the MATLAB2014b function *fminsearchbnd*) that refines the parameters' values until the final highest log-likelihood across conditions is achieved, i.e. until the model produces the best possible simulation of each participant's true responses.

2.2.3 Model comparison

Different models (e.g. Full Segregation, Forced Fusion and Bayesian Causal Inference models) can be fit to the same dataset. Model comparison allows identifying, among a set of candidate models, the one which most likely generated the observed data and should therefore be preferred (Lee & Wagenmakers, 2014). Crucially, a model's fit (i.e. how well is describes the current data) is weighted against its complexity (i.e. its number of parameters), in order to avoid overfitting and thus loss of generalisability (i.e. how well it will describe future data). To this aim, several summary indexes for model comparison have been developed, among which is the commonly used Bayesian Information Criterion (BIC; Schwarz, 1978). The BIC is an index that summarises a model's fit while introducing a penalty term for the number of parameters in the model, as described by

$$BIC = LL - 0.5 \times P \times \ln N$$

where LL = log-likelihood, P = number of parameters, N = number of data points (Raftery, 1995). Consequently, the model with the highest BIC is preferred[3].

A fixed-effect approach can be used to compare alternative models. First, a BIC is calculated for each model and participant; then, BICs are summed over participants for each model such that one group-level BIC is derived for each model; finally, alternative models are directly compared by taking the difference of their group-level BICs (relative BIC or relBIC). A major drawback of this fixed-effect approach is its incapability to deal with outliers and population heterogeneity, since it assumes a homogenous population with one (unknown) model (i.e. a fixed effect). Instead, a random-effect approach is robust to outliers and more generalisable, since it assumes a heterogeneous population with different models (i.e. a random effect) drawn from a fixed (unknown) distribution (Stephan et al., 2009).

Bayesian model selection (BMS) is a commonly used random-effect approach for model comparison that is grounded in Bayesian inference (Rigoux et al., 2014; Stephan et al., 2009). The aim of BMS is to estimate which model prevails at the population level. For each candidate model, a measure called *model evidence* is computed, which quantifies the probability of data D given model M, i.e. p(D | M). In other words, model evidence quantifies the quality of the model's predictions. Model evidence, also known as *marginal likelihood*, derives from the Bayes theorem for parameter estimation:

$$p(\theta|D,M) = \frac{p(D|\theta,M \times p(\theta|M))}{p(D|M)} = \frac{likelihood \times prior}{marginal\ likelihood}$$

where θ is a parameter, M is the model and D is the data (Lee & Wagenmakers, 2014).

[3] The BIC belongs to the interval scale of measurement and thus what matters is the difference in BIC values among models, not the absolute values.

Crucially, the marginal likelihood is computed by a weighted average of all the likelihoods across the parameter space, with weights represented by the parameter values' prior probabilities:

$$p(D|M) = \sum_{i=1}^{k} p(D|\theta_i, M) \times p(\theta_i|M)$$

with k = number of values that θ can assume. Analytically, the marginal likelihood can be approximated by the BIC and the maximisation of its logarithm finally produces BMS. As a result, it is possible to identify the model that generated the given data with the highest protected exceedance probability, namely the probability that a model is more likely than the other models, beyond differences due to chance (Rigoux et al., 2014)[4].

2.3 Functional magnetic resonance imaging

Functional neuroimaging is a class of invasive and non-invasive research techniques that is widely used in neuroscience to establish correlational or causal links between behaviour and brain functioning (Huettel et al., 2004). In Chapters 4 and 6 of the present thesis, I employed blood-oxygenation-level-dependent functional magnetic resonance imaging (BOLD fMRI) to non-invasively correlate participants' cognitive states with changes of brain metabolism, which is a proxy of neural activation with high spatial resolution (i.e. order of mm). In the following, I introduce the fundamentals of BOLD fMRI and I describe the fMRI techniques employed in this thesis. A supplementary glossary of fMRI terminology is provided at the end of the section.

[4] In the present work, subject-specific BICs for each alternative model were used as an approximation of model evidence to compute BMS as implemented in the SPM12 function *spm_BMS* (Rigoux et al., 2014; Stephan et al., 2009).

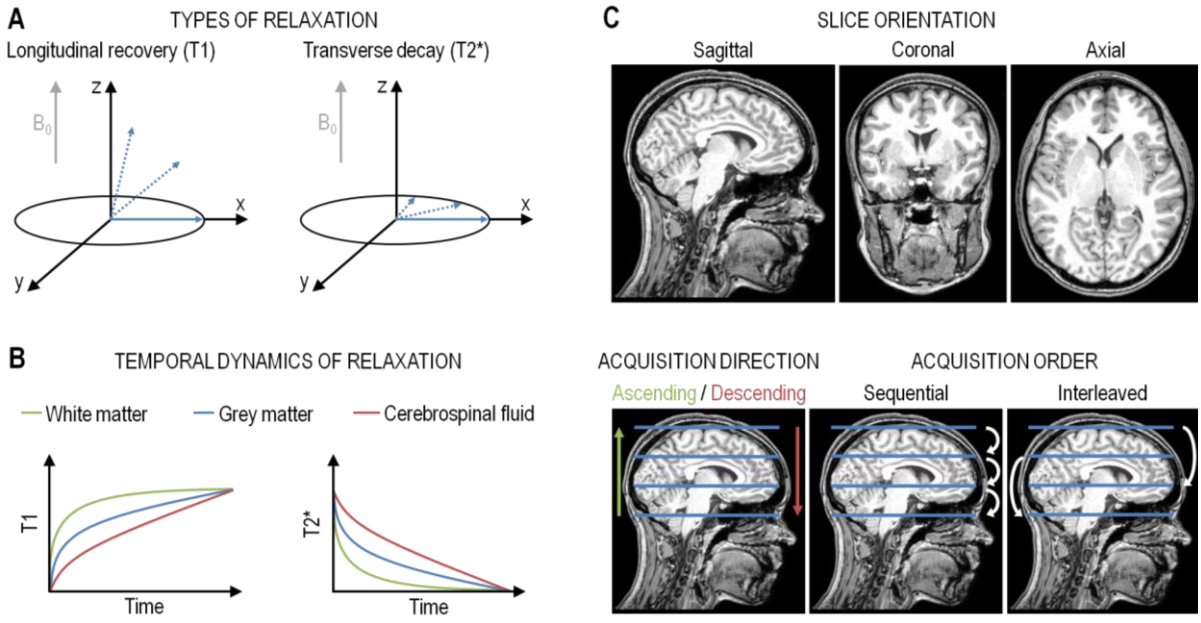

Figure 2.4: Principles of fMRI data acquisition

A) Types of relaxation: on the left, recovery of longitudinal direction as the spin system returns to the parallel state (T1 or longitudinal recovery); on the right, decrease of net magnetization amplitude due to loss of phase coherence of the spin system and local magnetic field inhomogeneities (T2* or transverse decay). B) T1 recovery and T2* decay exhibit different temporal dynamics and vary according to tissue type. Adapted from Pooley, 2005. C) Slice selection is characterised by a specific orientation (sagittal, coronal or axial), acquisition direction (ascending or descending) and order (sequential or interleaved). Here we show acquisition of axial slices; for interleaved acquisition order, even slices are acquired before or after odd slices. B_0: static magnetic field.

2.3.1 Principles of BOLD fMRI

Hydrogen atoms in water molecules, which are abundantly present in the human body, align their axis of spin either along (parallel to) or against (anti-parallel to) the static magnetic field (B_0) created via electromagnetism in MRI scanners (Huettel et al., 2004; Pooley, 2005). The majority of spins assume the parallel (low-energy) state, resulting in a net magnetisation of the spin system that is longitudinal to B_0. A radio-frequency (RF) pulse determines energy absorption by the spin system (excitation) such that many spins assume the anti-parallel (high-energy) state, resulting in a net magnetisation that is transversal to B_0. Upon termination

of the RF pulse, spins return to their state of equilibrium generating a magnetic flux (relaxation); the detection of the consequent electric current (reception) constitutes the MR signal. Two types of relaxation can be measured (Figure 2.4A): recovery of longitudinal direction as the spin system returns to the parallel state (T1 or longitudinal recovery); decrease of net magnetization amplitude due to loss of phase coherence of the spin system and local magnetic field inhomogeneities (T2* or transverse decay). T1 recovery and T2* decay exhibit different temporal dynamics and vary according to tissue type (Figure 2.4B). Thus, by specifying pulse sequences tuned to a specific relaxation type (T1 or T2* contrast, respectively) it is possible to optimise MR signal measurements to a particular physical property (functional or anatomical imaging, respectively). In particular, it is necessary to optimise the time interval between successive RF pulses (repetition time or TR) and the time interval between an RF pulse and reception (echo time or TE). Crucially, by applying additional magnetic fields that systematically vary B_0 strength over 3D space (spatial gradients), specific brain slices can be selectively excited over time (slice selection; Figure 2.4C) with a given orientation (i.e. sagittal, coronal or axial) and thickness (in mm). In addition, the combination of RF pulses and spatial gradients determines how the MR signal is acquired in terms of direction (i.e. ascending or descending) and order (i.e. sequential or interleaved). By changing spatial gradients rapidly after one RF pulse, it is possible to acquire an entire slice within a few or even a single TR (echo-planar imaging or EPI).

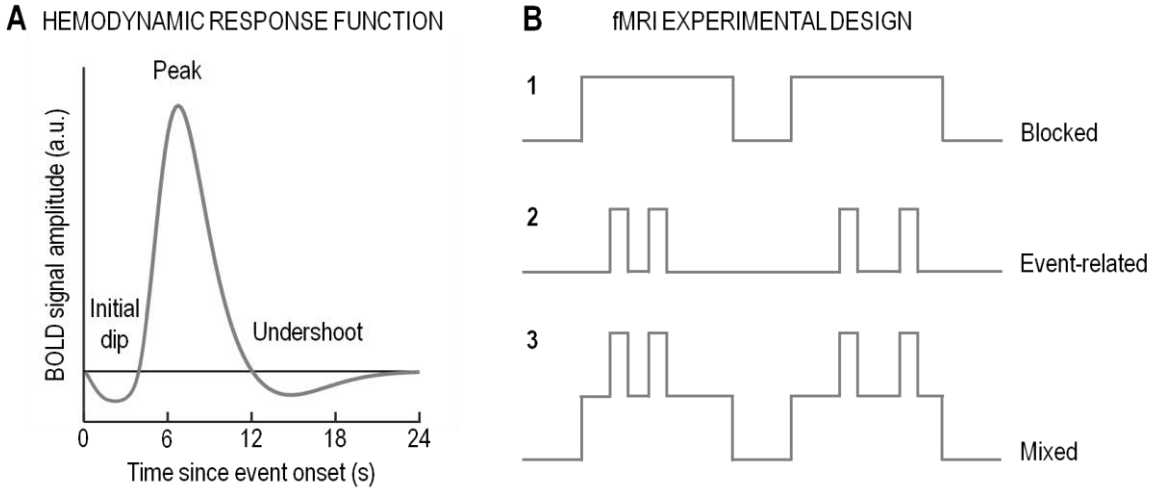

Figure 2.5: Hemodynamic response function (HRF) and fMRI experimental design

A) The HRF describes voxel-wise changes of BOLD signal over time. This results from relative changes of oxy/deoxygenated haemoglobin concentration, which alter local magnetic field inhomogeneities and thus modify MR signal intensity in T2* images. Specifically, the BOLD signal is directly proportional to the concentration of oxygenated haemoglobin. Three key phases summarise the HRF: local neuronal activity triggers initial oxygen consumption (initial dip); consequently, cerebral blood flow (CBF) provides local oxygen supply (peak); blood inflow is greater than blood outflow, causing a concurrent increase of cerebral blood volume (CBV); following cessation of neuronal activity, CBF decreases more rapidly than CBV causing a relative decrease of oxygen concentration (undershoot). B) fMRI experimental design: 1. blocked (experimental conditions are presented in separate blocks for an extended period); 2. event-related (short-duration trials of different experimental conditions are presented in intermixed order); 3. mixed (combination of blocked and event-related designs features). BOLD: blood-oxygenation-level-dependent.

Crucially, neural activity and MR signal correlate via neurovascular coupling (Logothetis et al., 2001), which is described by a cascade of events. First, neural activity determines local consumption of oxygen and glucose, which is turn triggers oxygen supply via cerebral blood flow. As a result, the ratio of blood's oxygenated and de-oxygenated haemoglobin changes over time. Since de-oxygenated haemoglobin is paramagnetic, it alters local magnetic field inhomogeneities, which in turn impact T2* decay and define the so-called blood-oxygenation-level-dependent (BOLD) signal. Changes of BOLD signal over time are described by the hemodynamic response function (HRF; Figure 2.5A). Crucially, the

HRF shows the properties of scaling (i.e. magnitude of system output is proportional to system input) and superposition (i.e. total response to a set of inputs is equal to summation of independent responses), which are critical assumption for fMRI design and data analysis (Huettel et al., 2004).

2.3.2 fMRI experimental design

There are two commonly used types of fMRI experimental design (Huettel et al., 2004): blocked design (experimental conditions are presented in separate blocks for an extended period, Figure 2.5B1) and event-related design (short-duration trials of different experimental conditions are presented in intermixed order, Figure 2.5B2). The selection of an fMRI design depends on research questions and task constraints. Based on superposition, blocked designs maximise the ability to measure changes of BOLD amplitude (detection) generated by state-related processes (e.g. sustained attention). Based on scaling, event-related designs maximise the ability to characterise the HRF time-course (estimation) generated by item-related processes (e.g. burst of noise at a given spatial position). Sometimes it is possible to combine features of blocked and event-related designs into mixed designs (Figure 2.5B3): some experimental conditions are organised into separate blocks (e.g. auditory versus visual attention), within which multiple short-duration trials of different conditions are intermixed (e.g. stimuli sampled from different spatial positions). As a result, mixed designs allow the investigation of both state-related and item-related processes. In order to optimise the fMRI procedure, it is possible to simulate the power of BOLD signal estimation in relation to a specific research question (i.e. design efficiency; Henson, 2006). This optimisation procedure was used in Chapter 4 of the present thesis to define the optimal order and length of events (in

seconds) for each experimental run. To this end, the power of BOLD signal was estimated for each experimental regressor relative to the baseline.

Importantly, fMRI experimental designs should include rest periods (such as no-task blocks or null-events), which provide a baseline level for the BOLD signal and thus allow the measurement of experiment-related activations. Moreover, the inclusion of control conditions allows isolating BOLD activity associated with specific conditions. Finally, factorial designs provide the highest degree of flexibility in terms of conditions comparison, as they allow testing for main effects and interactions (Friston et al., 1996; Price et al., 1997).

2.3.3 fMRI data analysis

Analyses of brain imaging data included in the present thesis are based on the following software packages and toolboxes: SPM12 (Friston et al., 1994a) was used for pre-processing of structural and functional MRI data and mass-univariate general linear modelling; The Decoding Toolbox 3.96 (Hebart et al., 2015) was employed for multivariate decoding analyses; FreeSurfer 5.3.0 (Fischl, 2012) was used for the definition of subject-specific anatomical regions of interest (ROIs). Custom code was developed in MATLAB2014b to supplement several analysis steps.

2.3.4 Pre-processing

Before statistical analysis, a set of computational procedures is applied to MRI data in order to minimise unwanted signal variability (Huettel et al., 2004).

Head motion between consecutive volumes creates a spatial mismatch in the volumes time series. *Spatial realignment* accounts for this mismatch via an interpolation procedure that spatially aligns all the volumes to a reference (usually the first or the middle volume in the

time series). Since the head's shape and size do not change between volumes, rigid-body transformations (translation and rotation in 3D directions) are applied. Conversely, head motion within a single volume acquisition interacts with local magnetic field inhomogeneities and distorts the shape and size of the image. *Unwarping* accounts for these distortions using information about head shape and size derived from spatial realignment in an iterative process until residual errors are minimised.

Since different slices are acquired at different time points with a given temporal frequency (TR), the HRF is not sampled instantaneously across the volume. *Slice-time correction* accounts for these temporal delays via an interpolation procedure that temporally aligns all the slices to a reference (usually the middle slice).

To allow precise anatomical identification of functional activity, *coregistration* applies rigid-body transformations to the anatomical image to spatially align it to a reference image (e.g. mean functional image derived from spatial realignment).

For group-level analyses, it is necessary to convert anatomical and functional volumes from participants' native space to a standard space, which can be done via a two-step procedure (Ashburner & Friston, 2005). First, *segmentation* partitions the brain tissue of each participant's anatomical image into different tissue types (including grey matter, white matter and cerebrospinal fluid). Specifically, segmentation uses information about tissue boundaries from tissue probability maps, which are derived from a reference sample of several anatomical images. Second, *normalization* employs image intensity and prior spatial information derived from the segmentation process to warp anatomical and functional volumes from different participants into the same reference space (conventionally, MNI

space). In particular, since different brains differ in size and shape, non-linear transformations are applied[5].

Given functional similarities of adjacent brain areas and the blurring of the vascular system, fMRI data inherently carry spatial smoothness, thus requiring the removal of high-frequency spatial noise. To this end, *spatial smoothing* is applied to functional images. In particular, each BOLD time series is convolved with a low-pass Gaussian filter (kernel) and the spatial extent of smoothing is proportional to the user-specified full-width-half-maximum (FWHM) of such kernel. Besides boosting the signal-to-noise ratio of functional data, spatial smoothing improves the validity of parametric statistical tests by increasing the normality of the error distribution; furthermore, it ameliorates the spatial overlap of functional activations across participants for group-level analyses.

Different pre-processing pipelines are employed depending on experimental design, data acquisition protocol and planned statistical analyses (Poldrack et al., 2011). In the present thesis, all the above steps were applied to functional EPI images entering random-effects general linear modelling in Chapter 4 (Ashburner et al., 2015). Normalisation and spatial smoothing were skipped prior to entering subject-specific functional EPI images into multivariate decoding in Chapter 4 (Hebart et al., 2015). Finally, slice-time correction was skipped in Chapter 6 given a multiband interleaved data acquisition protocol (Ashburner et al., 2015). Since multiband acquisition procedures allow instantaneous sampling of multiple slices within the same volume, they significantly decrease time delays for HRF sampling across the volume. Furthermore, interleaved acquisition protocols allow a more temporally distributed HRF sampling across the volume compared to sequential acquisition protocols.

[5] Since anatomical and functional volumes occupy the same space after coregistration, it is possible to apply the highly-detailed segmentation information to both anatomical and functional volumes, thus ensuring high spatial accuracy during normalisation.

For these reasons, temporal delays in HRF sampling were highly decreased and rendered slice-time correction redundant in Chapter 6.

2.3.5 Mass-univariate general linear modelling

Following the critical assumption that the BOLD signal is a linear system (i.e. it shows properties of scaling and superposition) with time invariance (i.e. input time delays equate to output time delays), a general linear model (GLM) is fitted to the time course of each voxel (mass-univariate approach; Friston et al., 1994a). Experimental conditions are included as regressors in the GLM design matrix. To account for noise generated by confounding factors, nuisance variables such as motion parameters from spatial realignment are also included as regressors. To construct the expected HRF in response to each condition, the corresponding onsets are convolved with a canonical HRF. For transient events, the duration is conventionally set to 0; for blocked conditions, duration equals to the block's length in seconds. To account for variations in latency and duration of the hemodynamic response in event-related designs, the temporal and dispersion derivatives of the HRF can be included as regressors in the design matrix for each experimental condition (Friston et al., 1998). The linear combination of all regressors in the design matrix determines the GLM, which is described as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

where Y is the measured voxel-wise HRF time series; X is a regressor; n is the total number of regressors; β_i is a parameter estimate, i.e. the contribution of the corresponding regressor to Y; β_0 reflects the total contribution of all factors kept constant in the experiment; ε is the error term (Friston et al., 1994a).

Critically, the BOLD signal belongs to the interval scale of measurement and thus what matters is the difference in explained BOLD variance across conditions. Consequently, parameter estimates of different conditions are contrasted for each voxel and the collection of voxel-wise outputs forms a contrast image. Following a hierarchical summary statistics approach, contrast images from each individual are entered as dependent variables in a new GLM, and statistical tests (e.g., ANOVA, t-test) are performed to evaluate which contrasts are significant at the population level (random-effect analysis; Friston et al., 1994a). Importantly, results are corrected for multiple comparisons at the peak level (i.e. voxel-wise activation threshold) or cluster level (i.e. spatial extent threshold given an auxiliary uncorrected voxel-wise activation threshold; Friston et al., 1994b).

2.3.6 Multivariate decoding

Mass-univariate GLM results are constrained by a pre-defined voxel-wise or cluster-wise activation threshold. Decoding techniques improve the sensitivity of fMRI analysis by detecting and exploiting relative changes of activation across distributed patterns of voxels (multivariate approach) to classify or predict experimental conditions (i.e. perceptual or cognitive states) irrespective of a nominal threshold (Haynes, 2015; Mur et al., 2009; Norman et al., 2006; Pereira et al., 2009). In other words, multivariate decoding targets individuation of distributed information content instead of localisation of activations (Hebart & Baker, 2018; Kriegeskorte et al., 2006; Kriegeskorte & Bandettini, 2007). Multivariate decoding can be performed within predefined regions of interest (ROIs) or across the whole brain via the so-called searchlight technique (Kriegeskorte et al., 2006; Kriegeskorte & Bandettini, 2007).

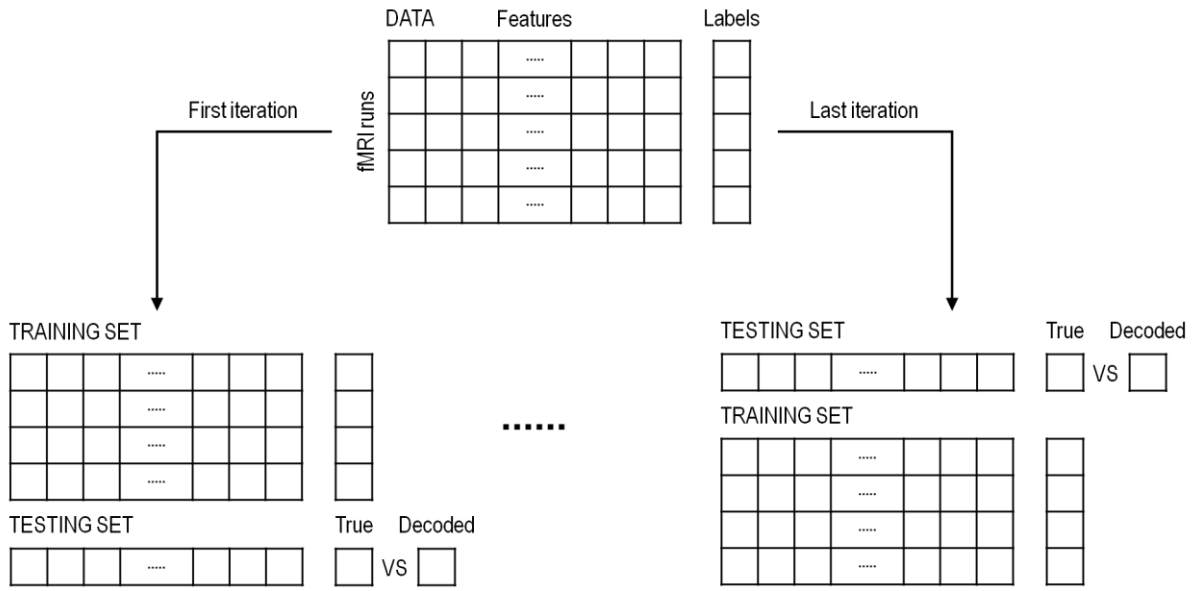

Figure 2.6: Cross-validation scheme for multivariate decoding

In each grid, rows represent fMRI runs and columns represent multi-voxel activation patterns (features). Features can be raw fMRI data or beta images from first-level GLM analyses. Multivariate decoding aims to map features to experimental conditions (labels). In this example, five fMRI runs were acquired and data were split accordingly, thus generating a five-iteration cross-validation. In each iteration (here we show the first and last), the classifier is trained on four runs and tested on the left-out run. The final output is the average decoded value across the five iterations.

In the following, we will focus on ROI-based multivariate decoding (used in this thesis). Machine learning is employed to develop algorithms that map multi-voxel activation patterns (features) to particular experimental conditions (labels) for each participant. For example, linear support vector machines (used in this thesis) consider features as points in a multidimensional space and attempt to find the boundary vector (hyperplane) that best separates these points (Chang & Lin, 2011). To this end, part of the fMRI dataset (training set) is used to train the algorithm and the remaining independent portion of dataset (testing set) is employed to verify the accuracy of the algorithm. Via an iterative process (cross-validation) each portion of the fMRI dataset (e.g. a run) is used to either train or test the algorithm (Figure 2.6) and the mean across iterations produces the final decoded output.

Importantly, while support vector classification provides a classification accuracy and a predicted output on a categorical scale (e.g. face versus house), support vector regression provides a correlation coefficient between features and labels and a predicted output on a continuous scale (e.g. spatial location along the azimuth). Finally, each participant's mean decoding output can be entered into group-level statistical tests to evaluate significance at the population level (Hebart et al., 2015).

2.3.7 Definition of anatomical regions of interest

Following a priori hypotheses based on previous literature, anatomical ROIs were employed in Chapter 4 to constrain the inclusion of functional data in multivariate decoding models. To this end, all ROIs were employed in participants' native space (Hebart et al., 2015).

Visual ROIs (i.e. V1, V2, V3 and intraparietal sulcus) were defined in MNI space via a volume-based probabilistic atlas based on retinotopic mapping of 53 healthy adults (Wang et al., 2015). Low-level auditory ROIs (i.e. A1-2) were defined in MNI space via a volume-based probabilistic atlas based on cytoarchitectonic studies of 10 adult post-mortem brains (Morosan et al., 2001), which is part of the SPM Anatomy Toolbox (Eickhoff et al., 2005). In both probabilistic atlases, the probability of belonging to each ROI is provided for each voxel. Thus, voxels can be assigned to a specific ROI via a user-defined probability threshold. To convert all the above ROIs from standardised MNI space to participants' native space, they were inverse-normalised using a transformation matrix that is produced during the segmentation and normalisation procedure of SPM.

Finally, the higher-level auditory ROI planum temporale (PT) was defined via automated parcellation of cortical surface in participants' native space as implemented in

Freesurfer 5.3.0 (Fischl, 2012) and subsequent extraction of the corresponding region using the Destrieux atlas (Destrieux et al., 2010) in Freesurfer.

2.3.8 Terminology of fMRI

The following list defines commonly used terms in fMRI studies (in alphabetical order).

- Block: a time interval that contains multiple trials
- Cluster: collection of functionally activated voxels in 3D space
- Echo-time (TE): time interval between a radio-frequency pulse and measurement of MR signal
- Field of view (FOV): total spatial extent of head coverage in 3D
- Flip angle: change of net magnetisation angle following a radio-frequency pulse
- GLM design matrix: specification of how regressors change over time
- Montreal Neurological Institute (MNI) space: standard space for normalisation of MRI data derived from the average of several anatomical images
- Region of interest (ROI): pre-determined set of voxels for MRI analyses defined via anatomical landmarks (anatomical ROI) or independent fMRI analysis (functional ROI)
- Repetition time (TR): time interval between successive radio-frequency pulses; it defines sampling frequency
- Run: uninterrupted acquisition of an MRI sequence
- Session: collection of runs within a single scanning visit
- Slice: collection of voxels within a single excitation slab
- Time series: volumes collected at different time points
- Trial: a single instance of an experimental condition
- Volume: collection of slices providing field of view; synonymous of image
- Voxel: 3D volume unit (analogous of pixel in 2D space)

CHAPTER 3

ATTENTION MODULATES SENSORY RELIABILITY AND IMPACTS RESPONSE SELECTION DURING MULTISENSORY PERCEPTUAL INFERENCE

Ambra Ferrari, Uta Noppeney

Computational Cognitive Neuroimaging lab, Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, B15 2TT Birmingham, UK

Citation:

Ferrari, A. & Noppeney, U. (in preparation). Attention modulates sensory reliability and impacts response selection during multisensory perceptual inference.

Authors contributions:

Experiment conceptualisation and design: Ambra Ferrari, Uta Noppeney.
Data collection: Ambra Ferrari.
Data analysis: Ambra Ferrari (supervised by Uta Noppeney).
Writing: Ambra Ferrari (supervised by Uta Noppeney).

Abstract

Effective interactions with our complex multisensory world require the integration of signals that originate from a common source and segregation of signals from separate sources. While it has become clear that endogenous attention impacts multisensory perceptual inference via selection of internal task-relevant representations, it is still a matter of debate whether attention can additionally modulate sensory reliability and thereby affect the relative weights for multisensory integration. The present study addressed such question via the manipulation of endogenous modality-specific attention in an audio-visual spatial ventriloquist task. Pre-stimulus focus and post-stimulus response selection were orthogonally manipulated in a cueing paradigm: participants were cued before stimuli presentation to attend to audition (or vision) and they were cued after stimuli presentation to report their perceived auditory (or visual) location. Psychophysics revealed that pre-stimulus focus and post-stimulus response selection additively affected the ventriloquist effect. Bayesian Causal Inference modelling unveiled the underlying computational mechanisms by showing increased sensory reliability in the attended modality. Collectively, our results demonstrate that attention impacts multisensory perceptual inference via modulation of reliability-weighed integration and selection of internal task-relevant representations.

Keywords

Multisensory integration, selective attention, causal inference, Bayesian, spatial localisation

3.1 Introduction

To form a solid representation of our world, it is essential to integrate sensory signals coming from a common source and to segregate signals from independent sources. In line with the principles of Bayesian Causal Inference (BCI), observers take into account the uncertainty about the causal structure of the environment (i.e. common or separate sources) in order to produce a final response (Körding et al., 2007; Rohe & Noppeney, 2015a, 2015b; Shams & Beierholm, 2010). Following the assumption of a common source, observers merge information from different senses weighted by their reliabilities (Alais & Burr, 2004; Ernst & Banks, 2002; Ernst & Bülthoff, 2004), such that more reliable (i.e. less variable or noisy) stimuli are attributed greater weights and thus bias behaviour to a greater extent.

Previous studies in unisensory contexts indicate that endogenous modality-specific attention (e.g. selectively attend to auditory stimuli) makes performance more efficient in the attended (e.g. auditory) relative to unattended (e.g. visual) modality. This is indexed by an increase of sensory discrimination accuracy and precision (Anton-Erxleben & Carrasco, 2013; Carrasco, 2011) and a decrease of response times (Boulter, 1977; Posner & Cohen, 1984; Spence & Driver, 1997; Spence et al., 2001). Thus, it is conceivable that attention modulates sensory reliability and thereby impacts sensory weighting during multisensory integration. However, evidence so far is mixed (Helbig & Ernst, 2008; Vercillo & Gori, 2015).

Furthermore, modality-specific response requirements (e.g. selectively respond to auditory stimuli) impact causal inference by encouraging observers to form a final perceptual estimate that is more biased towards the task-relevant modality (Aller & Noppeney, 2019; Cao et al., 2019; Odegaard et al., 2016; Rohe et al., 2019; Rohe & Noppeney, 2015a, 2016, 2018). In particular, the final response descends from a weighted average of the perceptual

estimate under one common source and the perceptual estimate in the task-relevant modality under two separate sources (according to the *model averaging* decision function of BCI, Rohe & Noppeney, 2015b). Overall, it is then plausible that modality-specific attention impacts multisensory perceptual inference via two complementary mechanisms: modulation of sensory reliability depending on pre-stimulus focus and selection of internal estimates based on task relevance. However, studies thus far have not been able to distinguish these two mechanisms, because they manipulated task-relevance only prior to stimulus presentation (Aller & Noppeney, 2019; Cao et al., 2019; Odegaard et al., 2016; Rohe et al., 2019; Rohe & Noppeney, 2015a, 2016, 2018).

The present study aimed to clarify whether and how endogenous modality-specific attention impacts multisensory perceptual inference, which was characterised in terms of localisation performance in a spatial ventriloquist paradigm. Crucially, we addressed the distinction between pre-stimulus focus and post-stimulus response selection. As recently discussed (Chun et al., 2011), the former reflects external attention (i.e. selection and modulation of sensory information), whereas the latter reflects internal attention (i.e. selection and modulation of internally generated representations). The two processes were orthogonally manipulated in an attention cueing paradigm (i.e. before stimuli presentation observers are cued to attend to vision or audition; after stimuli presentation they are cued to respond to vision or audition). On the one hand, we predicted pre-stimulus focus to change sensory reliability and thereby affect the relative weights for multisensory integration; on the other hand, we expected post-stimulus response selection to encourage the formation of a final perceptual estimate that is more biased towards the task-relevant sensory modality. Importantly, the orthogonal design allowed to contrast responses to attended versus unattended signals and thus expand previous work (Odegaard et al., 2016). Combining

psychophysics and computational modelling, we unveiled the computational principles driving the impact of attention on audio-visual spatial localisation.

3.2 Materials and methods

3.2.1 Participants

Thirty participants (10 males; mean age 22.7, range 18-32 years) were included in the experiment. Sample size was determined based on similar studies that targeted the role of attention in multisensory integration (Bertelson et al., 2000; Helbig & Ernst, 2008; Odegaard et al., 2016; Vercillo & Gori, 2015; Vroomen et al., 2001). Ten additional volunteers were excluded based on a priori exclusion criteria (see Section 3.2.5). All volunteers reported normal or corrected to normal vision, normal hearing and no history of neurological or psychiatric conditions. All volunteers provided written informed consent; they were naïve to the aim of the study; they received a reimbursement in the form of money or university credits for their participation in the experiment. The study was approved by the University of Birmingham Ethical Review Committee and was conducted in accordance with these regulations.

3.2.2 Stimuli

Stimuli were chosen based on previous work that investigated multisensory perceptual inference via the ventriloquist effect, using very similar experimental design and procedure (Aller & Noppeney, 2019; Rohe & Noppeney, 2015a, 2015b, 2016). The auditory stimulus consisted of a bursts of white noise (96,000 Hz sampling frequency; 65 dB sound pressure level; 5 ms on/off ramp) convolved with spatially-selective head-related transfer functions (HRTFs) based on the KEMAR dummy head of the MIT Media Lab48 (MIT Media

Laboratory, Gardner & Martin, 1995). HRTFs from the locations in the database were interpolated to obtain the locations required for the study. The visual stimulus consisted of a cloud of 20 white dots (luminance: 169 cd/m^2; dot diameter: 0.3° visual angle) sampled from a bivariate Gaussian distribution with a vertical standard deviation of 1° and a horizontal standard deviation of 5° presented on a grey background (17 cd/m^2). The white noise bursts and the cloud of white dots were generated independently for each experimental trial to prevent observers from learning non-specific cues.

3.2.3 Experimental design and procedure

The experiment combined spatial ventriloquism with a pre-/post-cueing attention paradigm (Figure 3.1A). Observers were pre-cued to attend to the auditory or visual modality (i.e. modality-specific attention). Next, they were presented with synchronous auditory and visual stimuli. Each stimulus was independently sampled from one of four positions along the azimuth (-9°, -3°, 3° or 9° visual angle), leading to four levels of audio-visual spatial disparity (0°, 6°, 12°, 18° visual angle). After stimuli presentation, observers were post-cued to report the perceived location of either the auditory or visual stimulus (i.e. modality-specific report). Hence, the task consisted of a 4 (Auditory location) × 4 (Visual location) × 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) factorial design. Crucially, the Attention × Report interaction generated valid and invalid attention conditions.

Every participant completed 10 runs over the course of two days (64 conditions × 3 trails / condition / run × 10 runs = 1920 trails in total). Each run was divided into 16 blocks of 12 trials. The experimental procedure was adapted from a very similar study that investigated the role of modality-specific attention in multisensory perceptual inference (Odegaard et al., 2016). At the beginning of each block (Figure 3.1B) a 2-second pre-cue (i.e. colour of the

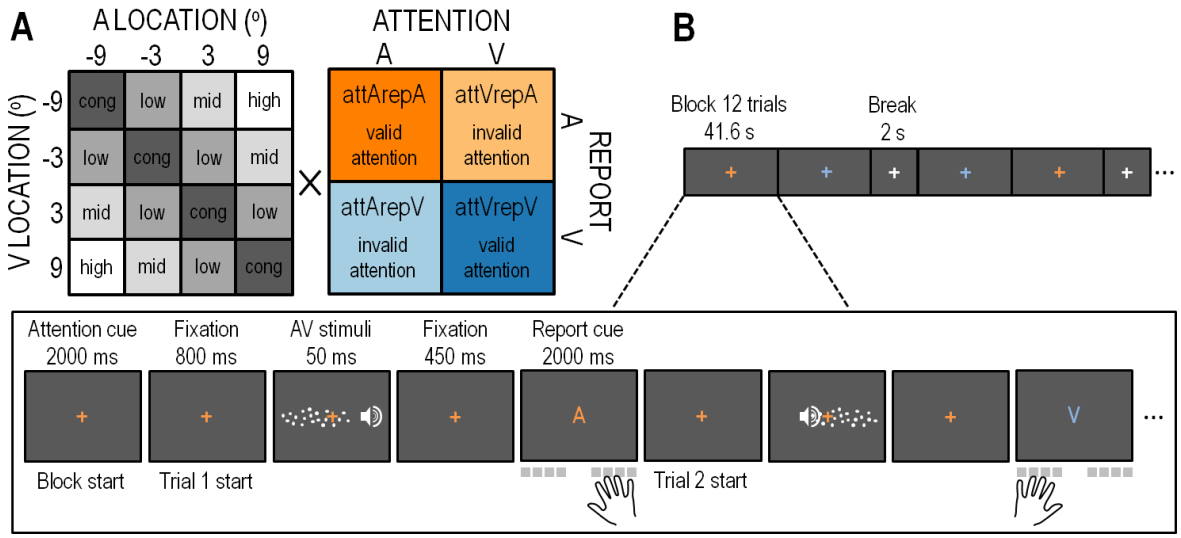

Figure 3.1: Experimental design and procedure

A) The experiment consisted of a 4 (Auditory location) x 4 (Visual location) x 2 (Attention: Auditory/Visual) x 2 (Report: Auditory/Visual) factorial design. Auditory and visual signals were independently sampled from 4 azimuthal locations (left hemispace: -9° and -3° visual angle; right hemispace: 3° and 9° visual angle), resulting in 16 AV spatial combinations with 4 levels of AV spatial disparity (cong: AV spatially congruent trials; low: 6° AV spatial disparity; mid: 12° AV spatial disparity; high: 18° AV spatial disparity). B) Experimental procedure: in a blocked design participants were cued before stimuli presentation (via colour of fixation cross) to attend to either the auditory or visual signal; on a trial-by-trial basis within each block, they were cued after stimuli presentation (via coloured letter: A for auditory; V for visual) to report their perceived auditory or visual location. They responded via button press with correspondent hand and key. A 2 s break period was inserted every two task blocks.

fixation cross) instructed participants to focus their attention on one sensory modality (e.g. to pay attention to the visual stimulus and to ignore the auditory stimulus). In each trial, after an 800 ms inter-trial interval, synchronous audio-visual spatial signals were presented for 50 ms. After a fixed 450 ms fixation interval, a post-cue (i.e. coloured letter) asked for the location of one of the two signals (i.e. “A” to locate the auditory stimulus; “V” to locate the visual stimulus) within a 2 seconds time interval. As for visual stimuli, subjects were instructed to consider the whole cloud of dots and estimate its middle point. While pre-stimulus focus was fixed within a block (with blocks’ order counterbalanced within and across participants), post-

stimulus response requests were pseudo-randomised within the block, with the constraint of (i) no more than 3 consecutive trials with the same post-cue and (ii) a 1:1 ratio of valid / invalid trials (i.e. report the location of the attended / unattended stimulus). We used these constraints to minimise two types of selection history effects (Awh et al., 2012; Theeuwes, 2018). In particular, the use of maximum 3 consecutive trials with the same target limited cumulative effects of inter-trial priming (e.g. Theeuwes & van der Burg, 2011); furthermore, the use of a 1:1 ratio of valid / invalid trials prevented participants from building response expectations prior to the appearance of the post-cue (e.g. Zuanazzi & Noppeney, 2018). A 2 s break period was inserted every two task blocks. Throughout the experiment, participants maintained their gaze on a fixation cross (1° diameter) in the centre of the screen. They were given two keypads, one per hand and sensory modality. Both keypads comprised four buttons, each mapping to one of the four possible stimuli positions along the azimuth. Participants reported their perceived location of the signal indicated by the post-cue as accurately as possible (within the 2 seconds time interval) using the corresponding keypad[1]. The mapping of hands (left/right), report modalities (auditory/visual) and cue colours (blue/yellow) was counterbalanced across participants. One preliminary practice run was used to familiarise participants with stimuli and procedure at the beginning of each testing day.

3.2.4 Experimental setup

The experiment was presented via Psychtoolbox version 3.0.11 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997) running under Matlab R2014a (The MathWorks, Inc.) on a Windows machine (Microsoft 7 2009). Auditory stimuli where presented via headphones (HD 280 PRO, Sennheiser, Wedemark-Wennebostel, Germany). Visual stimuli were

[1] A preliminary pilot study with 8 participants showed that a response period of 2 seconds was enough to provide an accurate response.

presented on a Gamma-calibrated LCD monitor (30" Dell UltraSharp U3014, USA; 2560 × 1600 pixels resolution; 60 Hz frame rate). Audio-visual synchrony was adjusted in the presentation software and confirmed by concurrently measuring auditory and visual onsets via a microphone and a photo-diode respectively. Participants sat in a dimly lit cubicle in front of the computer monitor at a viewing distance of 50 cm with their head positioned on a chin rest. Responses were collected via two keypads (Targus, USA), one per hand and report modality. Gaze position was monitored via Tobii Eyex eyetracking system (Tobii, Sweden).

3.2.5 Exclusion and inclusion criteria

Volunteers were post hoc excluded from the psychophysics analysis based on two criteria. Firstly, in a unisensory auditory or visual localization screening observers located either auditory or visual signals that were randomly presented at -9°, -3°, 3° or 9° visual angle along the azimuth. Participants completed 30 trials per condition (120 trials in total) for auditory and visual spatial localisation respectively, after being familiarized with stimuli and procedure via one preliminary practice run. Auditory and visual localization accuracies were quantified by the root-mean-square error (RMSE) between participants' reported location and signal's true location. Observers were excluded if their RMSE was greater than 5.5° for auditory localisation and 3.5° for visual localisation[2]. The analysis was limited to trials without missed or anticipated responses (i.e. no answer or response times < 100 ms respectively). A very limited number of trials were discarded both for auditory (across subjects mean ± SEM: 0.9% ± 0.3%) and visual (across subjects mean ± SEM: 0.8% ± 0.2%) localisation.

Secondly, observers were excluded if they did not show a significant cue validity effect (i.e. interaction between modality-specific attention and report) for response times in the

[2] Thresholds were defined as two standard deviations above the group mean RMSE (for auditory and visual localisation respectively) in a preliminary pilot study with 8 participants.

attention cueing paradigm. In other words, we expected observers to be significantly slower at reporting the location of unattended relative to attended stimuli (Donohue et al., 2015; Giessing et al., 2004; Natale et al., 2010). By assessing these attention shifting costs, the second criterion ensured that we included only volunteers who shifted their attention as instructed by the pre-cue (for follow-up analyses on included participants, see Sections 8.1.1 and 8.1.2). Analysis was limited to trials without missed, wrong or anticipated responses (i.e. no answer within 2s response time window, use of wrong keypad or response times < 100 ms respectively). A limited number of trials were discarded (across subjects mean ± SEM: 3.3% ± 0.6%).

3.2.6 Experimental data analysis

We employed psychophysics to test whether attention and report impact spatial localisation (model-free analysis). We further characterised the underlying computational principles by fitting three computational models to participants' localisation responses (model-based analysis). Analyses were limited to trials without missed, wrong or anticipated responses (i.e. no answer within 2 s response time window, use of wrong keypad or response times < 100 ms respectively). A limited number of trials were discarded (across subjects mean ± SEM: 3.3% ± 0.6%). Two-tailed p-values are reported for repeated-measures ANOVAs (Greenhouse-Geisser correction for violations of sphericity). When reporting simple contrasts, two-tailed parametric paired-sample t-tests are followed by two-tailed non-parametric Wilcoxon signed-ranks tests to account for occasional violations of normality assumptions. Bonferroni correction was used to account for multiple comparisons.

3.2.6.1 Eye movement analysis

We excluded trials without central fixation during stimuli presentation. Saccades were counted as significant eye movements if they fell outside a 1.3° circular area centred on subject's median of fixation, as defined in calibration trials (Blignaut, 2009). Participants successfully maintained fixation, with only a small number of rejected trials, in the ventriloquist paradigm (across subjects mean ± SEM: 1% ± 0.3%), unisensory auditory localisation (across subjects mean ± SEM: 0.9% ± 0.2%) and unisensory visual localisation (across subjects mean ± SEM: 0.4% ± 0.1%).

3.2.6.2 Model-free analysis: Audio-visual weight

For each participant and for each experimental trial where auditory and visual signals were spatially incongruent (i.e. AV spatial disparity greater than zero), we computed a measure called audio-visual weight (W_{AV}), which directly expresses the influence of the visual stimulus location (and complementarily, the influence of the auditory stimulus location) on the reported location. Thus, the W_{AV} index represents a quantitative measure of ventriloquist effect and it is defined as the distance between the reported location and the true auditory location, scaled by the distance between the true visual and auditory locations:

$$W_{AV} = \frac{Reported\ location - Auditory\ location}{Visual\ location - Auditory\ location}$$

A W_{AV} of 1 reflects full influence of the visual signal location on the localisation response (or in other words, no influence of the auditory signal location); a W_{AV} of 0 reflects no influence of the visual signal location on the localisation response (or in other words, full influence of the auditory signal location). We averaged the W_{AV} index across all combinations of AV locations at a particular level of AV spatial disparity and entered mean condition-specific W_{AV} for each participant into a 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual)

× 3 (AV spatial disparity: 6°, 12° or 18° visual angle, i.e. low, mid or high disparity) repeated measures ANOVA.

3.2.6.3 Model-free analysis: Response variance of spatially congruent trials

To further elucidate the influence of modality-specific attention, modality-specific report and spatial eccentricity on participants' localisation reliability (i.e. inverse of variance), we entered the standard deviation of responses for spatially congruent trials (i.e. AV spatial disparity equal to zero) into a 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 2 (Eccentricity: 3° or 9° visual angle, i.e. low or high eccentricity across sides) repeated measures ANOVA.

3.2.6.4 Model-based analysis: Bayesian Causal Inference

To unveil the computational principles underlying behaviour, we fitted three computational models to participants' localisation responses: (i) The Full Segregation model assumes independent processing of auditory and visual signals. (ii) The Forced Fusion model assumes mandatory integration of auditory and visual signals, each weighted by their respective reliabilities. (iii) The Bayesian Causal Inference model computes a final perceptual estimate of auditory (or visual) positions by averaging the spatial estimates under (i) and (ii) weighted by the posterior probabilities of their respective causal structure (i.e. we used the *model averaging* decision function, following Rohe & Noppeney, 2015b). Importantly, this step explicitly accounts for the effect of task relevance (see Section 2.2.1). The BCI model comprised the following free parameters: common-source prior p_{common} (binding tendency); spatial prior standard deviation σ_P (spread of the central bias); auditory standard deviation σ_A (inverse of auditory sensory reliability); visual standard deviation σ_V (inverse of visual sensory reliability). The Full Segregation and Forced Fusion models included a fixed

common-source prior of 0 and 1 respectively. For each model, parameters were fitted via simulation of 10,000 trials using MATLAB's *fminsearchbnd* function, which maximized the likelihood of the parameters after a preliminary grid search that refined the initial parameters space (for details of the generative model, see Körding et al., 2007). Both auditory standard deviation σ_A and visual standard deviation σ_V were fitted twice, separately for auditory and visual attention.

We first checked whether the BCI model outperformed the two alternative models in predicting participants' behaviour, as already shown in the past (Körding et al., 2007). In this way, we verified the presence of response selection based on task-relevance, as BCI is the only model that explicitly accounts for it. We compared the three models using the Bayesian Information Criterion (BIC) as an approximation to model evidence (Raftery, 1995). Importantly, the BIC depends both on model complexity and model fit, therefore establishing a fair comparison between models with a different number of parameters. For analysis at the group level, we applied both a fixed-effects approach (i.e. sum of individual BICs across subjects) and a random-effects approach (i.e. Bayesian model selection via SPM's *spm_BMS* function). Finally, we evaluated the effect of modality-specific attention on the sensory variance parameters of the winning model. After rejection of normality (Kolmogorov-Smirnov Test), non-parametric two-tailed Wilcoxon signed-ranks tests assessed pair-wise changes of σ_A and σ_V under auditory versus visual attention. We accounted for multiple comparisons via Bonferroni correction (α= 0.025).

3.3 Results

The present study evaluated whether and how endogenous modality-specific attention and report impact multisensory perceptual inference in a spatial ventriloquist paradigm. We

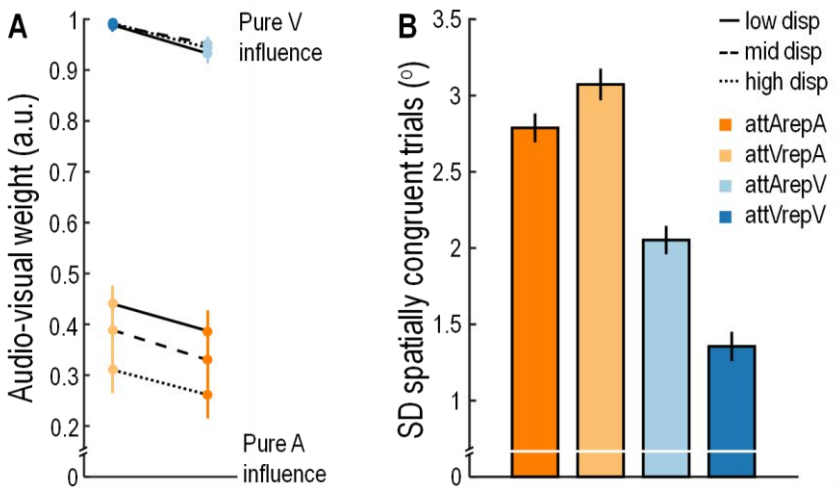

Figure 3.2: Model-free results

A) Mean (±SEM) audio-visual weight index W_{AV}: (Reported location – A location) ÷ (V location – A location) as a function of AV spatial disparity (Low/Mid/High: 6°/12°/18° visual angle), Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual). B) Mean (±SEM) standard deviation (SD) of responses for spatially congruent trials (i.e. AV spatial disparity equal to zero) in degrees of visual angle as a function of Attention and Report. A: auditory; V: visual.

employed psychophysics to test whether attention and report impact spatial localisation (model-free results). We further characterised the underlying mechanisms via computational modelling of participants' localisation responses (model-based results).

3.3.1 Model-free results: Audio-visual weight

The ventriloquist effect was quantified in terms of audio-visual weight (W_{AV}), which expresses the relative influence of visual and auditory signals on the reported location (W_{AV} = 1 reflects pure visual influence; W_{AV} = 0 reflects pure auditory influence). Results are shown in Figure 3.2A and summarised in Table 3.1. The 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 3 (AV spatial disparity: low/mid/high) repeated measures ANOVA with W_{AV} as dependent variable showed a significant main effect of Report ($F_{1,29}$ = 200.671, p <

0.001, $\eta p^2 = 0.874$), reflecting a greater W_{AV} for visual reports than auditory reports. In addition, we found a significant Report × AV disparity interaction ($F_{2,58} = 26.306$, $p < 0.001$, $\eta p^2 = 0.476$): the W_{AV} was smaller at higher than lower AV spatial disparities for auditory reports (high vs mid: $t_{29} = -6.317$, $p < 0.001$, Cohen's $d_{AV} = 0.288$, Wilcoxon signed-ranks $z = -4.288$, $p < 0.001$, $r = 0.554$; high vs low: $t_{29} = -6.202$, $p < 0.001$, Cohen's $d_{AV} = 0.554$, Wilcoxon signed-ranks $z = -4.165$, $p < 0.001$, $r = 0.538$; mid vs low: $t_{29} = -3.424$, $p = 0.002$, Cohen's $d_{AV} = 0.240$, Wilcoxon signed-ranks $z = -2.890$, $p = 0.004$, $r = 0.373$; Bonferroni-corrected $\alpha = 0.017$). In other words, the influence of visual signals on reported auditory locations decreased at higher AV spatial disparities, when signals are less likely to originate from a common source and the strength of multisensory interactions consequently decreases (Körding et al., 2007). Critically, we found a significant main effect of Attention ($F_{1,29} = 32.345$, $p < 0.001$, $\eta p^2 = 0.527$): the W_{AV} was greater under visual attention than auditory attention[3].

W_{AV} (a.u.) mean (±SEM)	**attArepA**	**attVrepA**	**attArepV**	**attVrepV**
Low disparity	0.387 (±0.040)	0.440 (±0.036)	0.933 (±0.019)	0.989 (±0.013)
Mid disparity	0.330 (±0.045)	0.389 (±0.046)	0.951 (±0.014)	0.989 (±0.008)
High disparity	0.261 (±0.046)	0.310 (±0.046)	0.944 (±0.014)	0.992 (±0.009)

Table 3.1: Audio-visual weight index (W_{AV})

Group mean (±SEM) as a function of AV spatial disparity (Low/Mid/High: 6°/12°/18° visual angle), Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

[3] As a sanity check, we verified that participants successfully located audio-visual congruent stimuli. Reported spatial locations were strongly correlated with the true audio-visual signals locations (across-participants mean ± SEM Fisher-z transformed Pearson correlation coefficient $z = 1.438$ (± 0.045), $p < 0.001$ for two-tailed one-sample Wilcoxon signed-ranks test against zero, after Fisher-z transformation of individual correlation coefficients).

3.3.2 Model-free results: Response variance of spatially congruent trials

To further elucidate the influence of modality-specific attention, modality-specific report and spatial eccentricity on sensory reliability, we examined response variance when observers were exposed to spatially congruent trials. Results are shown in Figure 3.2B and summarised in Table 3.2. The 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 2 (Eccentricity: Low/High) repeated measures ANOVA with the standard deviation of responses for spatially congruent trials as dependent variable showed a significant main effect of Report ($F_{1,29} = 72.373$, $p < 0.001$, $\eta p^2 = 0.714$): response variance was higher for auditory than visual reports. In other words, spatial localisation was more precise in the visual than auditory domain. Crucially, there was a significant Attention × Report interaction ($F_{1,29} = 32.484$, $p < 0.001$, $\eta p^2 = 0.528$): response variance decreased for valid versus invalid trials. Post-hoc t-tests (Bonferroni-corrected $\alpha = 0.025$) confirmed that the validity effect was significant both under auditory reports (valid vs invalid: $t_{29} = -3.344$, $p = 0.002$, Cohen's $d_{AV} = 0.431$, Wilcoxon signed-ranks $z = -2.910$, $p = 0.004$, $r = 0.376$) and visual reports (valid vs invalid: $t_{29} = -6.126$, $p < 0.001$, Cohen's $d_{AV} = 0.900$, Wilcoxon signed-ranks $z = -4.371$, $p < 0.001$, $r = 0.564$). In other words, response reliability increased for attended versus unattended stimuli, both for auditory and visual reports. Conversely, spatial eccentricity did not impact response variance ($p = 0.997$).

Response SD (°) mean (±SEM)	**attArepA**	**attVrepA**	**attArepV**	**attVrepV**
Low eccentricity	2.902 (±0.141)	3.202 (±0.139)	2.025 (±0.189)	1.138 (±0.174)
High eccentricity	2.674 (±0.175)	2.944 (±0.165)	2.080 (±0.144)	1.572 (±0.155)

Table 3.2: Response variance of spatially congruent trials

Group mean (±SEM) standard deviation (SD) as a function of Eccentricity (low: 3°, high: 9°), Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

3.3.3 Model-based results: Bayesian Causal Inference

Group summary statistics for each model and parameter (mean ± SEM) are reported in Table 3.3. First of all, Bayesian model comparison corroborated previous results (Körding et al., 2007) by revealing that the BCI model outperformed the Full Segregation and Forced Fusion models in predicting participants' localisation responses, both via fixed-effects analysis (highest sum of individual BICs across subjects) and random-effects analysis (highest protected exceedance probability, i.e. probability that a model is more likely than any other model, beyond differences due to chance). This result confirmed the presence of response selection based on task-relevance, as BCI is the only model that explicitly accounts for it.

Consequently, we evaluated the effect of modality-specific attention on the sensory variance parameters of the BCI model. The two-tailed Wilcoxon signed-ranks test contrasting auditory versus visual attention revealed that σ_A significantly decreased under auditory relative to visual attention (σ_A: $z = -4.165$, $p < 0.001$, $r = 0.538$) and σ_V significantly decreased under visual relative to auditory attention (σ_V: $z = -4.782$, $p < 0.001$, $r = 0.617$). In other words, sensory reliability increased for signals in the attended versus unattended sensory modality. In a follow-up investigation, we also fitted the common-source prior p_{common} and spatial prior standard deviation σ_P separately for auditory and visual attention. We found that

attention did not change such parameters, while the remaining results were virtually the same (and thus are not reported).

Mean (±SEM)	p_{common}	σ_P	σ_A (attA)	σ_A (attV)	σ_V (attA)	σ_V (attV)	**relBIC**	**pxp**
Bayesian Causal Inference	0.438 (±0.042)	15.443 (±1.714)	10.243 (±1.640)	12.460 (±1.701)	2.918 (±0.156)	2.166 (±0.111)	0	0.975
Forced Fusion	n/a	21.209 (±1.793)	10.614 (±0.910)	12.035 (±1.009)	6.344 (±0.303)	6.052 (±0.314)	1596.993	$1.00x10^{-6}$
Full segregation	n/a	13.139 (±1.255)	10.500 (±1.571)	12.443 (±1.518)	3.190 (±0.246)	2.149 (±0.119)	4126.234	0.025

Table 3.3: Model-based results

relBIC, Bayesian information criterion of a model summed over subjects (BIC = LL − 0.5 × P × ln(N), LL = log-likelihood, P = number of parameters, N = number of data points) relative to the BCI model (a model with smaller relBIC provides better data explanation); pxp, protected exceedance probability (probability that a model is more likely than the other models, beyond differences due to chance). attA: auditory attention; attV: visual attention.

3.4 Discussion

The extent to which multisensory perceptual inference is influenced by attention is still an open question (Helbig & Ernst, 2008; Macaluso et al., 2016; Odegaard et al., 2016; Rohe & Noppeney, 2016, 2018; Vercillo & Gori, 2015). The present study contributed to the debate by assessing whether and how endogenous modality-specific attention impacts audio-visual spatial localisation in a ventriloquist paradigm. Crucially, we dissociated pre-stimulus focus and post-stimulus response selection via orthogonal manipulation in a cueing paradigm: participants were pre-cued to attend to audition (or vision) and they were post-cued to report their perceived auditory (or visual) location. As a result, the effect of pre-stimulus focus on

sensory reliability could be dissociated from the effect of post-stimulus response selection on the final perceptual decision.

The audio-visual weight index revealed additive effects of pre-stimulus focus and post-stimulus response selection on audio-visual spatial inference. In agreement with previous evidence (Aller & Noppeney, 2019; Cao et al., 2019; Rohe et al., 2019; Rohe & Noppeney, 2015a, 2016, 2018), visual report (relative to auditory report) increased the extent to which the visual signal's location influenced participants' localisation responses. This result corroborates the view that modality-specific response requests encourage observers to segregate signals from different modalities and to select internal representations based on task relevance (Macaluso et al., 2016). Importantly, here we expand previous evidence by demonstrating that such effect is independent of pre-stimulus focus.

Crucially, visual attention (relative to auditory attention) additively increased the influence of visual signals on spatial localisation. This result suggests that pre-stimulus focus changed signals reliability and thereby affected the relative weights for multisensory integration prior to response selection (Macaluso et al., 2016). In line with this interpretation, the analysis of responses for spatially congruent trials revealed a decrease of variance (i.e. increase of reliability) for attended versus unattended stimuli. Importantly, these changes could not be explained by mere differences in spatial eccentricity (Charbonneau et al., 2013) and therefore they point to a pure attentional effect on spatial reliability[4]. Accordingly, fitting the Bayesian Causal Inference model (which outperformed alternative models in predicting participant's behaviour) revealed that modality-specific pre-stimulus focus increased sensory reliability in the attended relative to unattended modality. These results have direct functional implications for the ventriloquist effect: if the signal to be reported is more reliable due to

[4] Indeed, we used relatively central spatial positions in relation to the entire 40° visual angle field of view.

attentional focus, its perceived location is less attracted toward the position of a synchronous distractor and the final localisation response is more accurate. In accordance with this interpretation, previous research (Rohe & Noppeney, 2015b) has shown that physical reliability influences both implicit causal inference (i.e. spatial localisation) and explicit causal inference (i.e. common-source judgements): less integration takes place when the signal to be located is more reliable, as increases of spatial reliability sharpen the audio-visual spatial integration window; accordingly, common-source judgements decrease for spatially incongruent stimuli when sensory reliability increases. In this context, future studies should check whether modality-specific attention impacts explicit causal inference as well.

The present results partially corroborate a similar spatial ventriloquism study by Odegaard and colleagues (2016), which also showed an increase of visual reliability under visual attention, but did not show the same effect for the auditory modality. Importantly, the work of Odegaard and colleagues contrasted valid attention (e.g. attend to auditory modality, locate auditory stimulus) with divided attention (e.g. attend to visual and auditory modalities, locate auditory or visual stimulus); instead, the current study established the more extreme comparison between valid and invalid attention and allowed to uncover attentional effects on auditory reliability as well. In this context, it is worth noting qualitative differences in the magnitude of attentional modulation between vision and audition. In particular, the effect of attention on response variance of spatially congruent trials appeared stronger for visual than auditory reports (Figure 3.2B); similarly, the effect of attention was stronger on σ_V than σ_A in the fitted BCI model (Table 3.3). Differences in terms of task difficulty may provide an explanation. Given the well-known superiority of vision over audition in terms of spatial resolution (Freides, 1974), it is not surprising that in the present study participants were more precise during visual than auditory localisation, as indicated by lower response variance for

visual than auditory reports in spatially congruent trials (Figure 3.2B) and by lower σ_V than σ_A (combining attention conditions) in the fitted BCI model (Table 3.3). That is to say, task difficulty and therefore perceptual load (Lavie, 2005, 2010) were lower in vision than audition. The allocation of attention likely interacted with perceptual load (Lavie, 2005, 2010): when subjects were required to focus on the visual signal, there may have been residual attentional resources for paying attention to the concurrent auditory signal. As a consequence, auditory sensory reliability may have benefited from such attentional leak, thus decreasing the attended versus unattended difference. In the case of Odegaard and colleagues, auditory stimuli may have attracted more attentional resources to solve localisation across conditions, thus decreasing the difference between valid and divided attention. To put this hypothesis to test, future studies addressing the role of attention in multisensory perceptual inference may seek to balance task difficulty across vision and audition. A promising strategy would be to systematically decrease signal to noise ratio in the visual modality, given that sensory information must be comparably unreliable across modalities in order to boost attention-related effects on multisensory perceptual inference (Oruc et al., 2008).

Decreasing visual spatial reliability may also be useful to fully unveil additive effects of spatial disparity. Our current results show that the breakdown of integration was enhanced at higher audio-visual spatial discrepancies, when signals are less likely to originate from a common source, in line with the predictions of the BCI model (Körding et al., 2007) and in accordance with previous psychophysics (Rohe & Noppeney, 2015b) and neuroimaging (Aller & Noppeney, 2019; Rohe & Noppeney, 2015a, 2016) work. However, such disparity effect was significant under auditory reports only. A compatible qualitative pattern could be observed under visual report (Figure 3.2A), but the effect size was likely impacted by W_{AV} being already close to maximum. Therefore, future psychophysics work should corroborate

the present qualitative trend by better controlling for ceiling effects during visual spatial localisation.

Collectively, the current study elucidates how attentional control impacts multisensory perceptual inference. While pre-stimulus attentional focus modulates sensory reliability and thereby affects the relative weights for integration, post-stimulus selection of internal estimates biases responses towards task-relevant representations. Critically, it is still an open question how these distinct behavioural effects are implemented at the neural level. Hence, future neuroimaging research should aim to uncover the neural mechanisms behind the additive influence of pre-stimulus focus and post-stimulus response selection on multisensory perceptual inference. We hypothesise different effects along the visual and auditory dorsal cortical hierarchies, where distinct computational principles govern multisensory perceptual inference during audio-visual spatial localisation (Rohe & Noppeney, 2015a, 2016). On the one hand, it has been shown that low-level sensory areas encode modality-specific spatial representations, e.g. visual spatial locations in low-level visual areas (Rohe & Noppeney, 2015a, 2016). Pre-stimulus focus may act upon these representations in low-level sensory areas; in particular, it may express its effect on sensory reliability via changes of spatial representations' precision (Van Bergen et al., 2015). This may be implemented via internal noise reduction (Serences & Kastner, 2014), in line with attention-dependent sharpening of tuning functions (Martinez-Trujillo & Treue, 2004). On the other hand, post-stimulus response selection may impact the final perceptual response via higher-order association areas, which are known to compute flexible spatial representations based on task relevance (Aller & Noppeney, 2019; Cao et al., 2019; Rohe et al., 2019; Rohe & Noppeney, 2015a, 2016, 2018). This is in accordance with the presence of behaviourally relevant priority maps (Bisley & Goldberg, 2010; Sprague et al., 2018), i.e. maps that encode information relevant to

decision-making and response selection. Selective read-out mechanisms may tap into these priority maps and select internal representations from pools of neurons that encode task-relevant information (Pestilli et al., 2011; Serences & Kastner, 2014).

CHAPTER 4

ATTENTION IMPACTS MULTISENSORY PERCEPTUAL INFERENCE VIA DISTINCT COMPUTATIONAL PRINCIPLES ALONG THE CORTICAL HIERARCHIES

Ambra Ferrari, Uta Noppeney

Computational Cognitive Neuroimaging lab, Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, B15 2TT Birmingham, UK

Citation:

Ferrari, A. & Noppeney, U. (in preparation). Attention impacts multisensory perceptual inference via distinct computational principles along the cortical hierarchies.

Authors contributions:

Experiment conceptualisation and design: Ambra Ferrari, Uta Noppeney.
Data collection: Ambra Ferrari.
Data analysis: Ambra Ferrari (supervised by Uta Noppeney).
Writing: Ambra Ferrari (supervised by Uta Noppeney).

Abstract

In our natural environment we constantly process stimuli from different sensory modalities. Emerging evidence shows that multisensory interactions can be impacted by attentional control; however it is still a matter of debate whether attentional effects emerge already in early sensory areas or are deferred to higher-order association areas. The present fMRI study combined psychophysics and multivariate pattern decoding to characterize whether and how endogenous modality-specific attention impacts the weighting of audio-visual information to form spatial representations across the cortical hierarchies. At the behavioural level, we demonstrate that selection and modulation of sensory signals (i.e. pre-stimulus focus) affects the relative weights for multisensory integration via changes of sensory reliability, whereas selection of internal estimates (i.e. post-stimulus response selection) biases responses towards task-relevant representations. At the neural level, we show distinct effects of attentional control along the cortical hierarchies: in low-level visual areas, pre-stimulus focus biases the competition among spatial representations towards attended signals; in higher-order association areas, post-stimulus response selection biases the competition among spatial representations towards task-relevant signals. Collectively, the present study confirms that attentional control over multisensory interactions is pervasive along the cortical hierarchies, but is driven by distinct computational principles.

Keywords

Multisensory integration, selective attention, causal inference, Bayesian, spatial localisation, fMRI, cortical hierarchies

4.1 Introduction

Effective interactions with our complex multisensory world require the integration of signals coming from a common source and segregation of signals from different sources. Thus, the brain has to perform two computational tasks. First, it needs to compute the probability of common on independent sources based on current sensory evidence and prior binding knowledge (Körding et al., 2007; Shams & Beierholm, 2010), such as temporal synchrony (Lee & Noppeney, 2011a; Lewis & Noppeney, 2010; Magnotti et al., 2013; Maier et al., 2011; Munhall et al., 1996; Noesselt et al., 2007; Parise & Ernst, 2016; Parise et al., 2012; van Wassenhove et al., 2007) and spatial disparity (Lewald & Guski, 2003; Slutsky & Recanzone, 2001; Spence, 2013). Second, the brain needs to combine signals from a common source into the most precise representation via reliability-weighted integration, such that each component is weighted in proportion to its relative reliability, which is the inverse of sensory variance or noise (Alais & Burr, 2004; Ernst & Banks, 2002; Hillis et al., 2004). It has been a matter of debate whether reliability-weighted integration is automatic or whether instead it can be impacted by endogenous modality-specific attention (Helbig & Ernst, 2008; Vercillo & Gori, 2015). Emerging evidence (Ferrari & Noppeney, in preparation) suggests that valid attentional focus boosts the reliability of attended information and thereby increases the correspondent weight during integration. In addition, in line with Bayesian Causal Inference, modality-specific task contexts (e.g. selectively respond to auditory stimuli) impact causal inference by encouraging observers to form a final perceptual estimate that is more biased towards the task-relevant modality (Aller & Noppeney, 2019; Cao et al., 2019; Ferrari & Noppeney, in preparation; Odegaard et al., 2016; Rohe et al., 2019; Rohe & Noppeney, 2015a, 2016, 2018).

This raises the critical question of how such distinct influences are implemented at the neural level. Two different mechanisms are conceivable. First, directing attention to one particular sensory (e.g. visual) modality may increase the precision of the spatial representations in the corresponding early sensory (e.g. visual) cortices (Martinez-Trujillo & Treue, 2004; Serences & Kastner, 2014; Van Bergen et al., 2015) and thereby increase their influence on 'classic' multisensory integration areas such as parietal cortices. Second, posterior parietal cortices may form spatial estimates as predicted by Bayesian Causal Inference, namely averaging the forced fusion estimate with the full segregation estimate of the task-relevant sensory modality, each weighted by the probability of the respective causal structure (Rohe & Noppeney, 2015a, 2016). Consistent with this conjecture, there is evidence of flexible priority maps in parietal cortices (Bisley & Goldberg, 2010; Sprague et al., 2018), which allow selective read-out of representations according to task relevance (Pestilli et al., 2011; Serences & Kastner, 2014). Critically, past neuroimaging studies (Aller & Noppeney, 2019; Cao et al., 2019; Rohe et al., 2019; Rohe & Noppeney, 2015a, 2016) were not able to distinguish between these two fundamental neural mechanisms, because they manipulated the task-relevance of the sensory modality only prior to stimulus presentation. In fact, this could have affected both the reliability of sensory representations in early sensory cortices and the selective read-out in higher-order parietal areas.

The present functional magnetic resonance imaging (fMRI) study dissociated these two neural mechanisms by combining spatial ventriloquism with a novel attention cueing paradigm. Participants were presented with synchronous audio-visual spatial signals of variable spatial disparity. A pre-cue indicated the sensory modality that needed to be attended, a post-cue whether the auditory or visual location needed to be reported. Consequently, we addressed how endogenous modality-specific attention and report impact the formation of

perceptual and neural audio-visual spatial representations along the dorsal sensory cortical hierarchies. Critically, while pre-stimulus focus (i.e. Attention) may alter the reliability of sensory presentations in early sensory cortices, post-stimulus response selection (i.e. Report) would define selective read-out in association cortices.

4.2 Materials and methods

4.2.1 Participants

Twenty-seven participants (10 males; mean age 20.5, range 18-30 years) were included in the psychophysics experiment based on a priori power analysis (G*Power 3.1, Faul et al., 2007, 2009) with power (1-β) = 0.8, α = 0.05 and effect size Cohen's d_{AV} = 0.5. Estimation of effect size was derived from a one-tailed paired sample t-test that tested the effect of modality-specific attention on participants' localisation responses (in terms of audio-visual weight index, W_{AV}) in a previous study with a very similar experimental design (Ferrari & Noppeney, in preparation). Eight additional volunteers were excluded based on a priori exclusion criteria (see Section 4.2.6). All volunteers reported normal or corrected to normal vision, normal hearing and no history of neurological or psychiatric conditions.

Twelve participants of the psychophysics experiment took part in the subsequent fMRI study. The fMRI sample size was determined based on previous neuroimaging experiments that used similar experimental designs and analysis approaches (Aller & Noppeney, 2019; Rohe & Noppeney, 2015a, 2016). Participants included in the fMRI study (5 males; mean age 21.67 years, range 18-30 years) were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971; mean laterality index: 88.64; range: 60–100). All volunteers provided written informed consent and were naïve to the aim of the study; they received a

reimbursement in the form of money or university credits for their participation in the experiment. The study was approved by the University of Birmingham Ethical Review Committee and was conducted in accordance with these regulations.

4.2.2 Stimuli

The auditory stimulus consisted of a bursts of white noise (96,000 Hz sampling frequency; 65 dB sound pressure level; 5 ms on/off ramp) convolved with spatially-selective head-related transfer functions (HRTFs) based on the KEMAR dummy head of the MIT Media Lab48 (MIT Media Laboratory, Gardner & Martin, 1995). HRTFs from the locations in the database were interpolated to obtain the locations required for the study. The visual stimulus consisted of a cloud of 20 white dots (luminance: 169 cd/m^2; dot diameter: 0.3° visual angle) sampled from a bivariate Gaussian distribution with a vertical standard deviation of 1° and a horizontal standard deviation of 5° presented on a grey background (17 cd/m^2). White noise bursts and clouds of dots were generated independently for each experimental trial to prevent observers from learning non-specific cues.

4.2.3 Experimental design and procedure

We used the same experimental design and procedure for the psychophysics and the fMRI experiment. The experiment combined spatial ventriloquism with a pre-/post-cueing attention paradigm (Figure 4.1A). Observers were pre-cued to attend to the auditory or visual modality (i.e. modality-specific attention). Next, they were presented with synchronous auditory and visual stimuli. Each stimulus was independently sampled from one of three positions along the azimuth (-9°, 0° or 9° visual angle), leading to three levels of audio-visual spatial disparity (0°, 9°, 18° visual angle). After stimulus presentation, observers were post-cued to report the

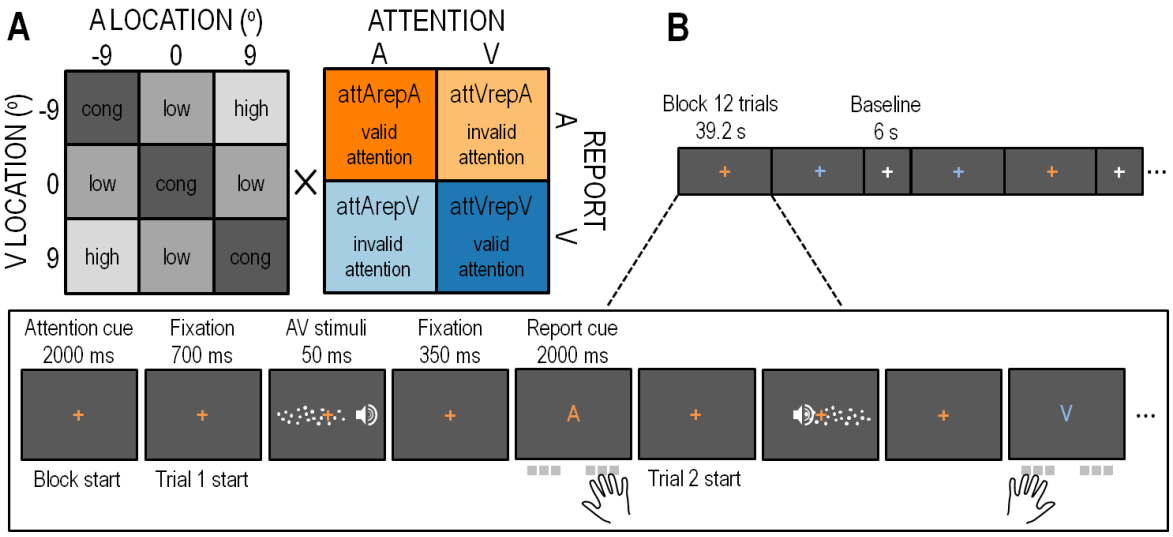

Figure 4.1: Experimental design and procedure

A) The experiment consisted of a 3 (Auditory location) x 3 (Visual location) x 2 (Attention: Auditory/Visual) x 2 (Report: Auditory/Visual) factorial design. Auditory and visual signals were independently sampled from 3 azimuthal locations (left hemispace: -9°; centre: 0°; right hemispace: 9° visual angle), resulting in 9 AV spatial combinations with 3 levels of AV spatial disparity (cong: AV spatially congruent trials; low: 9° AV spatial disparity; high: 18° AV spatial disparity). B) Experimental procedure: in different blocks participants were cued before stimuli presentation (via colour of fixation cross) to attend to either the auditory or visual signal; on a trial-by-trial basis within each block, they were cued after stimuli presentation (via coloured letter: A for auditory; V for visual) to report their perceived auditory or visual location. They responded via button press with correspondent hand and key. A 6 s baseline period was inserted every two task blocks.

perceived location of either the auditory or visual stimulus (i.e. modality-specific report). Hence, the task consisted of a 3 (Auditory location) × 3 (Visual location) × 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) factorial design. Crucially, the Attention × Report interaction generated valid and invalid attention conditions.

For the psychophysics experiment, every participant completed 3 runs in one day (36 conditions × 6 trails / condition / run × 3 runs = 648 trails in total). For the fMRI experiment, every participant completed 14 scanning runs over the course of four days (36 conditions × 6 trails / condition / run × 14 runs = 3024 trails in total). Each run was divided into 18 task

blocks (12 trials / block) and 9 fixation blocks (we presented one fixation block every two task blocks). At the beginning of each task block (Figure 4.1B) a 2-second pre-cue (i.e. colour of the fixation cross) instructed participants to focus their attention on one sensory modality (e.g. to pay attention to the visual stimulus and to ignore the auditory stimulus). In each trial, after a 700 ms inter-trial interval, synchronous audio-visual spatial signals were presented for 50 ms. After a fixed 350 ms fixation interval, a post-cue (i.e. coloured letter) asked for the location of one of the two signals (i.e. "A" to locate the auditory stimulus; "V" to locate the visual stimulus) within a 2 seconds time interval. As for visual stimuli, subjects were instructed to consider the whole cloud of dots and estimate its middle point. To increase design efficiency, auditory and visual spatial positions were sampled in a pseudo-randomized fashion, creating mini-blocks of 3, 2 or 1 trials with the same AV spatial combination. While pre-stimulus focus was fixed within a block (with blocks' order counterbalanced within and across participants), post-stimulus response requests were pseudo-randomised within the block, with the constraint of (i) no more than 3 consecutive trials with the same post-cue and (ii) a 1:1 ratio of valid/invalid trials (i.e. respond to the attended/unattended stimulus). We used these constraints to minimised two types of selection history effects (Awh et al., 2012; Theeuwes, 2018). In particular, the use of maximum 3 consecutive trials with the same target limited cumulative effects of inter-trial priming (e.g. Theeuwes & van der Burg, 2011); furthermore, the use of a 1:1 ratio of valid / invalid trials prevented participants from building response expectations prior to the appearance of the post-cue (e.g. Zuanazzi & Noppeney, 2018). Throughout the task, participants maintained their gaze on a fixation cross (1° diameter) in the centre of the screen. Participants were given two keypads, one per hand and sensory modality. On each keypad, a specific key corresponded to one of the three possible stimuli positions along the azimuth. Participants reported their perceived location of the signal

indicated by the post-cue as accurately as possible (within the 2 seconds time interval) using the corresponding keypad[1]. The mapping of hands (left/right), report modalities (auditory/visual) and colours (blue/yellow) was counterbalanced within subjects across MRI days. At the beginning of each day (both for psychophysics and MRI), participants were familiarized with stimuli and procedure via one preliminary practice run.

Besides the main experiment, we also verified whether participants successfully located unisensory auditory stimuli inside the scanner despite MR scanner noise. Please refer to Section 8.2.5 for paradigm description, analyses and results.

4.2.4 Experimental setup

4.2.4.1 Psychophysics experiment

The experiment was presented via Psychtoolbox version 3.0.11 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) running under MATLAB R2014a (MathWorks Inc.) on a Windows machine (Microsoft 7 2009). Auditory stimuli were played using headphones (HD 280 PRO, Sennheiser, Wedemark-Wennebostel, Germany). Visual stimuli were presented on a Gamma-calibrated LCD monitor (30” Dell UltraSharp U3014, USA; 2560 × 1600 pixels resolution; 60 Hz frame rate). We adjusted audio-visual latencies in the presentation software and confirmed their synchrony by recording and measuring their relative latencies using a microphone and a photo-diode. To mimic the perceptual environment to the fMRI experiment, scanner noise was played at 80 dB SPL through external loudspeakers positioned at each side of the monitor. Participants sat in a dimly lit cubicle in front of the computer monitor at a viewing distance of 50 cm with their head positioned on a chin rest. They gave responses via two

[1] A preliminary pilot study with 8 participants showed that a response period of 2 seconds was enough to provide an accurate response.

keypads (Targus, USA), one per hand and report modality. Gaze position was monitored via Tobii Eyex eyetracking system (Tobii, Sweden).

4.2.4.2 fMRI experiment

The experiment was presented via Psychtoolbox version 3.0.11 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) running under MATLAB R2011b (MathWorks Inc.) on a MacBook Pro (Mac OSX 10.6.8). Auditory stimuli were played using MR-compatible headphones (MR Confon HP-VS03, UK). Visual stimuli were back-projected onto a Plexiglas screen using a Barco Present-C F-Series projector (F35 WUXGA, UK; 1920 × 1024 pixels resolution; 60 Hz frame rate) and they were visible to the participants via a mirror mounted on the MR head-coil (horizontal visual field of ~40° visual angle at a viewing distance of ~68 cm). Participants gave responses via two MR-compatible keypads (NATA LXPAD 1×5-10M, BC Canada), one per hand and report modality.

4.2.5 MRI data acquisition

A 3T Philips Achieva MR scanner was used to acquire both a T1-weighted anatomical image (TR = 8400 ms, TE = 3.8 ms, flip angle = 8°, FOV = 288 mm × 232 mm, 175 sagittal slices acquired in sequential ascending direction, voxel size = 1 × 1 × 1 mm^3) and T2*-weighted axial echoplanar images (EPI) with blood-oxygenation-level-dependent contrast (gradient echo, TR = 2800 ms, TE = 40 ms, flip angle = 90°, FOV = 192 × 192 × 114 mm^2, 38 axial slices acquired in sequential ascending direction, voxel size = 2.5 × 2.5 × 2.5 mm^3 + 0.5 mm interslice gap). For each participant, a total of 276 volumes × 14 runs were acquired, after discarding the first four volumes of each run to allow for T1 equilibration effects. Functional data acquisition was performed over the course of 4 days and the anatomical image was acquired at the end of the last day.

4.2.6 Exclusion and inclusion criteria

4.2.6.1 Psychophysics experiment

Volunteers were post hoc excluded from the psychophysics analysis based on two criteria. Firstly, in a unisensory auditory or visual localization screening observers located either auditory or visual signals that were randomly presented at -9°, 0° or 9° visual angle along the azimuth. Participants completed 30 trials per condition (90 trials in total) for auditory and visual spatial localisation respectively, after being familiarized with stimuli and procedure via one preliminary practice run. Auditory and visual localization accuracies were quantified by the root-mean-square error (RMSE) between participants' reported location and signal's true location. Observers were excluded if their RMSE was greater than 5.5° for auditory localisation and 3.5° for visual localisation[2]. The analysis was limited to trials without missed or anticipated responses (i.e. no answer or response times < 100 ms respectively). A very limited number of trials were discarded both for auditory (across subjects mean ± SEM: 0.6% ± 0.3%) and visual (across subjects mean ± SEM: 1.6% ± 0.4%) localisation.

Secondly, observers were excluded if they did not show a significant cue validity effect (i.e. interaction between modality-specific attention and report) for response times in the attention cueing paradigm. In other words, we expected observers to be significantly slower at reporting the location of unattended stimuli relative to attended stimuli (Donohue et al., 2015; Giessing et al., 2004; Natale et al., 2010). By assessing these attention shifting costs, the second criterion ensured that we included only volunteers who shifted their attention as instructed by the pre-cue. Analysis was limited to trials without missed, wrong or anticipated responses (i.e. no answer within 2 s response time window, use of wrong keypad or response

[2] Thresholds were defined as two standard deviations above the group mean RMSE (for auditory and visual localisation respectively) in a preliminary pilot study with 8 participants.

times < 100 ms respectively). A limited number of trials were discarded (across subjects mean ± SEM: 3.4% ± 0.7%).

4.2.6.2 fMRI experiment

Participants of the psychophysics study were eligible for the subsequent fMRI experiment if they maintained central fixation throughout each run. We defined saccades as eye movements outside 1.3° circular area centred on subject's median of fixation based on calibration trials (Blignaut, 2009). Only participants who produced less than 20 saccades per run (i.e. 216 trials; threshold defined as two standard deviations above the group mean in a previous study, Ferrari & Noppeney, in preparation), were eligible to the fMRI experiment until we reached a pre-defined sample size (see Section 4.2.1).

4.2.7 Experimental data analysis

4.2.7.1 Psychophysics and fMRI experiments: behavioural analysis

Analyses were limited to trials without missed, wrong or anticipated responses (i.e. no answer within 2 s response time window, use of wrong keypad or response times < 100 ms respectively). A limited number of trials were discarded both for the psychophysics experiment (across subjects mean ± SEM: 3.4% ± 0.7%) and the fMRI experiment (across subjects mean ± SEM: 3.0% ± 1.0%). For psychophysics, we excluded trials without central fixation during stimuli presentation. Saccades were counted as significant eye movements if they fell outside a 1.3° circular area centred on subject's median of fixation, as defined in calibration trials (Blignaut, 2009). Participants successfully maintained fixation with only a small number of rejected trials (across subjects mean ± SEM: 0.4% ± 0.1%).

For each participant and for each experimental trial where auditory and visual signals were spatially incongruent (i.e. AV spatial disparity greater than zero), we computed a measure called audio-visual weight (W_{AV}), which directly expresses the influence of the visual stimulus location (and complementarily, the influence of the auditory stimulus location) on the reported location. Thus, the W_{AV} index represents a quantitative measure of ventriloquist effect and it is defined as the distance between the reported location and the true auditory location, scaled by the distance between the true visual and auditory locations:

$$W_{AV} = \frac{Reported\ location - Auditory\ location}{Visual\ location - Auditory\ location}$$

A W_{AV} of 1 reflects full influence of the visual signal location on the localisation response (or in other words, no influence of the auditory signal location); a W_{AV} of 0 reflects no influence of the visual signal location on the localisation response (or in other words, full influence of the auditory signal location). We averaged the W_{AV} index across all combinations of AV locations at a particular level of AV spatial disparity and entered mean condition-specific W_{AV} for each participant into a 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 2 (AV spatial disparity: 9° or 18° visual angle, i.e. low or high disparity) repeated measures ANOVA. Two-tailed p-values are reported for repeated-measures ANOVAs (Greenhouse-Geisser correction for violations of sphericity). When reporting simple contrasts, two-tailed parametric paired-sample t-tests are followed by two-tailed non-parametric Wilcoxon signed-ranks tests to account for occasional violations of normality assumptions. Bonferroni correction was used to account for multiple comparisons.

4.2.7.2 fMRI experiment: univariate analysis

MRI data were analysed using SPM12 (Wellcome Department of Imaging Neuroscience, London; www.fil.ion.ucl.ac.uk/spm; Friston et al., 1994a). Scans from each participant were

realigned (using the first scan as reference) and unwarped, slice time corrected to the central slice, spatially normalised into Montreal Neurological Institute (MNI) space using normalisation parameters from segmentation of the T1 structural image (Ashburner & Friston, 2005), resampled to a spatial resolution of 2 × 2 × 2 mm^3 and spatially smoothed with a Gaussian kernel of 8 mm full-width at half-maximum. A high-pass filter (1/128 Hz cutoff) was applied to the time series in each voxel.

In an event-related design, unit impulses representing stimuli onsets were convolved with a canonical hemodynamic response function and its first temporal derivative. The 36 experimental conditions of the 3 (Auditory location) × 3 (Visual location) × 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) factorial design were included as regressors in the design matrix. Onsets of the blocked pre-cues were included as separate regressors. Realignment parameters were also added as nuisance covariates to account for noise due to residual head motion artefacts. The voxel-wise magnitude of the BOLD signal in response to the audio-visual onsets was defined by the parameter estimates pertaining to the canonical hemodynamic response function. Following a hierarchical summary statistics approach, subject-specific images were entered into a first-level general linear model and contrasts (each experimental condition versus baseline summed over the fourteen runs) were passed to a second-level ANOVA, were contrasts of interest were defined. Following random effect analysis, inferences were made at the second level (Friston et al., 1994a).

The univariate analysis aimed to check whether participants were appropriately engaged with task requests, alongside response time and response errors effects (see Sections 8.2.1 and 8.2.2). At the group level, we pulled over left and right stimuli locations, then we tested for (in)validity effects (i.e. Attention × Report interaction [attVrepA & attArepV] > [attArepA & attVrepV] and vice-versa) and AV spatial (in)congruence effects (i.e. AVincongruent >

AVcongruent and vice-versa). For completeness, we also evaluated the main effect of Attention (i.e. attA > attV and vice-versa) and the main effect of Report (i.e. repA > repV and vice-versa). Whole-brain activations are reported at $p < 0.05$ (Family-Wise Error corrected) at the peak level (Friston et al., 1994b).

4.2.7.3 fMRI experiment: multivariate decoding analysis

Scans from each participant were realigned (using the first scan as reference), unwarped and slice time corrected to the central slice. A high-pass filter (1/128 Hz) was applied to the time series in each voxel. Unsmoothed images in participants' native space were entered into the same first-level design matrix specified for univariate analysis. The voxel-wise magnitude of the BOLD signal in response to the audio-visual onsets was defined by the parameter estimates pertaining to the canonical hemodynamic response function. First-level beta images were masked with a priori anatomically defined regions of interest (ROIs) along the visual and auditory dorsal sensory cortical hierarchies (see Section 4.2.7.4). The resulting spatial activation patterns were scaled to the range 0-1. Multivariate decoding was performed using The Decoding Toolbox 3.96 (TDT, Hebart et al., 2015). For each participant and ROI along the auditory and visual hierarchies, a linear support-vector regression model (SVR, as implemented in LIBSVM 3.17; Chang & Lin, 2011) was trained to learn the mapping from the fMRI activation patterns to the audio-visual congruent spatial locations from all but one run. Subsequently, it was asked to decode the spatially congruent and incongruent audio-visual spatial locations of the remaining run (Figure 4.2). Following a leave-one-run-out cross-validation scheme, this procedure was repeated for all runs.

The aim of our multivariate decoding analysis was two-fold. First, we investigated whether we could successfully decode spatial information in each ROI by evaluating the

Pearson correlation coefficient between decoded locations from fMRI activation patterns and true signals' locations in audio-visual spatially congruent conditions (which are non-ambiguous). We tested whether each ROI's Pearson correlation coefficient was significantly different from zero via two-tailed one-sample Wilcoxon signed-ranks test.

Second, we investigated how each ROI weights auditory and visual signals sampled from incongruent locations to form spatial representations. Using decoded spatial locations from spatially incongruent conditions, we computed a neural audio-visual weight (nW_{AV}) index, which directly expresses the influence of the visual stimulus location (and complementarily, the influence of the auditory stimulus location) on the decoded location. Similarly to behavioural analysis, the nW_{AV} index represents a quantitative measure of ventriloquist effect and it is defined as the distance between the decoded location and the true auditory location, scaled by the distance between the true visual and auditory locations:

$$nW_{AV} = \frac{Decoded\ location - Auditory\ location}{Visual\ location - Auditory\ location}$$

An nW_{AV} of 1 reflects full influence of the visual signal location on the decoded location (or in other words, no influence of the auditory signal location); an nW_{AV} of 0 reflects no influence of the visual signal location on the decoded location (or in other words, full influence of the auditory signal location). We averaged the nW_{AV} index across all combinations of AV locations at a particular level of AV spatial disparity and entered mean condition-specific nW_{AV} for each participants into a 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 2 (AV spatial disparity: 9° or 18° visual angle, i.e. low or high disparity) repeated measures ANOVA. Two-tailed p-values are reported for repeated-measures ANOVAs (Greenhouse-Geisser correction for violations of sphericity). When reporting simple contrasts, two-tailed parametric t-tests are followed by two-tailed non-

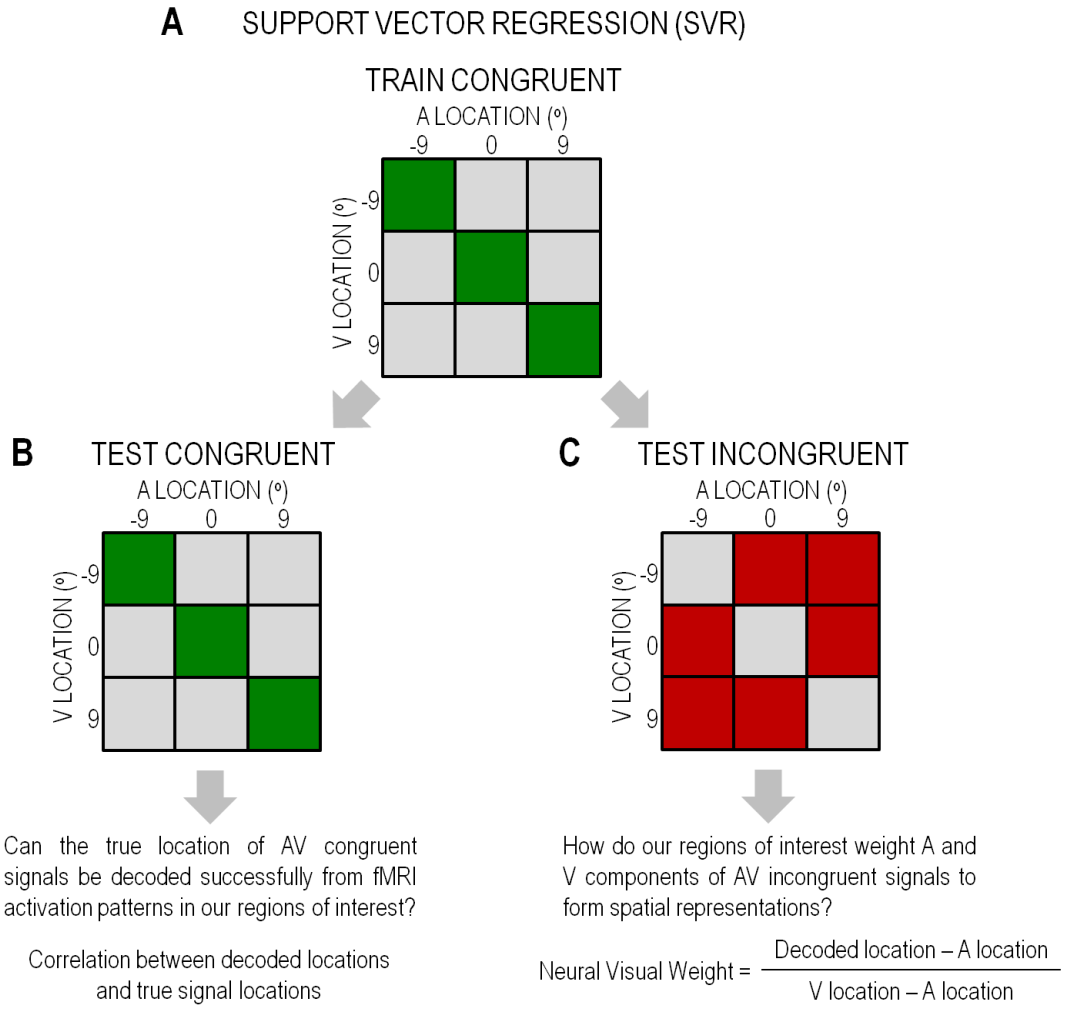

Figure 4.2: Overview of multivariate decoding analysis

A) Neural spatial estimates were obtained by training a SVR model to learn the mapping from fMRI activation patterns (beta images) to external spatial locations for audio-visual congruent conditions (green boxes). This learnt mapping was then used to decode spatial locations from fMRI activation patterns of B) the spatially congruent conditions in a leave-one-run-out cross-validation scheme and C) the spatially incongruent conditions (red boxes) in a leave-one-run-out generalisation scheme.

parametric Wilcoxon signed-ranks tests to account for potential violation of normality assumptions. Bonferroni correction was used to account for multiple comparisons.

4.2.7.4 Regions of interest definition

Based on previous studies (Rohe & Noppeney, 2015a, 2016, 2018), we focused our analysis on a specific set of visual and auditory ROIs along the dorsal sensory hierarchies (Goodale & Milner, 1992; Rauschecker & Tian, 2000; Rauschecker, 2018): low-level visual cortex (V1-3), posterior intraparietal sulcus (pIPS), anterior intraparietal sulcus (aIPS), low-level auditory cortex (A1-2), planum temporale (PT). Visual ROIs were defined using volume-based full-probability maps (threshold: 80 percentile) from a probabilistic atlas for visual topography (Wang et al., 2015). Low-level visual cortex comprised ventral and dorsal areas V1-3; posterior intraparietal sulcus comprised areas IPS0, IPS1 and IPS2; anterior intraparietal sulcus comprised areas IPS3, IPS4, SPL1 (Swisher et al., 2007). Low-level auditory cortex (A1-2) comprised areas TE1.0, TE1.1 and TE1.2 from the Anatomy Toolbox (Eickhoff et al., 2005). Planum temporale was defined using the corresponding region from the Destrieux atlas of Freesurfer 5.3.0 (Destrieux et al., 2010). All regions of interest were defined bilaterally, i.e. via combination of corresponding areas from left and right hemispheres.

4.3 Results

Firstly, we report the behavioural results of the psychophysics and the fMRI experiment, where we tested the effect of attention and report on the ventriloquist effect. Secondly, we report the multivariate decoding and univariate results of the fMRI experiment, where we addressed the respective neural mechanisms.

4.3.1 Psychophysics and fMRI experiments: behavioural results

The ventriloquist effect was quantified in terms of audio-visual weight (W_{AV}), which expresses the relative influence of visual and auditory signals on the reported location (W_{AV} =

1 reflects pure visual influence; W_{AV} = 0 reflects pure auditory influence). Results of psychophysics and fMRI experiments are shown in Figure 4.3 and summarised in Table 4.1. For both experiments, the 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 2 (AV spatial disparity: low/high) repeated measures ANOVA with W_{AV} as dependent variable showed a significant main effect of Report (psychophysics: $F_{1,26} = 273.384$, $p < 0.001$, $\eta p^2 = 0.913$; fMRI: $F_{1,11} = 172.725$, $p < 0.001$, $\eta p^2 = 0.940$), reflecting a greater W_{AV} for visual reports than auditory reports. In addition, we found a significant Report × AV disparity interaction (psychophysics: $F_{1,26} = 83.030$, $p < 0.001$, $\eta p^2 = 0.762$; fMRI: $F_{1,11} = 143.284$, $p < 0.001$, $\eta p^2 = 0.929$): the W_{AV} was smaller at high than low AV spatial disparities for auditory reports (psychophysics: $t_{26} = -9.113$, $p < 0.001$, Cohen's $d_{AV} = 0.940$, Wilcoxon signed-ranks $z = -4.469$, $p < 0.001$, $r = 0.608$; fMRI: $t_{11} = -11.545$, $p < 0.001$, Cohen's $d_{AV} = 1.032$, Wilcoxon signed-ranks $z = -3.059$, $p < 0.001$, $r = 0.624$). In other words, the influence of visual signals on reported auditory locations decreased at higher AV spatial disparities, when signals are less likely to originate from a common source and therefore multisensory interactions are weakened (Körding et al., 2007). Critically, we found a significant main effect of Attention (psychophysics: $F_{1,26} = 5.933$, $p = 0.022$, $\eta p^2 = 0.186$; fMRI: $F_{1,11} = 8.477$, $p = 0.014$, $\eta p^2 = 0.435$): the W_{AV} was greater under visual attention than auditory attention[3].

[3] As a sanity check, we verified that participants successfully located audio-visual congruent stimuli. Reported spatial locations were strongly correlated with the true audio-visual signals location (psychophysics: across-participants mean ± SEM Fisher-z transformed Pearson correlation coefficient $z = 2.128$ (± 0.070), $p < 0.001$ for two-tailed one-sample Wilcoxon signed-ranks test against zero, after Fisher-z transformation of individual correlation coefficients; fMRI: $z = 2.031$ (± 0.106), $p < 0.001$). For details of unisensory auditory localisation within the MR scanner, please see Section 8.2.5.

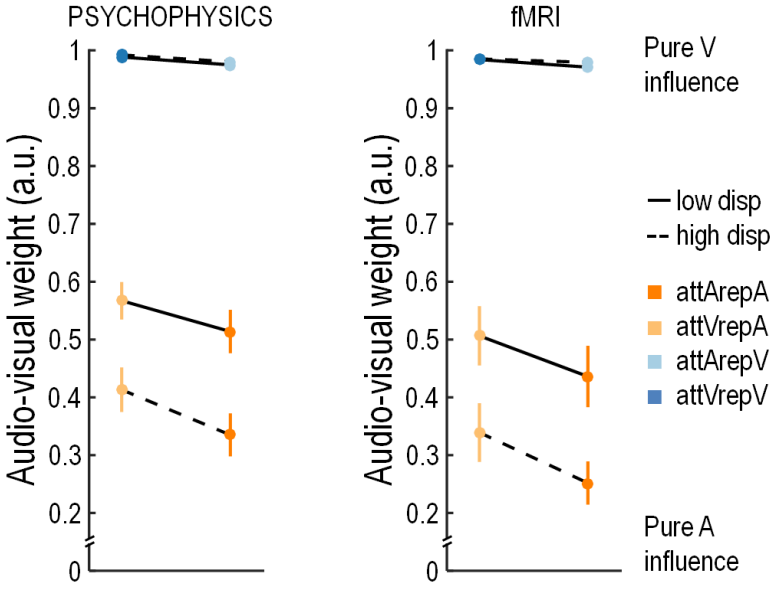

Figure 4.3: Audio-visual weight index (W_{AV}) in the psychophysics and fMRI experiments

Group mean W_{AV} (±SEM) is plotted as a function of AV spatial disparity (Low/High: 9°/18° visual angle), Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

W_{AV} (a.u.) mean (±SEM)	**attArepA**	**attVrepA**	**attArepV**	**attVrepV**
Psychophysics				
Low disparity	0.514 (±0.038)	0.567 (±0.032)	0.975 (±0.006)	0.988 (±0.003)
High disparity	0.335 (±0.037)	0.413 (±0.038)	0.980 (±0.006)	0.992 (±0.002)
fMRI				
Low disparity	0.436 (±0.053)	0.506 (±0.051)	0.971 (±0.008)	0.984 (±0.007)
High disparity	0.252 (±0.037)	0.339 (±0.051)	0.979 (±0.004)	0.985 (±0.005)

Table 4.1: Audio-visual weight index (W_{AV}) in the psychophysics and fMRI experiments

Group mean (±SEM) as a function of AV spatial disparity (Low/High: 9°/18° visual angle), Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

4.3.2 fMRI experiment: multivariate decoding results

To characterize how brain regions across the visual and auditory dorsal cortical hierarchies integrate auditory and visual signals into spatial representations, we combined fMRI with multivariate pattern decoding. A linear support vector regression model was trained on AV spatially congruent trials and subsequently used to decode spatial locations in AV spatially congruent and incongruent trials (Figure 4.2). Firstly, we checked that we could decode spatial locations for congruent trials significantly better than chance in all regions of interest. Two-tailed one-sample Wilcoxon signed-ranks tests confirmed that the Pearson correlation coefficient between decoded locations and true signals' locations was significantly different from zero in each ROI (Table 4.2). In other words, we successfully decoded AV congruent spatial locations along the visual and auditory dorsal cortical hierarchies.

Region	**V1-3**	**pIPS**	**aIPS**	**PT**	**A1-2**
Fisher-z transformed Pearson's correlation coefficient	1.356 (±0.082)	0.630 (±0.040)	0.415 (±0.041)	0.164 (±0.045)	0.124 (±0.035)
p-value	< 0.001	< 0.001	< 0.001	0.009	0.007

Table 4.2: Decoding of audio-visual congruent locations

Group mean (±SEM) Pearson's correlation coefficient between decoded locations and true signals' locations in audio-visual congruent trials for each ROI. Individual participants' correlation coefficients were Fisher-z transformed before entering them into two-tailed one-sample Wilcoxon signed-ranks tests against 0 (p-values show statistical significance).

Secondly, we investigated how each ROI weights auditory and visual signals sampled from incongruent locations to form spatial representations. Mirroring the analysis of behavioural localisation responses, we computed a neural audio-visual weight (nW_{AV}), which expresses the relative influence of visual and auditory signals on the decoded location (nW_{AV} = 1 reflects pure visual influence; nW_{AV} = 0 reflects pure auditory influence). Results are

shown in Figure 4.4 and summarised in Table 4.3 (we pulled over AV spatial disparity as it did not show any significant effects). The 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 2 (AV spatial disparity: low/high) repeated measures ANOVA with nW_{AV} as dependent variable showed a significant main effect of Attention in V1-3 ($F_{1,11} = 6.795$, $p = 0.024$, $\eta p^2 = 0.382$), reflecting a greater nW_{AV} for visual attention than auditory attention. Moreover, we found a significant Attention × Report interaction in pIPS ($F_{1,11} = 10.839$, $p = 0.007$, $\eta p^2 = 0.496$). Post-hoc t-tests revealed that nW_{AV} was greater for visual report than auditory report under auditory attention ($t_{11} = 2.163$, $p = 0.031$, Cohen's $d_{AV} = 0.624$, Wilcoxon signed-ranks $z = 2.040$, $p = 0.041$, $r = 0.416$). Finally, we found a main effect of Report at the top of the dorsal visual hierarchy, in aIPS ($F_{1,11} = 12.202$, $p = 0.005$, $\eta p^2 = 0.526$), reflecting a greater nW_{AV} for visual report than auditory report. Similarly, we found a main effect of Report at the top of the dorsal auditory hierarchy, in PT ($F_{1,11} = 3.957$, $p = 0.036$, $\eta p^2 = 0.265$), reflecting a greater nW_{AV} for visual report than auditory report.

nW_{AV} (a.u.) mean (±SEM)	**attArepA**	**attVrepA**	**attArepV**	**attVrepV**
V1-3	0.971 (±0.032)	1.018 (±0.026)	0.978 (±0.032)	0.994 (±0.036)
pIPS	0.757 (±0.060)	0.897 (±0.063)	0.902 (±0.065)	0.868 (±0.079)
aIPS	0.534 (±0.061)	0.685 (±0.061)	0.775 (±0.059)	0.743 (±0.054)
PT	0.041 (±0.052)	0.242 (±0.053)	0.315 (±0.061)	0.309 (±0.046)
A1-2	0.406 (±0.078)	0.254 (±0.085)	0.416 (±0.048)	0.437 (±0.073)

Table 4.3: Neural audio-visual weight index (nW_{AV})

Group mean (±SEM) as a function of Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

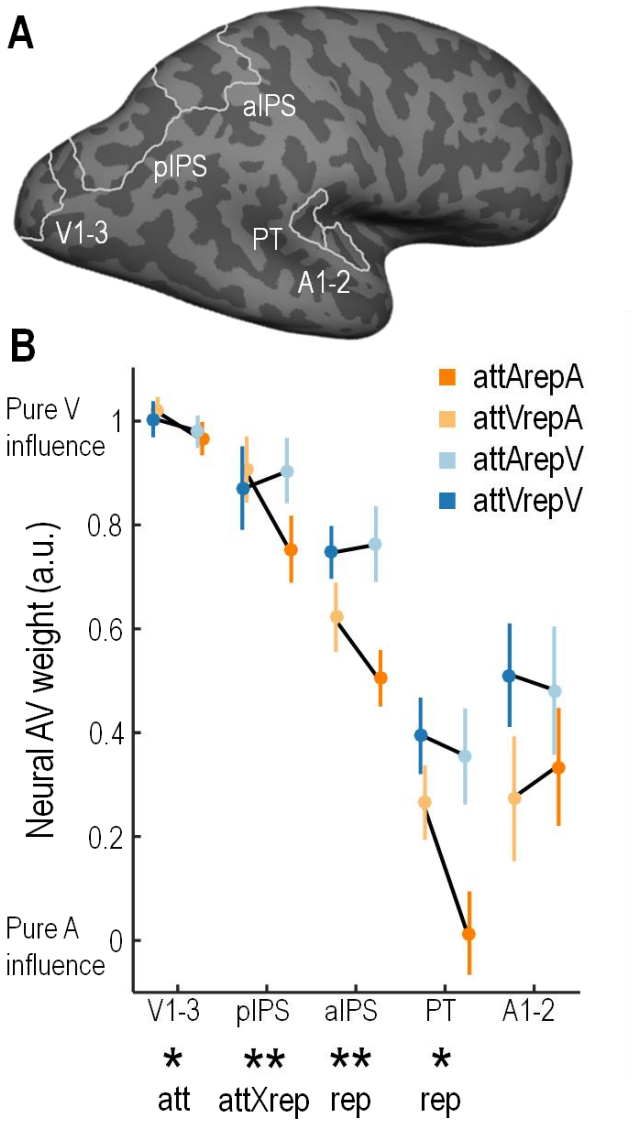

Figure 4.4: fMRI multivariate decoding results

A) fMRI voxel response patterns were obtained from anatomical ROIs along the visual and auditory dorsal cortical hierarchies: low-level visual cortex (V1-3), posterior intraparietal sulcus (pIPS), anterior intraparietal sulcus (aIPS), planum temporale (PT), low-level auditory cortex (A1-2). B) Group mean (±SEM) neural audio-visual weight index nW_{AV}: (Decoded location – A location) ÷ (V location – A location) as a function of Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual). * $p < 0.05$, ** $p < 0.01$

4.3.3 fMRI experiment: univariate results

To check whether participants were correctly engaged with task requests, we evaluated changes of brain activations, alongside response times and response errors (see Sections 8.2.1 and 8.2.2).

Reorienting of modality-specific attention across sensory modalities (i.e. attention invalidity as expressed by Attention × Report interaction [attVrepA & attArepV] > [attArepA & attVrepV]) increased activations in a bilateral fronto-parietal system encompassing the superior frontal gyrus, intraparietal sulcus, precuneus and middle frontal gyrus (Figure 4.5A). Significant activation increases were also found in the left inferior frontal gyrus, bilateral dorsal anterior cingulate gyrus, fusiform gyrus, cerebellum and calcarine cortex (Table 4.4)[4].

Similarly, AV spatial incongruence (i.e. AVincongruent > AVcongruent) increased activations in a bilateral fronto-parietal system encompassing the superior frontal gyrus, superior parietal lobule, intraparietal sulcus and inferior frontal gyrus (Figure 4.5B). Significant activation increases were also found in the bilateral anterior insula (Table 4.5). In a follow-up investigation, we notably found common increases of activation for attention invalidity and AV spatial incongruence (i.e. via a logical "AND" conjunction over the two) in the same bilateral fronto-parietal system, which included the superior frontal gyrus, superior parietal lobule and intraparietal sulcus (Figure 4.5C). Significant common activations were also found in the bilateral dorsal anterior cingulate gyrus and left anterior insula (Table 4.6). Collectively, the present results suggest the recruitment of a widespread domain-general executive control network (Duncan, 2010; Fedorenko et al., 2013) for reorienting of attention

[4] In a supplementary analysis (Section 8.2.4), we verified the effect of attention reorienting separately for auditory reports (i.e. attVrepA > attArepA) and visual reports (i.e. attArepV > attVrepV). The same bilateral fronto-parietal system was recruited for shifts of attention from vision to audition and from audition to vision (Figure 8.4; Table 8.6).

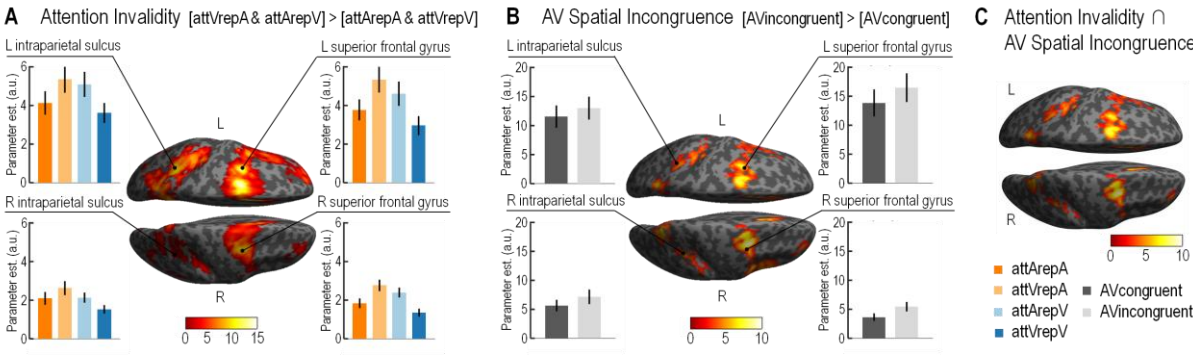

Figure 4.5: fMRI univariate results

Increases of BOLD response associated with A) attention invalidity B) AV spatial incongruence C) additive effects of attention invalidity and AV spatial incongruence (i.e. conjunction). Activation increases are rendered on an inflated canonical brain ($p < 0.001$ uncorrected at peak level for visualisation purposes, extent threshold $k > 0$ voxels). Bar plots represent group mean (±SEM) parameter estimates in non-dimensional units (corresponding to percentage whole-brain mean). Source of anatomical labels: Duvernoy (1999). L: left; R: right; A: auditory; V: visual; attA: auditory attention; attV: visual attention; repA: auditory report; repV: visual report.

across sensory modalities and for resolving audio-visual incongruence during spatial localisation.

Finally, we evaluated the main effect of modality-specific attention and report. On the one end, modality-specific attention (i.e. attA > attV and vice-versa) did not induce significant changes of activation in any brain areas. On the other hand, modality-specific report determined changes of activation depending on sensory modality (Table 4.7). Auditory relative to visual localisation (i.e. repA > repV) activated a bilateral system encompassing anterior insula, inferior frontal gyrus, dorsal anterior cingulate gyrus, superior frontal gyrus/sulcus and intraparietal sulcus. Visual relative to auditory localisation (i.e. repV > repA) induced activations in a bilateral system comprising the posterior cingulate gyrus, precuneus, angular gyrus, frontopolar gyrus and hippocampus. Significant activation increases were also found in the bilateral middle frontal gyrus, middle temporal gyrus and amygdala.

Collectively, auditory relative to visual localisation determined changes of BOLD-response that typically follow the performance of cognitive-demanding tasks, namely the concurrent activation of executive control areas (Duncan, 2010; Fedorenko et al., 2013) and deactivation of default mode areas (Raichle, 2000, 2015).

Brain regions	**MNI coordinates (mm)**			**Cluster size (voxels)**	**z-score (peak)**	**$p_{FWE\text{-}corrected}$ (peak)**
	x	y	z			
Attention invalidity						
L superior frontal gyrus	-4	8	52	7113	> 8	0.000
L superior frontal gyrus	-24	-6	56		> 8	0.000
R superior frontal gyrus	24	-4	52		> 8	0.000
L inferior frontal gyrus (pars opercularis)	-46	2	32		> 8	0.000
R anterior cingulate gyrus	10	18	36		7.41	0.000
L anterior cingulate gyrus	-10	18	32		5.11	0.007
L intraparietal sulcus	-28	-54	46	4129	> 8	0.000
L precuneus	-6	-62	48		> 8	0.000
L superior parietal lobule	-14	-68	50		> 8	0.000
R intraparietal sulcus	34	-42	44	43	5.40	0.002
L middle frontal gyrus	-28	48	12	472	> 8	0.000
R middle frontal gyrus	32	44	26	109	5.93	0.000
R fusiform gyrus	34	-54	-20	635	> 8	0.000
R cerebellum	38	-54	-32		6.34	0.000
L fusiform gyrus	-32	-50	-18	613	7.77	0.000
L cerebellum	-30	-56	-32		6.40	0.000
R calcarine cortex	8	-76	8	771	6.17	0.000
L calcarine cortex	-10	-80	6		5.85	0.000
Attention validity						
R inferior frontal gyrus (pars triangularis)	46	40	4	36	5.23	0.004
R lateral orbital gyrus	34	38	-10	13	5.00	0.011
L lateral orbital gyrus	-34	-36	-12	7	4.80	0.028

Table 4.4: fMRI univariate results: Attention (in)validity

Effect of attention invalidity (attVrepA&attArepV > attArepA&attVrepV) and attention validity (attArepA&attVrepV > attVrepA&attArepV). p-values are FWE-corrected at the peak level for multiple comparisons within the entire brain. Source of anatomical labels: Duvernoy (1999). attA: auditory attention; attV: visual attention; repA: auditory report; repV: visual report; L: left; R: right.

Brain regions	**MNI coordinates (mm)**			**Cluster size (voxels)**	**z-score (peak)**	**$p_{FWE\text{-}corrected}$ (peak)**
	x	y	z			
AV spatial incongruence						
R superior frontal gyrus	22	0	52	574	> 8	0.000
L superior frontal gyrus	-28	-6	60	512	7.57	0.000
L superior frontal gyrus	-2	14	48	778	> 8	0.000
R anterior cingulate gyrus	10	20	38		7.12	0.000
R anterior insula	34	20	4	274	7.22	0.000
L anterior insula	-30	24	0	194	6.72	0.000
R superior parietal lobule	16	-70	54	138	6.24	0.000
L superior parietal lobule	-16	-70	52	70	5.73	0.000
R inferior frontal gyrus (pars opercularis)	44	6	28	380	6.06	0.000
L inferior frontal gyrus (pars opercularis)	-54	4	20	44	5.63	0.001
R intraparietal sulcus	34	-44	44	68	5.51	0.001
L intraparietal sulcus	-42	-36	42	65	5.31	0.003
AV spatial congruence						
L angular gyrus	-52	-68	26	10	4.95	0.014
R frontopolar gyrus	6	62	-6	7	4.85	0.022

Table 4.5: fMRI univariate results: AV spatial (in)congruence

Effect of AV spatial incongruence (AVincongruent > AVcongruent) and AV spatial congruence (AVcongruent > AVincongruent). p-values are FWE-corrected at the peak level for multiple comparisons within the entire brain. Source of anatomical labels: Duvernoy (1999). AV: audio-visual; L: left; R: right.

Brain regions	MNI coordinates (mm)			Cluster size (voxels)	z-score (peak)	$p_{FWE\text{-}corrected}$ (peak)
	x	y	z			
Attention invalidity ∩ AV spatial incongruence						
R superior frontal gyrus	22	0	52	492	> 8	0.000
L superior frontal gyrus	-24	-4	54	512	7.23	0.000
L anterior cingulate gyrus	-2	14	48	615	> 8	0.000
R anterior cingulate gyrus	8	18	40		7.07	0.000
L anterior insula	-30	26	2	75	6.45	0.000
R superior parietal lobule	14	-68	54	96	6.14	0.000
L superior parietal lobule	-16	-70	52	70	5.73	0.000
R intraparietal sulcus	34	-44	46	25	5.14	0.006
L intraparietal sulcus	-34	-46	46	65	4.91	0.017

Table 4.6: fMRI univariate results: Attention invalidity and AV spatial incongruence

Common effect of attention invalidity and AV spatial incongruence given by the conjunction [attVrepA&attArepV > attArepA&attVrepV] ∩ [AVincongruent > AVcongruent]. p-values are FWE-corrected at the peak level for multiple comparisons within the entire brain. Source of anatomical labels: Duvernoy (1999). AV: audio-visual; attA: auditory attention; attV: visual attention; repA: auditory report; repV: visual report; L: left; R: right.

Brain regions	MNI coordinates (mm)			Cluster size (voxels)	z-score (peak)	$p_{FWE\text{-}corrected}$ (peak)
	x	y	z			
Modality-specific report						
repA > repV						
R anterior insula	34	22	0	1927	> 8	0.000
R inferior precentral sulcus	44	8	26		7.14	0.000
R inferior frontal gyrus (pars opercularis)	52	20	28		6.93	0.000
L anterior insula	-34	20	0	680	> 8	0.000
L superior frontal gyrus	-1	20	46	1200	> 8	0.000
R superior frontal gyrus	8	22	38		> 8	0.000
R anterior cingulate gyrus	8	22	38		> 8	0.000
L inferior frontal gyrus (pars triangularis)	-40	26	22	1306	> 8	0.000
L inferior precentral sulcus	-50	8	26		> 8	0.000

L superior frontal sulcus	-24	-4	58	156	5.99	0.000
R superior frontal sulcus	24	2	54	2	4.74	0.036
L intraparietal sulcus	-32	-48	38	49	5.33	0.002
R intraparietal sulcus	38	-44	44	30	5.23	0.004
repV > repA						
L posterior cingulate gyrus	-4	-38	42	2377	> 8	0.000
L precuneus	-6	-54	18		5.10	0.000
L angular gyrus	-54	-56	28	1531	> 8	0.000
L middle occipital gyrus	-30	-72	24		5.46	0.001
R angular gyrus	52	-50	32	806	> 8	0.000
R frontopolar gyrus	6	62	-6	1876	7.21	0.000
L frontopolar gyrus	-12	62	12		6.70	0.000
L middle frontal gyrus	-32	20	42	1184	7.17	0.000
R middle frontal gyrus	36	28	42	345	5.89	0.000
L postcentral gyrus	-40	-26	60	772	7.11	0.000
R middle occipital gyrus	36	-80	10	796	6.93	0.000
L middle temporal gyrus	-62	-46	-4	398	7.41	0.000
R middle temporal gyrus	66	-20	-8	352	7.35	0.000
L inferior frontal gyrus (pars triangularis)	-50	28	-2	175	6.41	0.000
L hippocampus	-28	-22	-18	136	5.96	0.000
R hippocampus	30	-22	-16	87	5.90	0.000
L Amygdala	-22	-2	-16	86	5.75	0.000
R Amygdala	22	-2	-16	25	5.32	0.002

Table 4.7: fMRI univariate results: Modality-specific report

Auditory report relative to visual report (repA > repV) and vice-versa (repV > repA). p-values are FWE-corrected at the peak level for multiple comparisons within the entire brain. Source of anatomical labels: Duvernoy (1999). repA: auditory report; repV: visual report; L: left; R: right.

4.4 Discussion

The present study combined psychophysics and fMRI analyses to characterize whether and how endogenous modality-specific attention impacts audio-visual spatial interactions and the underlying neural representations along the sensory cortical hierarchies. In particular, we evaluated whether attentional effects emerge already in early sensory areas or are deferred to higher-order association areas (Rohe & Noppeney, 2016). Critically, we dissociated between pre-stimulus focus (i.e. Attention) and post-stimulus response selection (i.e. Report) in an orthogonal cueing paradigm: participants were pre-cued to attend to audition (or vision) and they were post-cued to report their perceived auditory (or visual) location. Therefore, we could dissociate between two possible computational accounts of attentional control over audio-visual interactions. On the one hand, attention may change sensory noise and thereby affect reliability-weighed integration, under the assumption of a common cause (Alais & Burr, 2004; Ernst & Banks, 2002; Hillis et al., 2004); on the other hand, report may determine the selection of internal task-relevant spatial representations to produce a final response, in line with Bayesian Causal Inference (Körding et al., 2007; Shams & Beierholm, 2010).

At the behavioural level, we replicated previous findings (Ferrari & Noppeney, in preparation) by showing additive effects of attention and report on audio-visual spatial interactions. Visual attention (relative to auditory attention) and visual report (relative to auditory report) additively increased the influence of the visual signal's location on participants' localisation responses, as reflected by the main effect of attention on the W_{AV} index. Thus, we corroborate the view that selection and modulation of sensory signals (i.e. attention) affects the relative weights for multisensory integration via changes of sensory reliability (for more extensive discussion, see Ferrari & Noppeney, in preparation; for

complementary evidence, see model-based analysis in Section 8.2.3); moreover, selection of internal estimates (i.e. report) biases responses towards task-relevant sensory representations (Aller & Noppeney, 2019; Cao et al., 2019; Rohe et al., 2019; Rohe & Noppeney, 2015a, 2016, 2018). Furthermore, in line with the predictions of the BCI model (Körding et al., 2007; Shams & Beierholm, 2010) and in accordance with previous psychophysics (Rohe & Noppeney, 2015b) and neuroimaging (Aller & Noppeney, 2019; Rohe & Noppeney, 2015a, 2016) studies, we replicated decreased integration at higher AV spatial disparities, where signals are less likely to originate from a common source. However, the effect was present under auditory reports only. As previously discussed (Ferrari & Noppeney, in preparation), the expression of disparity effects under visual report was likely prevented by the audio-visual weight being already close to maximum. Such ceiling effect can be explained by the well-known superiority of vision over audition in driving spatial localisation responses, due to the generally higher spatial reliability of visual signals (Freides, 1974). In line with this interpretation, here we corroborate greater task difficulty for auditory than visual localisation via fMRI univariate analysis. Auditory relative to visual localisation triggered the activation of a widespread bilateral network implicated in control of domain-general task difficulty (Duncan, 2010; Fedorenko et al., 2013) and the concurrent deactivation of default mode areas, consistently with the execution of cognitively demanding tasks (Raichle, 2000, 2015).

For multivariate decoding, we first ensured successful decoding of spatial estimates in all regions of interest, as indexed by the significant correlation coefficient between true spatial location and decoded location for audio-visual congruent trials, which are spatially non-ambiguous. Second, we demonstrated that attention and report govern audio-visual interactions at different levels of the sensory cortical hierarchies. In primary visual cortices, visual attention relative to auditory attention boosted the influence of visual signals on the

formation of spatial representations, as reflected by increases of the nW_{AV} index. At the top of the hierarchy, in the anterior intraparietal sulcus (aIPS) and planum temporale (PT), visual report relative to auditory report increased the nW_{AV} index. Importantly, this result not only corroborates but also expands previous evidence (Aller & Noppeney, 2019; Cao et al., 2019; Rohe et al., 2019; Rohe & Noppeney, 2015a, 2016, 2018) by dissociating the effect of attention and report. However, unlike previous work (Aller & Noppeney, 2019; Rohe & Noppeney, 2015a, 2016), we did not find any effects of spatial disparity on the nW_{AV} index. Here, the use of limited spatial eccentricities may have impacted the signal-to-noise ratio for decoding spatial positions during spatially incongruent trials, where estimates are generally noisier than in the case of congruent trials. Moreover, the Attention × Report interaction found in the posterior intraparietal sulcus (pIPS) likely reflects the absence of a strong effect of either factor on the decoded spatial representations. Indeed, recent work suggests that perceptual salience and behavioural relevance independently impact activation profiles along the visual dorsal cortical hierarchy (Sprague et al., 2018).

As a whole, the present study confirms that attentional control over multisensory interactions is pervasive along the sensory cortical hierarchies, but is driven by distinct computational principles. In low-level visual areas, pre-stimulus focus biases the competition among spatial representations towards attended stimuli, irrespective of response requests[5]. Thus, the selection and modulation of sensory information increases the weight of attended representations at the bottom of the visual hierarchy. Although the current study cannot draw conclusions about the underlying neural computations, we propose that pre-stimulus focus may change spatial representations' precision via sharpening of tuning functions (Martinez-

[5] Null effects in low level auditory areas must be treated with caution, as low signal-to-noise ratio (which already emerged for decoding of AV spatially congruent trials, see Table 4.2) might have prevented reliable decoding of spatial representations in case of AV spatially incongruent trials.

Trujillo & Treue, 2004), possibly via modulation of internal noise (Serences & Kastner, 2014). In high-level sensory areas, post-stimulus response selection biases the competition among spatial representations towards task-relevant stimuli, irrespective of pre-stimulus focus. Thus, context-dependent behavioural relevance determines the selective read-out of spatial representations at the top of the sensory hierarchies (Pestilli et al., 2011; Serences & Kastner, 2014), in line with the presence of priority maps that are scaled by task relevance (Bisley & Goldberg, 2010; Rohe & Noppeney, 2016; Sprague et al., 2018). Due to the sluggishness of the BOLD-response, here we could not investigate how attentional influences on audio-visual spatial interactions evolved over time. Future electrophysiological studies using similar multivariate decoding techniques (Aller & Noppeney, 2019; Rohe et al., 2019) should characterise such temporal dynamics while also corroborating the differential impact of pre-stimulus focus and post-stimulus response selection along the sensory cortical hierarchies.

Critically, the univariate analysis of fMRI data provided evidence that participants properly engaged with task instructions, as reflected by brain activations for the attention x report interaction (i.e. validity effect). Invalid relative to valid trials (regardless of sensory modalities) increased the BOLD-response in a widespread bilateral fronto-parietal system (encompassing the superior frontal gyrus and superior parietal lobule / intraparietal sulcus), which is known to control intermodal (re)orienting of attention based on cueing paradigms (Corbetta et al., 2008; Corbetta & Shulman, 2002; Santangelo et al., 2010; Shomstein & Yantis, 2004). Moreover, audio-visual spatially incongruent trials relative to congruent trials activated similar bilateral fronto-parietal regions, the anterior cingulate gyrus and the anterior insula, which are central nodes of a so-called salience network implicated in conflict detection and cognitive control (Menon & Uddin, 2010). In the current study, these areas may have

acted as a control hub to detect and resolve conflicts arising in case of attention invalidity and AV spatial incongruence, which required enhanced cognitive effort as indexed by response times costs and response errors (see Sections 8.2.1 and 8.2.2). Another recent fMRI study (Love et al., 2018) points to the same conclusion: more cognitively demanding multisensory tasks (i.e., audio-visual temporal-order judgement relative to synchrony judgement) induced activations of middle frontal cortex, precuneus and superior medial frontal cortex. Consistently, previous neuroimaging work (Duncan, 2010; Fedorenko et al., 2013) has shown that domain-general task difficulty (which determines increased response times and decreased accuracy) engages a widespread system encompassing the superior frontal gyrus, intraparietal sulcus, anterior cingulate gyrus and anterior insula. In summary, the present fMRI univariate results confirm that participants were shifting attention according to pre- and post-cues and were resolving audio-visual spatial incongruence to perform spatial localisation, in line with task requests.

One may wonder whether the absence of a main effect of attention (i.e. pre-stimulus focus) on the BOLD-response hinders the conclusion that participants were directing their focus in accordance with pre-cues. In particular, it could be expected that sustained modality-specific attention determines BOLD-response increases in sensory areas associated with the attended modality and deactivations in sensory areas associated with the unattended modalities (Johnson & Zatorre, 2005, 2006; Mozolic et al., 2008b). Accordingly, previous electrophysiological studies have shown alpha-band power decreases (which reflect a state of heightened local excitability, Jensen & Mazaheri, 2010) over the sensory area subserving the attended modality, and vice-versa alpha-band power increases over sensory areas subserving the unattended modalities (Foxe et al., 1998; Fu et al., 2001; Gomez-Ramirez et al., 2011; Mazaheri et al., 2014). A crucial difference in terms of task design explains the lack of intra-

modal activation and cross-modal deactivation effects due to sustained modality-specific attention in the current study. In order to minimise selection history effects (Awh et al., 2012; Theeuwes, 2018), we employed a 1:1 ratio of valid / invalid trials (i.e. respond to the attended / unattended stimulus), which determined frequent attentional shifts based on pre- and post-cues and therefore disrupted the deployment of sustained modality-specific attention over the course of each attention block. Thus, our design was not optimised to detect intra-modal activation and cross-modal deactivation effects in sensory areas due to sustained attention, but instead it determined strong recruitment of a fronto-parietal system implicated in (re)orienting of modality-specific attention (Corbetta et al., 2008; Corbetta & Shulman, 2002; Santangelo et al., 2010; Shomstein & Yantis, 2004). Future fMRI studies aiming to detect the effect of sustained modality-specific attention on the BOLD response could use long (10-12 s) intervals between attention cue and stimuli to isolate attention-related activations in sensory areas (Kastner et al., 1999). Alternatively, one could consider using time-resolved methods such as M/EEG (Aller & Noppeney, 2019; Cao et al., 2019).

Collectively, our results demonstrate that attentional control over multisensory perceptual inference is pervasive in human multisensory neo-cortex, but it impacts different computational tasks along the sensory cortical hierarchies.

CHAPTER 5

CROSS-MODAL BINDING CAPTURES ATTENTION WITHIN A COCKTAIL-PARTY SCENARIO

Ambra Ferrari, Giulio Degano, Uta Noppeney

Computational Cognitive Neuroimaging lab, Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, B15 2TT Birmingham, UK

Citation:

Ferrari, A., Degano, G. & Noppeney, U. (in preparation). Cross-modal binding captures attention within a cocktail-party scenario.

Authors contributions:

Experiment conceptualisation and design: Ambra Ferrari, Giulio Degano, Uta Noppeney.
Data collection: Ambra Ferrari.
Data analysis: Ambra Ferrari (supervised by Uta Noppeney).
Writing: Ambra Ferrari (supervised by Uta Noppeney).

Abstract

In cluttered environments, concurrent auditory inputs compete for our attention. Crucially, naturalistic listening is aided by coherent cross-modal information, such as lip movements. This raises the critical question of whether multisensory objects enhance selective attention during competition for processing resources. The present study investigated whether cross-modal binding captures attention, free of linguistic confounds. Critically, we first independently assessed that participants were able to perceive multisensory congruence. Subsequently, we asked them to perform a target detection task in an auditory cocktail-party scenario. We evaluated whether cross-modal binding enhanced selective attention towards one of concurrent competing auditory streams and therefore impacted target detectability. Perceptual sensitivity changed as a function of multisensory congruence. In particular, d' decreased when targets were not presented in the stream of information containing cross-modal coherent information, showing that cross-modal binding captures attention. In addition, response criterion became more liberal for targets presented concurrently across modalities, confirming that cross-modal redundancy impacts decision strategies. In summary, the present study demonstrates that multisensory objects promote attentional selection under competition for processing resources.

Keywords

Multisensory integration, selective attention, cocktail-party scenario, target detection

5.1 Introduction

In our complex and dynamic world, we are constantly bombarded with a myriad of sensory inputs that tax our limited processing resources. For example, while following a conversation in a noisy environment, concurrent auditory streams compete for our attention. Since the pioneering work on the so-called cocktail-party problem (Cherry, 1953), numerous studies have described our remarkable ability to selectively attend to one source of information and ignore concurrent inputs (Bronkhorst, 2015). Crucially, selective listening is enhanced by lip-reading (Bernstein et al., 2004; Grant & Seitz, 2002; Sumby & Pollack, 1954), especially in noisy conditions (Crosse et al., 2016; Ross et al., 2007; van de Rijt et al., 2019) and in the presence of simultaneous competing speech (Helfer & Freyman, 2005; Zion Golumbic et al., 2013). To account for this phenomenon, it has been suggested that cross-modal binding (i.e. the automatic grouping of coherent cross-modal features such as voice and lip movements into a unified object or event, Bizley et al., 2016) promotes attentional selection in cocktail-party conditions. Accordingly, recent work (Maddox et al., 2015) leveraged cross-modal binding to spread object-based attention from vision to audition (Busse et al., 2005) and therefore orient selective attention toward one of competing auditory streams (i.e. the one matching the visual stream). However, the authors did not independently establish that cross-modal binding occurred in the first place and consequently triggered cross-modal spread of object-based attention, resulting in changes of auditory perception. Instead, they used changes of auditory perception to simultaneously probe cross-modal binding and its impact on auditory scene analysis. Given the absence of an independent assay of cross-modal binding, the interpretation of these results remains controversial.

Crucially, it has been shown that objects (i.e. perceptual units resulting from organization of elements via Gestalt factors) are salient entities that capture attention (Humphreys & Riddoch, 2003; Kimchi et al., 2007; Yeshurun et al., 2009). Favouring perceptual units may be advantageous in order to quickly identify and react to objects in the surrounding environment. Accordingly, it has been shown that multisensory redundancies generated by a common source scaffold the development of selective attention in human infants (Bahrick & Lickliter, 2000; Bahrick et al., 2004) and still guide attention in adulthood when performing challenging tasks in relation to the abilities of the perceiver (Lickliter & Bahrick, 2013). Building on this body of evidence, here we addressed the hypothesis that cross-modal binding captures attention and consequently impacts perception in a cocktail-party scenario.

Critically, the use of speech stimuli confounds perceptual and linguistic processes during naturalistic listening. Alongside the potential influence of cross-modal binding on the allocation of selective attention, semantic context supports speech tracking and comprehension (Broderick et al., 2019; Davis & Johnsrude, 2007; Hannemann et al., 2007; Kuperberg & Jaeger, 2016; Mattys et al., 2012). Here, we sought to isolate the former process by use of custom-composed unknown music pieces because (i) they allow the use of cross-modal temporal coherence to elicit cross-modal binding (Bizley et al., 2016; Noppeney & Lee, 2018; Shamma et al., 2011) and (ii) they avoid linguistic confounds that could impact speech intelligibility. The use of music stimuli also enabled us to further test the hypothesis that multisensory integration plays a crucial role for the temporal parsing of naturalistic music, which may be especially amplified in expert musicians (Jicol et al., 2018; Lee & Noppeney, 2011a; Petrini et al., 2009, 2011). Moreover, we paired auditory and tactile information with the purpose of eliciting cross-modal binding because in everyday situations

they provide redundant information about vibratory events (Soto-Faraco & Deco, 2009), which are particularly useful to parse real-life stimuli such as music (Huang et al., 2012; Tranchant et al., 2017) and speech (Drullman & Bronkhorst, 2004; Fletcher et al., 2018; Huang et al., 2017; Riecke et al., 2019). Hence, audio-tactile music stimuli represented a promising choice for the study of cross-modal binding during naturalistic listening, beyond the use or artificial and transient stimuli (Petrini et al., 2014; Soto-Faraco & Deco, 2009; Stanley et al., 2019). Importantly, we first assessed that participants perceived audio-tactile congruence (via a preliminary screening) and we subsequently exploited cross-modal binding to unambiguously evaluate whether it enhances selective attention in a cocktail-party scenario[1].

In experiment 1, we presented participants with two simultaneous auditory streams (one signal, one masker), which they had to track in order to perform a target detection task[2]. We then measured detection of targets (which appeared in the signal stream) under competition for auditory attention (due to the concurrent masker stream). We investigated whether and how a tactile stream matching the auditory signal stream ("match-signal") enhances target detection relative to a tactile stream matching the auditory masker stream ("match-masker") and relative to no tactile stream ("no-touch"). We hypothesised that audio-tactile congruence would direct attention to the signal stream in the match-signal condition, to the masker stream in the match-masker condition and that attention would be divided across the signal and masker streams in the no-touch condition (Figure 5.1). Since selective attention is known to amplify the detectability of attended features (Theeuwes & Chen, 2005; Theeuwes et al.,

[1] Perceived multisensory congruence such as temporal coherence determines cross-modal binding (Bizley et al., 2016; Noppeney & Lee, 2018).
[2] For a description of the target, see Section 5.2.2 and Figure 5.2A.

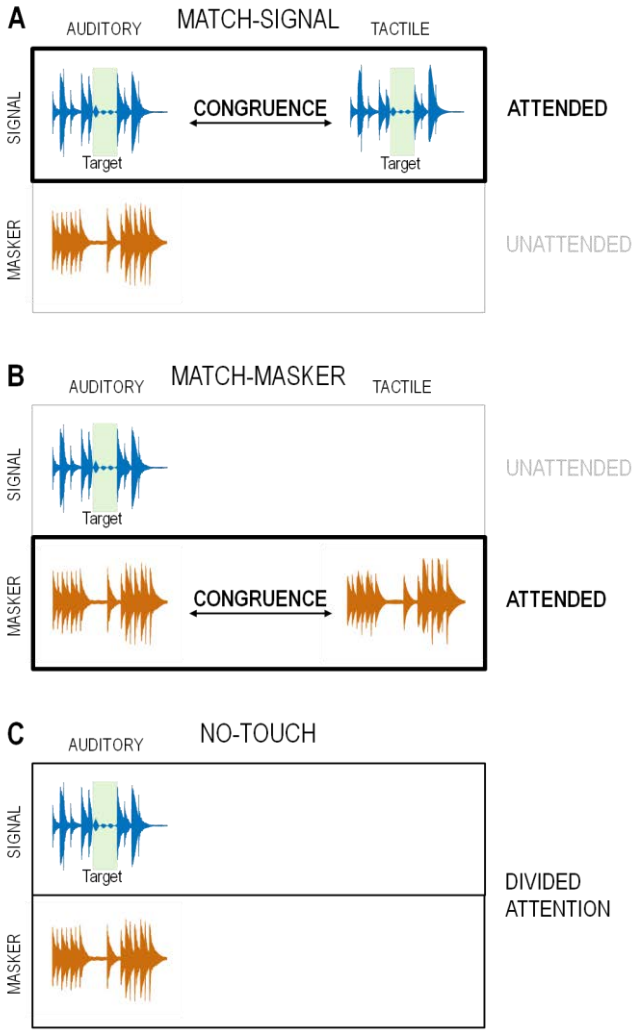

Figure 5.1: Rationale of the study

A) In the match-signal condition, the tactile stream presents envelope and frequency information matched to the auditory stream containing the target (signal stream); we hypothesise that audio-tactile congruence directs attention to the signal stream. B) In the match-masker condition, the tactile stream is matched to the auditory stream not containing the target (masker stream); we hypothesise that audio-tactile congruence directs attention to the masker stream. C) In the no-touch condition, there is no tactile stream; we hypothesise that attention is divided between the signal and masker streams.

2004), we expected changes in target detection performance in terms of perceptual sensitivity (i.e. d').

Moreover, we compared both match-signal and match-masker conditions to the no-touch condition to evaluate whether changes of performance resulted from a beneficial attentional enhancement of the signal stream or from an interference of the masker stream relative to a neutral baseline (i.e. divided attention).

Critically, when a target appeared in the match-signal condition, it was presented both in the auditory and tactile modalities to preserve full congruence among the two streams. Conversely, in the match-masker and no-touch conditions we presented only one target in the auditory modality (Figure 5.1). Thus, when comparing the match-signal condition to the match-masker and no-touch conditions, differences in target detection might have originated from the summation of auditory and tactile target information relative to auditory information alone. In experiment 2 we directly addressed this interpretation by investigating whether and how audio-tactile stimulation per se determines a multisensory benefit for target detection relative to auditory stimulation alone.

5.2 Materials and methods

5.2.1 Participants

Twenty-four participants (3 males; mean age 22, range 18-30 years) were included in the psychophysics experiment based on a priori power analysis (G*Power 3.1, Faul et al., 2007; 2009) with power $(1-\beta) = 0.8$, $\alpha = 0.05$ and effect size Cohen's $d_{AV} = 0.53$. Estimation of effect size was derived from a preliminary pilot study with 9 participants[3]. Two additional volunteers were excluded based on a priori inclusion criteria (see Section 0). All volunteers reported normal or corrected to normal vision, normal hearing and touch and no history of

[3] We used Cohen's d_{AV} of a one-tailed paired sample t-test that evaluated changes of d' between match-signal and match-masker conditions in experiment 1.

neurological or psychiatric conditions. They had never received any formal music training and were classified as non-musicians via the Music USE (MUSE) Questionnaire (Chin & Rickard, 2012), based on duration, frequency and regularity of instrument playing[4]. All volunteers provided written informed consent and were naïve to the aim of the study; they received a reimbursement in the form of money or university credits for their participation in the experiment. The study was approved by the University of Birmingham Ethical Review Committee and was conducted in accordance with these regulations.

5.2.2 Stimuli

Auditory stimuli consisted of 8 s monophonic music pieces, custom-composed from an online database (Disbergen et al., 2018; https://www.zlab.mcgill.ca/) and synthesized from Musical Instrument Digital Interface (MIDI) files using Linux MultiMedia Studio 1.1.3 (LMMS, https://lmms.io/) with a piano sound font (grand-piano-YDP-20160804). Tactile stimuli consisted of the same 8 s music pieces, synthesized using LMMS with sinusoidal oscillations (TripleOscillator). As a result, we obtained corresponding envelope and frequency information across audition and touch for each music piece (mean sound pressure level: 65 dB; frequency range: 1-500 Hz). Auditory and tactile synthesized stimuli were recorded at 44100 Hz with 16-bit resolution, normalised and saved as WAV files using Audacity 2.1.2. For cocktail-party conditions, auditory monophonic pieces were combined into two-stream polyphonic pieces via simultaneous recording. In both experiments, participants performed a yes-no target detection task. The target consisted of a 2 Hz sinusoidal modulation of envelope

[4] Index of Music Instrument Playing (IMIP) < 0.4, where IMIP = [Years of instrument playing x Hours of practice per day / Regularity of practice] and Regularity of practice ("How long since you last regularly played music?") scored as follows: 1 for "less than one week", 2 for "less than one month", 4 for "less than one year", 8 for "between 1 and 5 years", 16 for "between 5 and 10 years", 32 for "more than 10 years".

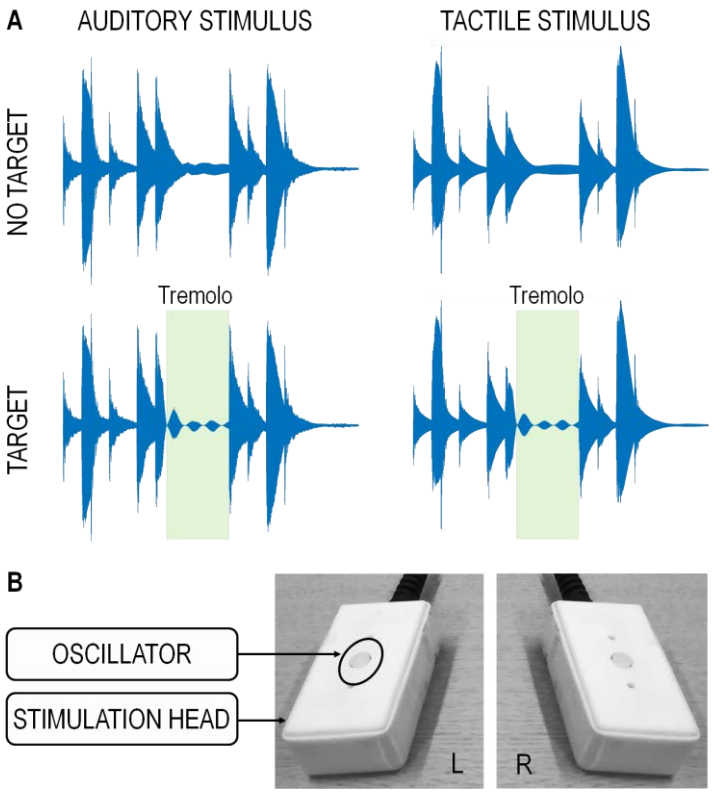

Figure 5.2: Experimental stimuli

A) Auditory and tactile signals provided correspondent envelope and frequency information for each music piece; the target consisted of a 2 Hz sinusoidal modulation of envelope intensity called 'tremolo' (highlighted in green). B) Piezoelectric system. A stimulation head (rectangular box) was applied to each hand, with the fingertip of each index finger in correspondence with the oscillator (encircled). The oscillator provided vibrations by moving up and down. L: left; R: right.

intensity called "tremolo" (Figure 5.2A), which was inserted 300 ms after the onset of a note using Audacity (1700 ms duration, 100 ms fade-out).

5.2.3 Experimental setup

The experiment was presented via Psychtoolbox version 3.0.11 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) running under MATLAB R2011b (MathWorks Inc.) on a MacBook Pro (Mac OSX 10.6.8). Stimuli were extracted from WAV files and played via MATLAB custom-code. Auditory stimuli were presented through headphones (HD 280 PRO,

Sennheiser, Wedemark-Wennebostel, Germany) via the laptop's built-in soundcard. Tactile stimuli were presented through a piezoelectric system (PTS-C2, Dancer Design, UK) via an external sound-card (Asus Xonar U7, Taiwan). A piezoelectric stimulation head (Figure 5.2B) was applied to each hand, with the fingertip of each index finger in correspondence with the stimulation oscillator. For multisensory conditions, we adjusted audio-tactile latencies in the presentation software and confirmed their synchrony by recording and measuring their relative latencies using two microphones. Participants were instructed to sit still in a dimly lit cubicle with their eyes closed and their head positioned on a chin rest. Responses were collected via two pedals (SODIAL, Shenzhen IMC Digital Technology Co.), one in correspondence of each foot.

5.2.4 Inclusion criteria

Via a preliminary screening (see Section 5.2.5) we independently verified that participants perceived multisensory congruence. Volunteers were pre-selected based on the ability to perceptually discriminate audio-tactile congruence (i.e. same piece across audition and touch) and incongruence (i.e. two different pieces across audition and touch). Volunteers who showed d' > 2.8 (see Section 5.2.5) were included in the study (group mean ± SEM d' for included participants = 5.253 ± 0.370; $criterion_{center}$ = -0.053 ± 0.123; proportion of correct responses = 0.969 ± 0.006)[5].

5.2.5 Screening

In a yes-no congruence judgement paradigm (Figure 5.3), we presented synchronous auditory and tactile stimuli that were either congruent or incongruent. Background white noise was

[5] Threshold was defined as two standards deviations below the group mean d' in a preliminary pilot study with 9 participants.

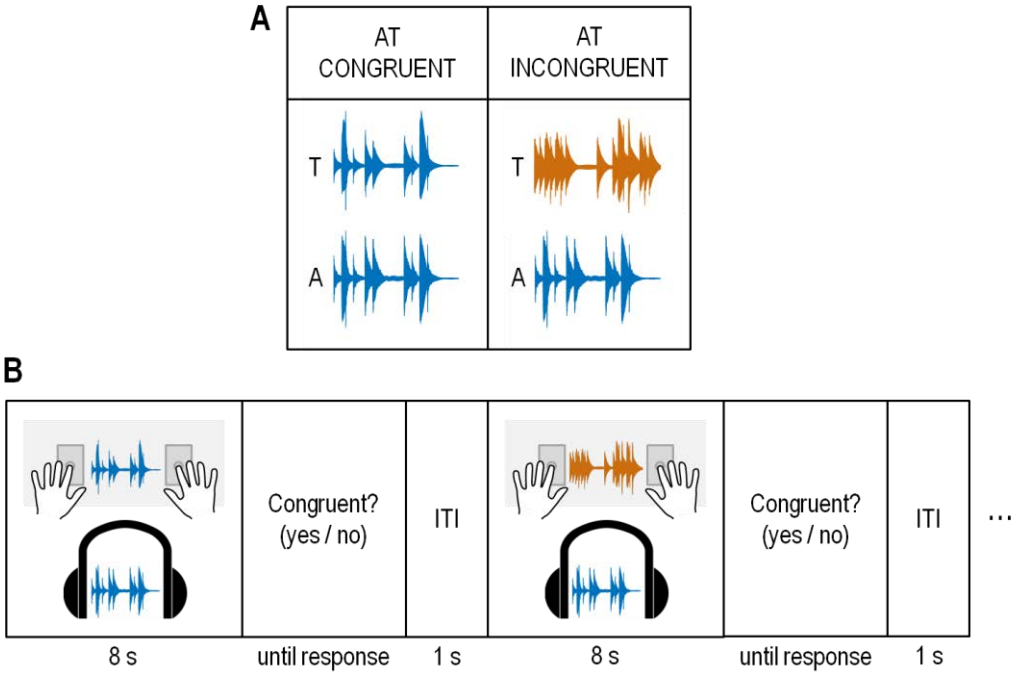

Figure 5.3: Screening design and procedure

A) The experimental design comprised 2 conditions: audio-tactile congruence (i.e. same piece across audition and touch) and incongruence (i.e. two different pieces across audition and touch). B) Experimental procedure: after synchronous auditory and tactile stimuli presentation (8 s piano pieces), participants reported via pedal press whether stimuli were congruent (left foot) or not (right foot). A: auditory; T: tactile; AT: audio-tactile; ITI: inter-trial interval.

additionally played through the headphones (65 dB sound pressure level) to mask the sound of tactile vibrations. After stimuli presentation, participants reported whether they perceived the same music piece through audition and touch via pedal press (yes: left pedal; no: right pedal). Focus was put on accuracy and there was no time limit for the response. The experimental setup was the same as in experiments 1 and 2. After familiarization with stimuli and procedure via one preliminary practice run, each participant completed 2 experimental runs (2 conditions × 15 trials / condition / run × 2 runs = 60 trials in total). Based on signal detection theory (Wickens, 2002), ‘yes’ and ‘no’ responses in congruent trials were classified as hits and misses, whereas ‘yes’ and ‘no’ responses in incongruent trials were classified as

false alarms and correct rejections. Accordingly, we calculated d' (Z(hit rate) - Z(false alarm rate)), correct responses (proportion hits + proportion correct rejections) and $criterion_{center}$ (-[Z(false alarm rate) + Z(hit rate)] / 2).

5.2.6 Experimental design and procedure

5.2.6.1 Experiment 1

In a yes-no target detection paradigm, we created a 3 (match-signal / match-masker / no-touch) × 2 (target / catch) experimental design (Figure 5.4A). The same catch trials were shared across match-signal and match-masker conditions, thus producing 5 experimental conditions in total. Targets (Figure 5.2A) appeared in the first or second half of a stream with a 1:1 ratio to minimise any target onset expectations. For each participant, stream identity (stream 1 / stream 2) and target position (first half / second half) were counterbalanced across conditions and the order of presentation was randomised for each experimental run. At the onset of each trial (Figure 5.4B), participants were presented with two simultaneous auditory streams (one target, one masker) and they were instructed to pay attention to both in order to detect targets. In match-signal and match-masker conditions, tactile streams were simultaneously presented at the fingertip of each index finger. Stimuli presentation was accompanied by background white noise played through the headphones (65 dB sound pressure level), which served a two-fold aim: it masked the sound of tactile vibrations and it prevented ceiling effects for auditory target detection[6]. After stimuli presentation, participants reported whether they perceived a target via pedal press (yes: left pedal; no: right pedal). Focus was put on accuracy and there was no time limit for the response. After familiarization with stimuli, target and procedure via one preliminary practice run, each participant

[6] We defined signal-to-noise ratio (SNR = 0) based on a preliminary pilot study with 9 participants which showed 65% correct responses (hits + correct rejection) for unisensory auditory target detection.

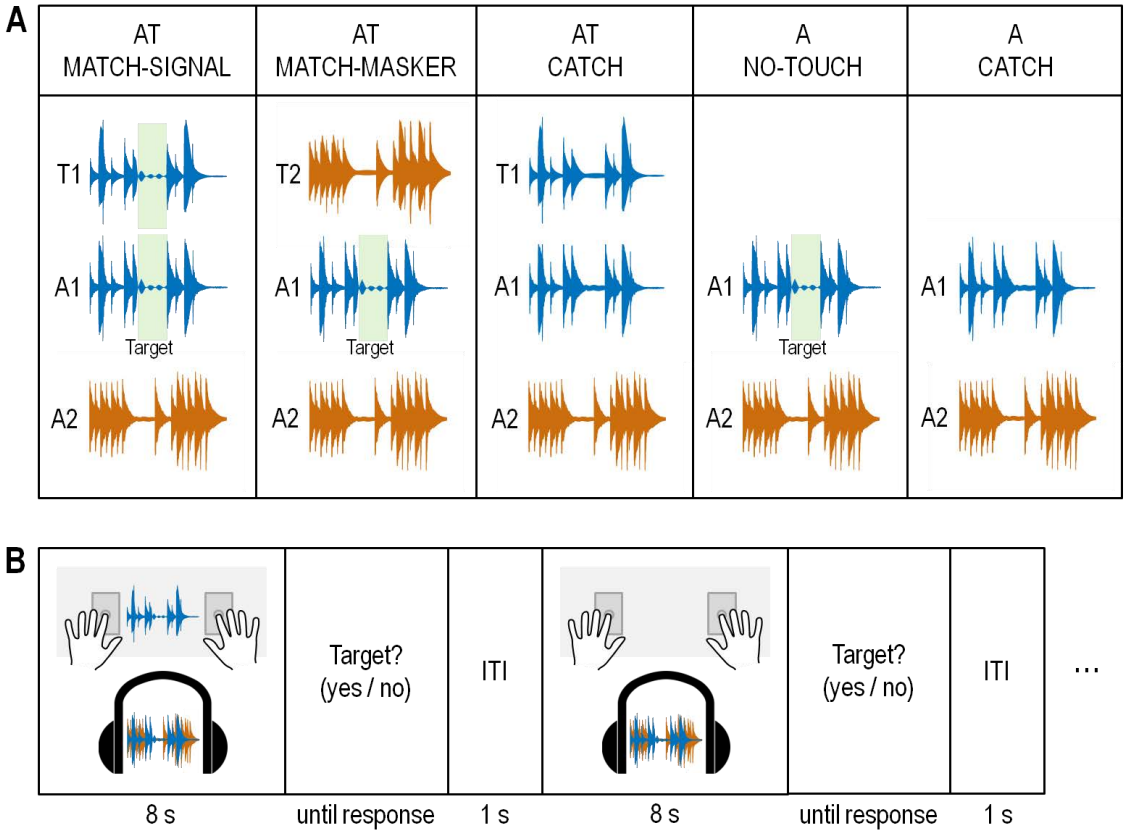

Figure 5.4: Experiment 1 design and procedure

A) The experimental design comprised 3 (match-signal / match-masker / no-touch) x 2 (target / catch) conditions. The same catch trials were shared across match-signal and match-masker conditions, thus producing 5 experimental conditions in total. Targets are highlighted in green. B) Experimental procedure: after stimuli presentation (8 s piano pieces), participants reported via pedal press whether a target was present (left foot) or not (right foot). A; auditory; T: tactile; AT: audio-tactile; ITI: inter-trial interval.

completed 7 experimental runs (5 conditions × 6 trials / condition / run × 7 runs = 210 trials in total).

5.2.6.2 Experiment 2

In a yes-no target detection paradigm, we created a 2 (audio-tactile stimulation / auditory stimulation) × 2 (target / catch) experimental design (Figure 5.5A). Targets (Figure 5.2A) appeared in the first or second half of a stream with a 1:1 ratio to minimise any target onset

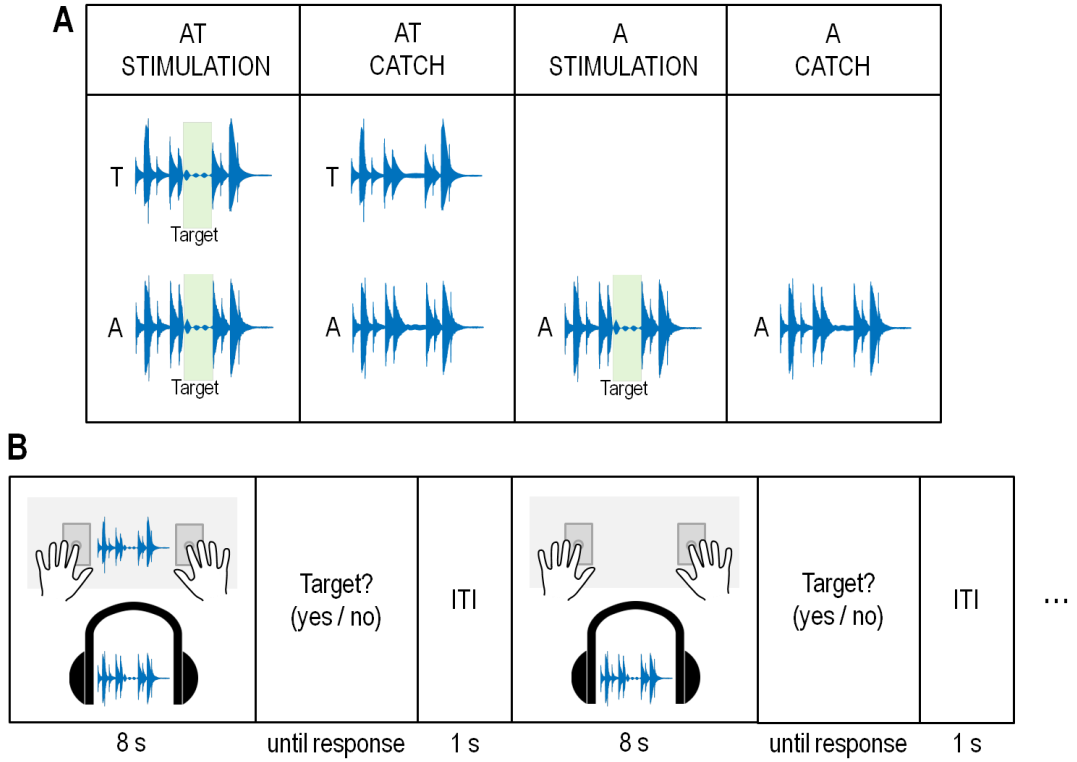

Figure 5.5: Experiment 2 design and procedure

A) The experimental design comprised 2 (audio-tactile stimulation / auditory stimulation) x 2 (target / catch) conditions. Targets are highlighted in green. B) Experimental procedure: after stimuli presentation (8 s piano pieces), participants reported via pedal press whether a target was present (left foot) or not (right foot). A; auditory; T: tactile; ITI: inter-trial interval.

expectations. For each participant, target position (first half / second half) was counterbalanced across conditions and the order of presentation was randomised for each experimental run. At the onset of each trial (Figure 5.5B), participants were presented with simultaneous auditory and tactile streams or with an auditory stream alone and were instructed to detect auditory targets. Stimuli presentation was accompanied by background white noise played through the headphones (65 dB sound pressure level), which served a two-fold aim: it masked the sound of tactile vibrations and it prevented ceiling effects for auditory target detection. After familiarization with stimuli, target and procedure via one preliminary

practice run, each participant completed 5 experimental runs (4 conditions × 8 trials / condition / run × 5 runs = 160 trials in total).

5.2.7 Experimental data analysis

For both experiments, we performed the following analyses. Based on signal detection theory (Wickens, 2002), 'yes' and 'no' responses in target trials were classified as hits and misses, whereas 'yes' and 'no' responses in no target trials were classified as false alarms and correct rejections. For each participant and experimental condition, we calculated proportion of correct responses (proportion hits + proportion correct rejections), d' (Z(hit rate) - Z(false alarm rate)) and $criterion_{center}$ (- [Z(false alarm rate) + Z(hit rate)] / 2).

After rejection of normality (Kolmogorov-Smirnov Test), individual proportion of correct responses, d' and $criterion_{center}$ were entered into group pair-wise comparisons across conditions via two-tailed non-parametric Wilcoxon signed-ranks tests. Bonferroni correction was used to account for multiple comparisons (experiment 1: $\alpha = 0.017$; experiment 2: $\alpha = 0.025$).

5.3 Results

For both experiments, first we report proportion of correct responses to evaluate whether performance changed across experimental conditions. Second, we report d' and $criterion_{center}$ to evaluate how responses changed, i.e. we dissociate perceptual sensitivity and decision bias.

5.2.8 Experiment 1

Results are shown in Figure 5.6 and summarized in Table 5.1. Proportion of correct responses was the lowest for match-masker, followed by no-touch and finally the highest for match-

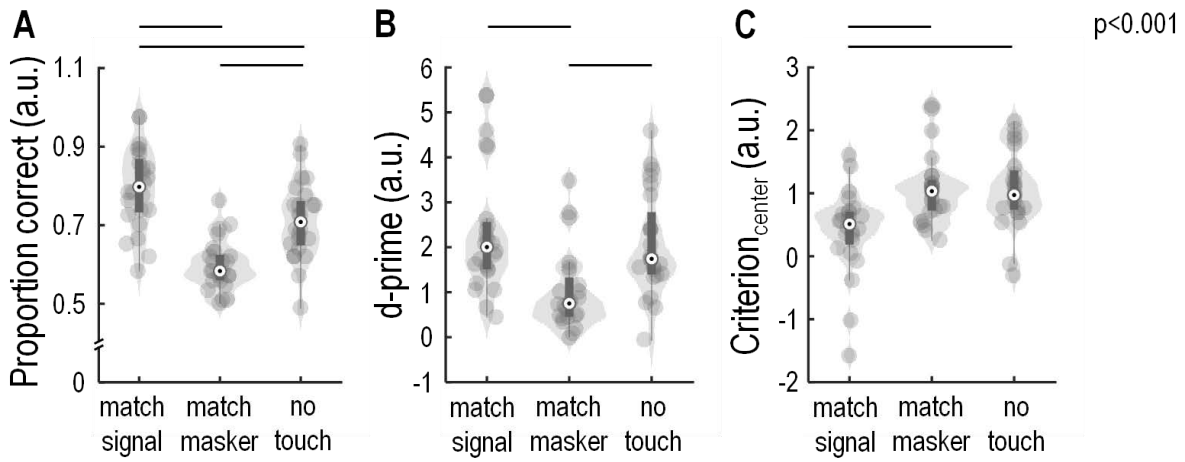

Figure 5.6: Results of experiment 1

A) Proportion of correct responses (proportion hits + proportion correct rejections). B) Perceptual sensitivity (d-prime). C) Decision bias ($criterion_{center}$). Grey dots represent values of individual participants. Inside each violin plot, the encircled black dot reflects the median, thick box indicates quartiles, thin line indicates quartiles ± 1.5 × inter-quartile range.

signal (match-masker vs. no-touch: $z = -4.260$, $p < 0.001$, $r = 0.615$; no-touch vs. match-signal: $z = -3.720$, $p < 0.001$, $r = 0.537$; match-masker vs. match-signal: $z = -4.287$, $p < 0.001$, $r = 0.619$). Crucially, we found significantly lower sensitivity (d') for match-masker relative to no-touch ($z = -3.571$, $p < 0.001$, $r = 0.515$) and relative to match-signal ($z = -4.286$, $p < 0.001$, $r = 0.619$), but no significant difference between no-touch and match-signal ($z = -1.095$, $p = 0.274$, $r = 0.158$). Thus, audio-tactile congruence produced a deleterious attentional enhancement of the masker stream relative to baseline (divided attention), but there was no additional sensitivity benefit beyond baseline when attention was directed to the signal stream. However, there was a significant left shift of decision bias ($criterion_{center}$) for match-signal relative to match-masker ($z = -4.286$, $p < 0.001$, $r = 0.619$) and no-touch ($z = -3.619$, $p < 0.001$, $r = 0.522$). In other words, decision bias moved towards zero and consequently participants were less biased towards “no” responses when the tactile stream matched the auditory signal stream (for a plot of hits and false alarms, see Figure 5.7).

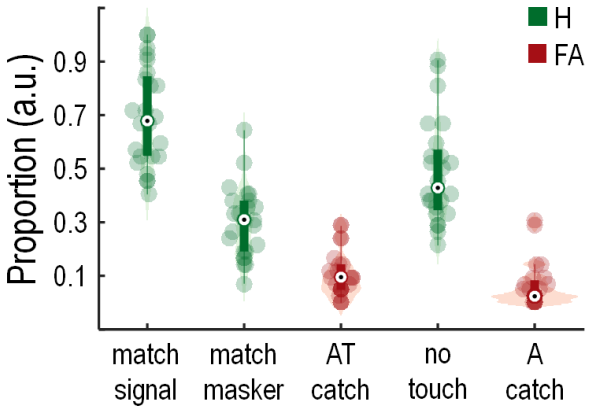

Figure 5.7: Hits and false alarms of experiment 1

Proportion of hits (H) in green and proportion of false alarms (FA) in red for each experimental condition. Coloured dots represent values of individual participants. Inside each violin plot, the encircled black dot reflects the median, thick box indicates quartiles, thin line indicates quartiles ± 1.5 × inter-quartile range. AT: audio-tactile; A: auditory.

Condition	**Proportion correct mean (±SEM)**	**d-prime mean (±SEM)**	**Criterion$_{central}$ mean (±SEM)**
Match-signal	0.795 (±0.021)	2.338 (±0.287)	0.400 (±0.139)
Match-masker	0.598 (±0.013)	1.010 (±0.179)	1.064 (±0.112)
No-touch	0.711 (± 0.020)	2.061 (± 0.234)	1.050 (± 0.125)

Table 5.1: Results of experiment 1

Group mean (±SEM) proportion of correct responses (proportion hits + proportion correct rejections), perceptual sensitivity (d') and decision bias (criterion$_{center}$) for each experimental condition.

Importantly, the match-masker and no-touch conditions were balanced in terms of number of targets (which appeared only in the auditory modality) and therefore changes of perceptual sensitivity can be unambiguously attributed to attention orienting. On the contrary, when comparing the match-signal condition to the match-masker and no-touch conditions, differences in sensitivity or decision bias might have originated from the summation of

auditory and tactile target information relative to auditory information alone, rather than from deployment of attention towards the signal stream via audio-tactile congruence. In experiment 2 we addressed this possibility.

5.2.9 Experiment 2

Results are shown in Figure 5.8 and summarized in Table 5.2. Proportion of correct responses was significantly higher for audio-tactile stimulation compared to auditory stimulation alone ($z = 2.778$, $p = 0.005$, $r = 0.401$). Crucially, this was not mirrored by changes of sensitivity ($z = 0.200$, $p = 0.841$, $r = 0.029$). Instead, there was a significant left shift of criterion for audio-tactile stimulation relative to auditory stimulation ($z = -4.286$, $p < 0.001$, $r = 0.619$). In other words, decision bias moved towards zero and consequently participants were less biased towards “no” responses when they were exposed to redundant audio-tactile stimuli relative to unisensory auditory stimuli (for a plot of hits and false alarms, see Figure 5.9).

Condition	**Proportion correct mean (±SEM)**	**d-prime mean (±SEM)**	**Criterion$_{central}$ mean (±SEM)**
AT stimulation	0.761 (±0.022)	1.944 (±0.251)	0.253 (±0.140)
A stimulation	0.677 (±0.019)	1.942 (±0.253)	1.144 (±0.146)

Table 5.2: Results of experiment 2

Group mean (±SEM) proportion of correct responses (proportion hits + proportion correct rejections), perceptual sensitivity (d') and decision bias (criterion$_{center}$) for each experimental condition. AT: audio-tactile; A: auditory.

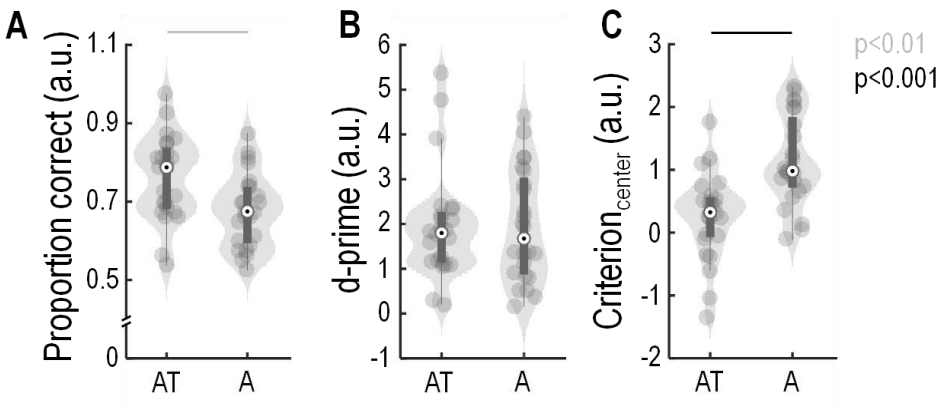

Figure 5.8: Results of experiment 2

A) Proportion of correct responses (proportion hits + proportion correct rejections). B) Perceptual sensitivity (d-prime). C) Decision bias (criterion$_{center}$). Grey dots represent values of individual participants. Inside each violin plot, the encircled black dot indicates the median, thick box indicates quartiles, thin line indicates quartiles ± 1.5 × inter-quartile range. AT: audio-tactile stimulation; A: auditory stimulation.

These results have important implications for experiment 1. On the one hand, they confirm that the summation of auditory and tactile target information determines a more liberal response criterion relative to auditory targets alone. Thus, these results sustain the view that the more liberal responses in the match-signal condition in experiment 1 originated from the summation of auditory and tactile target information. On the other hand, the present results clarify that redundant auditory and tactile target information cannot determine changes of sensitivity. Hence, changes of d' between match-signal and match-masker conditions in experiment 1 cannot be explained by summation of auditory and tactile target information. Instead, changes of d' truly reflected attentional capture via cross-modal binding.

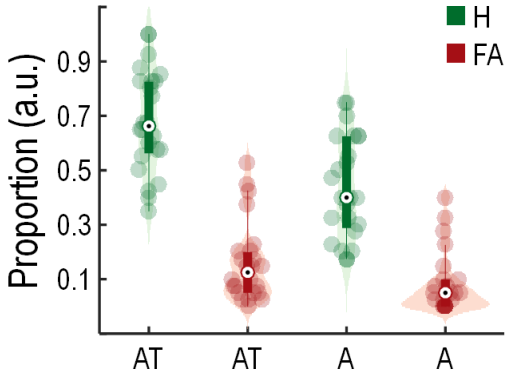

Figure 5.9: Hits and false alarms of experiment 2

Proportion of hits (H) in green and proportion of false alarms (FA) in red for each experimental condition. Coloured dots represent values of individual participants. Inside each violin plot, the encircled black dot reflects the median, thick box indicates quartiles, thin line indicates quartiles ± 1.5 × inter-quartile range. AT: audio-tactile stimulation; A: auditory stimulation.

5.4 Discussion

The present study investigated the role of cross-modal binding in resolving attentional competition during naturalistic listening in a cocktail-party scenario. Specifically, we first assessed that participants were able to perceive audio-tactile congruence (via a preliminary screening) and we subsequently used cross-modal binding to evaluate whether it enhanced selective attention in a cocktail-party scenario, via direct evaluation of perceptual changes in target detection. Moreover, we employed custom-composed unknown music pieces to dissociate perceptual from linguistic factors (which are confounded when employing speech stimuli). Importantly, the perception of a cross-modal object relies on the automatic binding of sensory signals via corresponding features, such us temporal coherence (Atilgan et al., 2018; Bizley et al., 2016; Maddox et al., 2015; Noppeney & Lee, 2018; Shamma et al., 2011). Here, we show that corresponding envelope and frequency information across time in the auditory and tactile modalities produce strong congruence judgments (in our preliminary

screening) and thus support cross-modal binding. Our results are in line with evidence that audition and touch convey redundant temporal and spectral information (Soto-Faraco & Deco, 2009), which can be integrated to better parse naturalistic vibratory stimuli such as music (Huang et al., 2012; Tranchant et al., 2017) and speech (Drullman & Bronkhorst, 2004; Fletcher et al., 2018; Huang et al., 2017; Riecke et al., 2019).

In experiment 1, where we established competition for auditory attention via use of concurrent signal and masker streams, target detectability changed as a function of cross-modal binding. Specifically, audio-tactile congruence produced a deleterious attentional enhancement of the masker stream relative to baseline divided attention and relative to the match-signal condition; on the contrary, there was no beneficial attentional enhancement of the signal stream beyond baseline. Nevertheless, we do not rule out this possibility under different experimental contexts. Firstly, increasing the level of background noise may render target detection during divided attention more challenging, hence leaving space for a beneficial attentional enhancement via cross-modal binding (Bizley et al., 2016; Crosse et al., 2016; Ross et al., 2007; van de Rijt et al., 2019). Secondly, an alternative baseline condition that includes a tactile stream unmatched to the signal and the masker streams could represent a more balanced perceptual scenario against which to compare congruence-driven attentional enhancement of target detectability (Maddox et al., 2015).

Analysis of decision bias revealed less conservative responses when participants perceived targets redundantly in the auditory and tactile modalities, as shown in the match-signal condition in experiment 1 and as directly confirmed in experiment 2. Accordingly, previous evidence shows more liberal response criteria for simultaneous congruent multisensory relative to unisensory stimulation (Frassinetti et al., 2002a; Lovelace et al., 2003; Marks et al., 2003; Odgaard et al., 2003), underscoring the impact of cross-modal

redundancy on uncertainty reduction (Ernst & Bülthoff, 2004). Hence, future experiments targeting perceptual benefits of multisensory congruence should employ designs that control for the impact of cross-modal redundancy on decision strategies (Alais et al., 2010). At the same time, it is worth remembering that multisensory congruence can also boost perceptual sensitivity; crucially, this is optimised via employment of subject-specific near-threshold target stimuli (Eramudugolla et al., 2011; Frassinetti et al., 2002a; Hoefer et al., 2013; Lovelace et al., 2003; Noesselt et al., 2008), in line with the principle of inverse effectiveness (Stein et al., 2009).

As a whole, the present study demonstrates that not only unisensory (Humphreys & Riddoch, 2003; Kimchi et al., 2007; Yeshurun et al., 2009) but also multisensory perceptual units are salient entities that capture attention. In particular, we show that cross-modal binding, free of linguistic confounds, enhances auditory selective attention and consequently impacts listening in a cocktail-party scenario. We suggest that object-based spread of attention may additionally enhance the detectability of all the features belonging to the same cross-modal object (Bizley et al., 2016; Maddox et al., 2015), however future studies targeting such process should better control for the establishment of cross-modal binding in the first place. Nevertheless, here we reinforce the idea that multisensory objects represent salient events promoting object-based attentional selection during naturalistic listening under competition for processing resources. Our results are also in line with an abundant body of evidence suggesting pre-attentive capture of spatial attention by multisensory cues under high competition for attentional resources due to dual-task (Ho et al., 2009; Santangelo et al., 2008; Santangelo & Spence, 2007) and visual search (Matusz & Eimer, 2011; Van der Burg et al., 2008; 2011) conditions. Importantly, naturalistic vision represents a similar computational challenge in terms of competition for attention (Kaiser et al., 2019; Peelen & Kastner, 2014).

Thus, future studies should address the extent to which cross-modal binding effects on complex scene analysis generalise across sensory modalities. Critically, the strength of perceptual organization into a coherent unit determines the strength of object-based attentional capture in unisensory contexts (Kimchi et al., 2016). Thus, it would be important to replicate the same finding in the present cross-modal context by parametrically modulating the degree of multisensory congruence (e.g. temporal asynchrony, Riecke et al., 2019) and subsequently testing for gradual changes of target detectability in cocktail-party conditions. Furthermore, future neuroimaging studies should aim to unveil the neural mechanisms underlying cross-modal binding and its influence on attention. Initial evidence in anesthetised rodents suggests the causal involvement of direct cortico-cortical connections between early sensory areas in the automatic formation of multisensory objects (Atilgan et al., 2018). Thus, it would be important to corroborate such effects in the human brain. Critically, in order to support the present behavioural evidence that cross-modal binding captures attention, it will be necessary to probe the recruitment of the fronto-parietal network for control of object-based attention (Corbetta et al., 2008; Corbetta & Shulman, 2002; Serences et al., 2004; Shomstein & Behrmann, 2006).

CHAPTER 6

CROSS-MODAL BINDING IN AUDITORY CORTEX RECRUITS THE ATTENTION NETWORK WITHIN A COCKTAIL-PARTY SCENARIO

Ambra Ferrari, Giulio Degano, Uta Noppeney

Computational Cognitive Neuroimaging lab, Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, B15 2TT Birmingham, UK

Citation:

Ferrari, A., Degano, G. & Noppeney, U. (in preparation). Cross-modal binding in auditory cortex recruits the attention network within a cocktail-party scenario.

Authors contributions:

Experiment conceptualisation and design: Ambra Ferrari, Giulio Degano, Uta Noppeney.
Data collection: Ambra Ferrari.
Data analysis: Ambra Ferrari (supervised by Uta Noppeney).
Writing: Ambra Ferrari (supervised by Uta Noppeney).

Abstract

In everyday situations, we are constantly bombarded with a myriad of sensory inputs that tax our limited processing resources. Crucially, coherent cross-modal information captures attention and amplifies selective tracking of all the features pertaining to the multisensory perceptual unit. The present fMRI study evaluated the neural underpinnings of cross-modal binding and its relationship with attentional selection in the human brain, during naturalistic listening free of linguistic confounds. Participants passively perceived naturalistic music pieces either in the unisensory auditory or tactile modality or via the combination of the two. Importantly, we first independently assessed that participants perceived multisensory congruence. Subsequently, we tested for superadditivity as a neural marker of cross-modal binding; moreover, we compared conditions with or without competition for auditory attention (i.e. within or outside an auditory cocktail-party scenario) to probe the influence of cross-modal binding on attentional selection. Different superadditive integration profiles were identified in bilateral auditory cortices and in a network of association areas implicated in music perception, irrespective of the auditory scenario. Crucially, cross-modal binding recruited a bilateral posterior parietal network for control of object-based attention selectively within a cocktail-party scenario. As a whole, the present study suggests that early interactions at the bottom of the sensory hierarchy promote further analysis in association areas and trigger attentional orienting during competition for processing resources.

Keywords

Multisensory integration, selective attention, cocktail-party scenario, fMRI, superadditivity

6.1 Introduction

In cocktail-party scenarios (Cherry, 1953) human listeners show the remarkable ability to selectively track one stream of information among those which concurrently compete for limited attentional resources (Bronkhorst, 2015). It is well established that cross-modal information delivered by lip movements enhances selective listening of the corresponding speech stream (Bernstein et al., 2004; Grant & Seitz, 2002; Sumby & Pollack, 1954). In particular, the benefits of lip-reading increase under highly challenging listening conditions (Crosse et al., 2016; Helfer & Freyman, 2005; Ross et al., 2007; van de Rijt et al., 2019; Zion Golumbic et al., 2013), suggesting that cross-modal information may reduce perceptual ambiguity and support selective tracking of the observed speaker (Zion Golumbic, et al., 2013). Accordingly, emerging evidence (Maddox et al., 2015; Ferrari et al., in preparation) shows that the automatic grouping of coherent cross-modal features into a multisensory perceptual unit (i.e. cross-modal binding, Bizley et al., 2016) promotes attentional selection in cocktail-party scenarios. In particular, multisensory objects capture attention (Ferrari et al., in preparation) and thus enhance selective listening of the auditory stream containing coherent cross-modal information. Moreover, multisensory objects trigger cross-modal spread of object-based attention (Maddox et al., 2015) and therefore amplify selective tracking of all the features pertaining to the multisensory object (for a cautionary note, see Ferrari et al., in preparation).

Such results raise the critical question of how cross-modal binding occurs and consequently impacts attention at the neural level. It has been widely demonstrated that cross-modal interactions emerge already at the bottom of the cortical hierarchies (Driver & Noesselt, 2008; Foxe & Schroeder, 2005; Ghazanfar & Schroeder, 2006; Kayser &

Logothetis, 2007; Noppeney et al., 2018; Schroeder & Foxe, 2005). Importantly, driving or modulatory effects of cross-modal stimuli in low-level sensory areas are supported by direct cortico-cortical anatomical connections (Musacchia & Schroeder, 2009; Schroeder et al., 2003) and reciprocal stimulus-driven entrainment (Kayser et al., 2010, 2008; Lakatos et al., 2007; Senkowski et al., 2008). Such early multisensory interactions may boost the perceptual salience of coherent cross-modal inputs during sensory segmentation (Laurienti et al., 2002; Lewis & Noppeney, 2010; Stanford & Stein, 2007; Werner & Noppeney, 2010) and thus orient subsequent elaboration of more complex representations in association areas (Foxe & Schroeder, 2005; Lewis & Noppeney, 2010; Noppeney et al., 2018; Werner & Noppeney, 2010). In particular, multisensory interactions in early sensory areas may promote figure-ground segregation processes via detection of cross-modal temporal coherence (Shamma et al., 2011). As a whole, it is then conceivable that early multisensory interactions between sensory cortices represent the neural underpinning of cross-modal binding (Bizley et al., 2016). Consistently, a recent electrophysiological investigation with anesthetised ferrets (Atilgan et al., 2018) has demonstrated that cortico-cortical interactions among visual and auditory areas causally determine the automatic representation of a perceptual scene in auditory cortex. Specifically, visually-induced phase entrainment of local field potentials reinforced the neural representation of coherent cross-modal features in auditory cortex and such effect disappeared when deactivating primary visual areas via a cooling procedure. Accordingly, it has been shown that lip movements enhance the tracking of a congruent speech stream in human auditory cortex (Crosse et al., 2015, 2016; Zion Golumbic et al., 2013). However, the use of speech stimuli limits the interpretation of such results in terms of pure cross-modal binding, as perceptual and linguistic processes are intimately confounded (Maddox et al., 2015; Ferrari et al., in preparation).

The present functional magnetic resonance imaging (fMRI) study aimed to characterise the neural implementation of cross-modal binding in humans, during naturalistic listening free of linguistic confounds. To this end, we employed custom-composed unknown music pieces because (i) they allow to elicit cross-modal binding via multisensory temporal coherence (Bizley et al., 2016; Noppeney & Lee, 2018; Shamma et al., 2011) and (ii) they avoid linguistic confounds that could influence speech intelligibility (Broderick et al., 2019; Davis & Johnsrude, 2007; Hannemann et al., 2007; Kuperberg & Jaeger, 2016; Mattys et al., 2012). Furthermore, we aimed to elicit cross-modal binding via paired auditory and tactile information based on a two-fold rationale. Firstly, audio-tactile interactions are functionally relevant in relation to real-life vibratory events (Soto-Faraco & Deco, 2009), such as music (Huang et al., 2012; Tranchant et al., 2017) and speech (Drullman & Bronkhorst, 2004; Fletcher et al., 2018; Huang et al., 2017; Riecke et al., 2019). Secondly, numerous anatomical studies with primates (Cappe & Barone, 2005; de la Mothe et al., 2006a, 2006b; Hackett et al., 2007; Smiley et al., 2007) and functional studies with primates and humans (Foxe et al., 2000, 2002; Fu et al., 2003; Hoefer et al., 2013; Kayser et al., 2005; Lakatos et al., 2007; Schroeder & Foxe, 2002; Schroeder et al., 2001; Schürmann et al., 2006) indicate convergence and modulatory effects of somatosensory inputs in auditory cortex. Therefore, it is plausible that audio-tactile interactions in auditory areas support the emergence of cross-modal binding, mirroring the case of audio-visual pairings (Atilgan et al., 2018).

In the context of fMRI, response non-linearities such as superadditivity (i.e. response to multisensory stimuli greater than the sum of responses to unisensory stimuli) are considered the most stringent indicator of multisensory integration (James & Stevenson, 2012; Noppeney, 2012). In particular, there is evidence of superadditivity in early auditory cortices in response to audio-visual (Calvert et al., 2000; Laurienti et al., 2002; Werner & Noppeney,

2010) and audio-tactile (Hoefer et al., 2013) stimulation and such effects are sustained by direct effective connectivity among the corresponding sensory areas (Hoefer et al., 2013; Werner & Noppeney, 2010). Based on these premises, we tested for superadditive integration profiles as a neural marker of cross-modal binding during naturalistic listening. Importantly, we first assessed that participants perceived audio-tactile congruence (via a preliminary screening) and we subsequently exploited cross-modal binding to unambiguously characterise its neural signature[1]. Critically, we also investigated the relationship between cross-modal binding and attention at the neural level, in order to elucidate the neural mechanisms underlying attentional capture by multisensory objects (Ferrari et al., in preparation). In particular, we hypothesised that under conditions of attentional competition cross-modal binding may recruit the fronto-parietal network for control of object-based attention (Corbetta et al., 2008; Corbetta & Shulman, 2002; Serences et al., 2004; Shomstein & Behrmann, 2006). To test this hypothesis, we compared cross-modal binding (i.e. superadditivity) in conditions with or without competition for auditory attention (i.e. within or outside an auditory cocktail-party scenario).

6.2 Materials and methods

6.2.1 Participants

Twelve participants (3 males; mean age 27.75, range 22-34 years) were included in the experiment based on a priori inclusion criteria (see Section 6.2.7). No extra volunteers were excluded. Sample size was determined based on previous neuroimaging experiments that used similar experimental designs and analysis approaches to those planned for the present

[1] Perceived multisensory congruence such as temporal coherence determines cross-modal binding (Bizley et al., 2016; Noppeney & Lee, 2018).

experiment (Alluri et al., 2012, 2013; Hoefle et al., 2018; Sankaran et al., 2018; Santoro et al., 2017; Toiviainen et al., 2014). All volunteers reported normal or corrected to normal vision, normal hearing and touch and no history of neurological or psychiatric conditions; they were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971; mean laterality index: 85; range: 60–100); they had never received any formal music training and were classified as non-musicians via the Music USE (MUSE) Questionnaire (Chin & Rickard, 2012), based on duration, frequency and regularity of instrument playing[2]. Participants provided written informed consent and were naïve to the aim of the study; they received a monetary reimbursement for their participation in the experiment. The study was approved by the University of Birmingham Ethical Review Committee and was conducted in accordance with these regulations.

6.2.2 Stimuli

Auditory stimuli (Figure 6.1A) consisted of 28 s monophonic music pieces, custom-composed in collaboration with a composer and synthesized from Musical Instrument Digital Interface (MIDI) files using Linux MultiMedia Studio 1.1.3 (LMMS, https://lmms.io/) with a piano sound font (grand-piano-YDP-20160804). Tactile stimuli consisted of the same 28 s music pieces, synthesized using LMMS with sinusoidal oscillations (TripleOscillator). As a result, we obtained corresponding envelope and frequency information across audition and touch for each music piece (mean sound pressure level: 75dB; frequency range: 1-500Hz).

[2] Index of Music Instrument Playing (IMIP) < 0.4, where IMIP = [Years of instrument playing x Hours of practice per day / Regularity of practice] and Regularity of practice ("How long since you last regularly played music?") scored as follows: 1 for "less than one week", 2 for "less than one month", 4 for "less than one year", 8 for "between 1 and 5 years", 16 for "between 5 and 10 years", 32 for "more than 10 years".

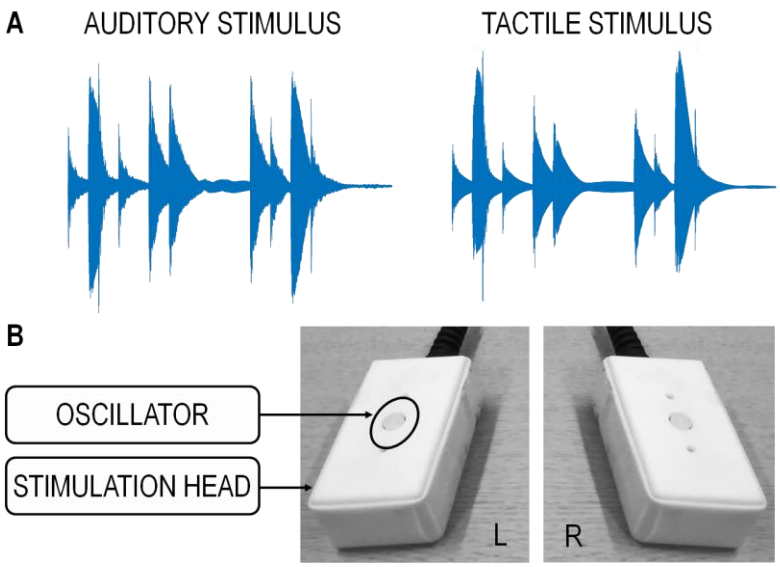

Figure 6.1: Experimental stimuli

A) Auditory and tactile signals provided correspondent envelope and frequency information for each music piece. B) Piezoelectric system. A stimulation head (rectangular box) was applied to each hand, with the fingertip of each index finger in correspondence with the oscillator (encircled). The oscillator provided vibrations by moving up and down. L: left; R: right.

Auditory and tactile synthesized stimuli were recorded at 44100 Hz with 16-bit resolution, normalised and saved as WAV files using Audacity 2.1.2. For cocktail-party conditions, auditory monophonic pieces were combined into two-stream polyphonic pieces via simultaneous recording.

6.2.3 Experimental design and procedure

The experimental design comprised five experimental conditions (Figure 6.2A): auditory ("A"), with a monophonic piece in the auditory modality; tactile ("T"), with a monophonic piece in the tactile modality; audio-tactile ("AT"), with the same monophonic piece concurrently presented in the auditory and tactile modalities; auditory cocktail-party ("Acp"), with two paired monophonic pieces concurrently presented in the auditory modality; audio-tactile cocktail-party ("AcpT"), with a monophonic piece in the tactile modality matching one

of two paired monophonic pieces concurrently presented in the auditory modality. Each run comprised 15 stimulation blocks (duration: 28 s) interleaved with 15 fixation blocks (duration: 6 s). In each stimulation block (Figure 6.2B), participants were exposed to one of the five experimental conditions. For each participant and experimental run, the order of stimulation blocks was pseudo-randomised with the following constraints: consecutive blocks always contained different experimental conditions; if condition “Y” followed condition “X” in one run, the reverse sequence was presented in another run. In this way, we sought to minimise participants' anticipation and habituation processes, as well as counterbalancing the effect of fatigue across experimental conditions.

Participants were instructed to lay still inside the scanner and passively experience the stimulation with their eyes closed. To monitor participants’ vigilance, we employed a visual oddball task (Figure 6.2C): during stimulation blocks, participants reported the occasional appearance of full-screen light-grey flashes (luminance: 85 cd/m^2; duration: 50 ms) via pedal press with their right foot. Luminance was adjusted in order to optimise flashes' visibility with eyes closed. Participants were instructed to press the pedal as soon as they noticed a flash and keep their concentration on the audio-tactile stimulation. In each run, 5 stimulation blocks contained flashes (3 blocks with 1 flash; 2 blocks with 2 flashes). We randomised flashes' onsets with the following constraints: no flashes within the first and last 2 seconds of a block; a minimum gap of 2 seconds between two flashes appearing in the same block. The number of flashes was counterbalanced across conditions within each participant. We checked if a response occurred between 100 ms and 2000 ms after each flash onset (Crosse et al., 2015, 2016). Participants reported flashes appearance with high accuracy (group mean ± SEM proportion of hits = 0.893 ± 0.002).

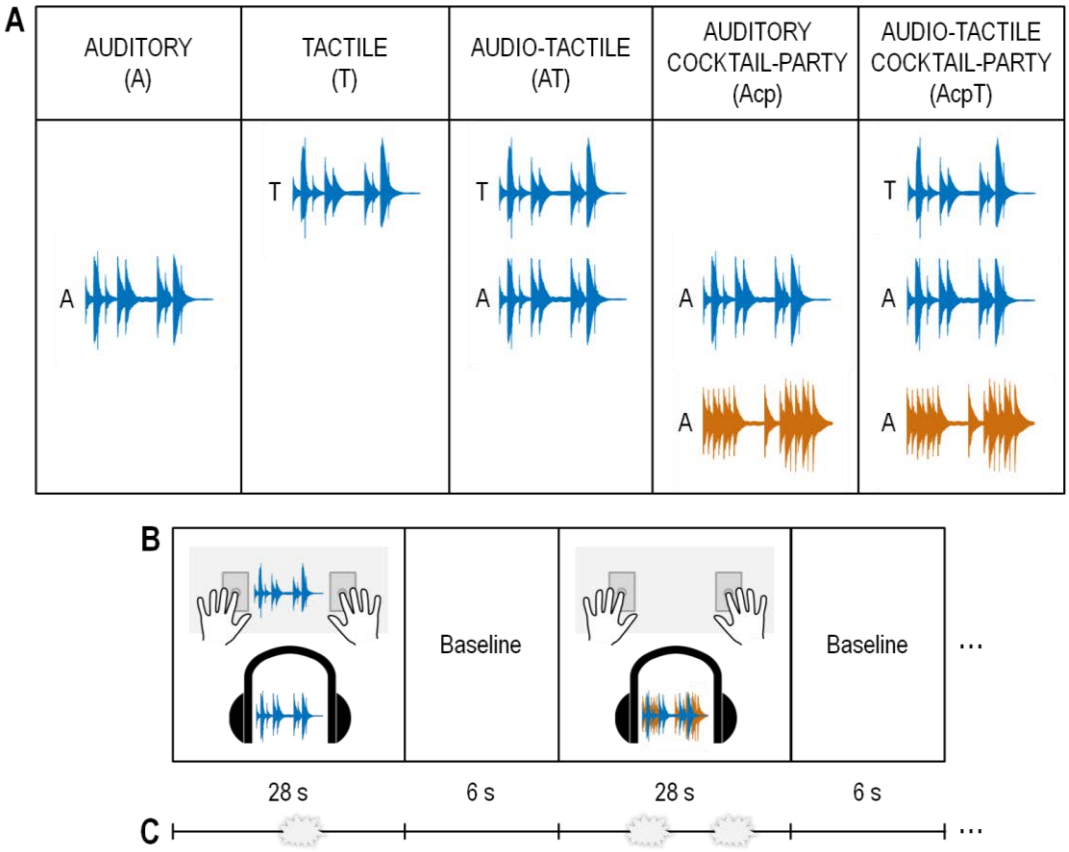

Figure 6.2: Experimental design and procedure

A) Different music pieces are depicted with different colours. B) Experimental procedure: in each run, 15 stimulation blocks (28 s duration) were interleaved with 15 baseline blocks (6 s). C) A visual oddball task was employed to monitor participants' vigilance: during stimulation blocks, participants reported the appearance of full-screen light-grey flashes (duration: 50 ms) via pedal press with their right foot.

Each participant completed 16 scanning runs over the course of 2 days (5 conditions × 3 stimulation blocks / condition / run × 16 runs = 240 stimulation blocks in total), after familiarization with stimuli and procedure via one preliminary practice run at the beginning of each scanning day.

6.2.4 Experimental setup

The experiment was presented via Psychtoolbox version 3.0.15 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) running under MATLAB R2018b (MathWorks Inc.) on a Linux machine (Ubuntu 18.04.2 LTS). Stimuli were extracted from WAV files and played via MATLAB custom-code. Auditory stimuli were played through an MR-compatible system (SOUNDPixx MRI pneumatic transducer and amplifier VPX-ACC-8100, QC Canada; MRIaudio in-ear headphones, USA) controlled via the stimulation PC's built-in soundcard. Tactile stimuli were presented through a piezoelectric system (PTS-C2, Dancer Design, UK) controlled via an external sound-card (Asus Xonar U7, Taiwan). A piezoelectric stimulation head (Figure 6.1B) was applied to each hand, with the fingertip of each index finger in correspondence with the stimulation oscillator. For multisensory conditions, we adjusted audio-tactile latencies in the presentation software and confirmed their synchrony by recording and measuring their relative latencies using two microphones. Visual stimuli were back-projected onto a Plexiglas screen using a Barco Present-C F-Series projector (F35 WUXGA, UK; 1920 × 1024 pixels resolution; 60 Hz frame rate) and they were visible to the participants via a mirror mounted on the MR head-coil (horizontal visual field of ~40° visual angle at a viewing distance of ~68 cm). Participants gave responses by pressing any keys of an MR-compatible keypad (NATA LXPAD 1×5-10M, BC Canada) attached to the right foot with elastic cohesive bandage and secured via foam supports.

6.2.5 MRI data acquisition

A 3T Siemens Prisma MR scanner was used to acquire both a T1-weighted anatomical image (TR = 2000 ms, TE = 2.03 ms, flip angle = 8°, FOV = 256 mm × 256 mm, 208 sagittal slices acquired in sequential ascending direction, voxel size = 1 × 1 × 1 mm^3) and T2*-

weighted axial echoplanar images (EPI) with blood-oxygenation-level-dependent contrast (gradient echo, multiband factor of 2, TR = 1550 ms, TE = 35 ms, flip angle = 71°, FOV = 210 × 210 × 150 mm^2, 60 axial slices acquired in interleaved ascending direction, voxel size = 2.5 × 2.5 × 2.5 mm^3, no interslice gap). For each participant, a total of 400 volumes × 16 runs were acquired, after discarding the first four volumes of each run to allow for T1 equilibration effects. Functional data acquisition was performed over the course of 2 days and the anatomical image was acquired at the end of the first day.

6.2.6 Experimental data analysis

MRI data were analysed using SPM12 (Wellcome Department of Imaging Neuroscience, London; www.fil.ion.ucl.ac.uk/spm; Friston et al., 1994a). Scans from each participant were realigned (using the first scan as reference) and unwarped, spatially normalised into Montreal Neurological Institute (MNI) space using normalisation parameters from segmentation of the T1 structural image (Ashburner & Friston, 2005), resampled to a spatial resolution of 2 × 2 × 2 mm^3 and spatially smoothed with a Gaussian kernel of 8 mm full-width at half-maximum. A high-pass filter (1/128 Hz cutoff) was applied to the time series in each voxel.

In a blocked design, unit impulses representing stimulation blocks onsets (duration: 28 s) were convolved with a canonical hemodynamic response function. The 6 experimental conditions were included as regressors in the design matrix. Onsets of all flashes (duration: 0 s) were included as a separate nuisance regressor (see Section 8.3.1 for the corresponding fMRI results). Realignment parameters were also added as nuisance covariates to account for noise due to residual head motion artefacts. The voxel-wise magnitude of the BOLD signal in response to each stimulation block was defined by the parameter estimates pertaining to the canonical hemodynamic response function. Following a hierarchical summary statistics

approach, subject-specific images were entered into a first-level general linear model and contrasts (each experimental condition versus baseline summed over the sixteen runs) were passed to a second-level ANOVA, where contrasts of interest were defined. Following random effect analysis, inferences were made at the second level (Friston et al., 1994a).

To assess cross-modal binding at the neural level, we tested for superadditive integration profiles (James & Stevenson, 2012; Noppeney, 2012). Specifically, we evaluated whether the BOLD response to bisensory audio-tactile stimulation was greater than the sum of BOLD responses to unisensory auditory and tactile stimulation. In other words, we set the null hypothesis of linear response additivity (i.e. superposition, see Section 2.3.1), which represents mere response convergence of independent unisensory neuronal populations, and we tested for the presence of response non-linearity (in particular, superadditivity), which characterises proper multisensory neural populations (Laurienti et al., 2005). The contrast AT > A+T evaluated superadditivity outside a cocktail-party scenario, i.e. in the absence of competition for auditory attention; the contrast AcpT > Acp+T evaluated superadditivity within a cocktail-party scenario, i.e. in the presence of competition for auditory attention. Whole-brain activations are reported at $p < 0.05$ (Family-Wise Error corrected) at the cluster level, with an auxiliary uncorrected peak-level threshold of $p < 0.001$ (Friston et al., 1994b).

6.2.7 Inclusion criteria

Via a preliminary screening (see Section 6.2.8) we independently verified that participants perceived multisensory congruence. Volunteers were pre-selected based on the ability to perceptually discriminate audio-tactile congruence (i.e. same music piece across audition and touch) and incongruence (i.e. two different music pieces across audition and touch). Volunteers who showed d' > 2.8 were included in the study (group mean ± SEM d' for

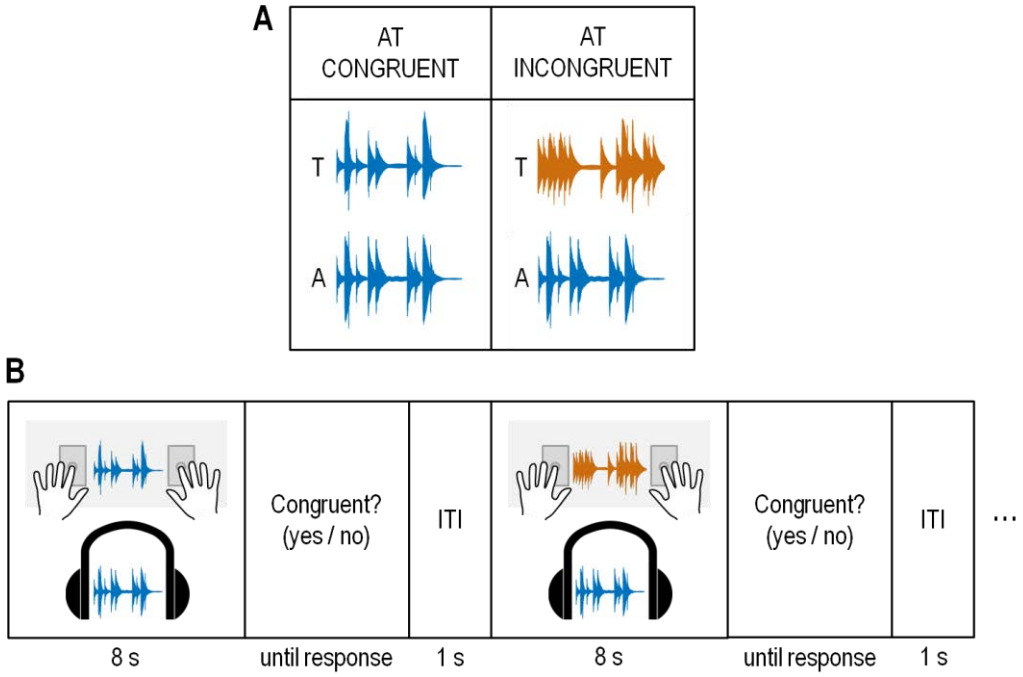

Figure 6.3: Screening design and procedure

A) The experimental design comprised 2 conditions: audio-tactile congruence (i.e. same piece across audition and touch) and incongruence (i.e. two different pieces across audition and touch). B) Experimental procedure: after synchronous auditory and tactile stimuli presentation (8 s piano pieces), participants reported via pedal press whether stimuli were congruent (left foot) or not (right foot). A: auditory; T: tactile; AT: audio-tactile; ITI: inter-trial interval.

included participants = 5.409 ± 0.533; $criterion_{center}$ = 0.093 ± 0.206; proportion of correct responses = 0.966 ± 0.011)[3].

6.2.8 Screening

In a yes-no congruence judgement paradigm (Figure 6.3), we presented synchronous auditory and tactile stimuli that were either congruent or incongruent (duration: 8 s). Background white noise was additionally played through the headphones (65 dB sound pressure level) to mask the sound of tactile vibrations. After stimuli presentation, participants reported whether they

[3] Threshold was defined as two standards deviations below the group mean d' in a preliminary pilot study with 9 participants.

perceived the same music piece through audition and touch via pedal press (yes: left pedal; no: right pedal). Focus was put on accuracy and there was no time limit for the response. Participants were instructed to sit still in a dimly lit cubicle with their eyes closed and their head positioned on a chin rest. Responses were collected via two pedals (SODIAL, Shenzhen IMC Digital Technology Co.), one in correspondence of each foot. After familiarization with stimuli and procedure via one preliminary practice run, each participant completed 2 experimental runs (2 conditions × 15 trials / condition / run × 2 runs = 60 trials in total). Based on signal detection theory (Wickens, 2002), 'yes' and 'no' responses in congruent trials were classified as hits and misses, whereas 'yes' and 'no' responses in incongruent trials were classified as false alarms and correct rejections. Consequently, we calculated d' (Z(hit rate) - Z(false alarm rate)), $\text{criterion}_{\text{center}}$ (- [Z(false alarm rate) + Z(hit rate)] / 2) and proportion of correct responses (proportion hits + proportion correct rejections).

6.3 Results

To investigate the neural mechanisms underlying cross-modal binding and its relationship with attention, we performed the following analyses. Firstly, we characterised superadditivity separately within or outside a cocktail-party scenario. Secondly, we evaluated the effect of superadditivity jointly across these two auditory contexts. Finally, we determined the effect of superadditivity selectively within a cocktail-party scenario relative to outside a cocktail-party scenario.

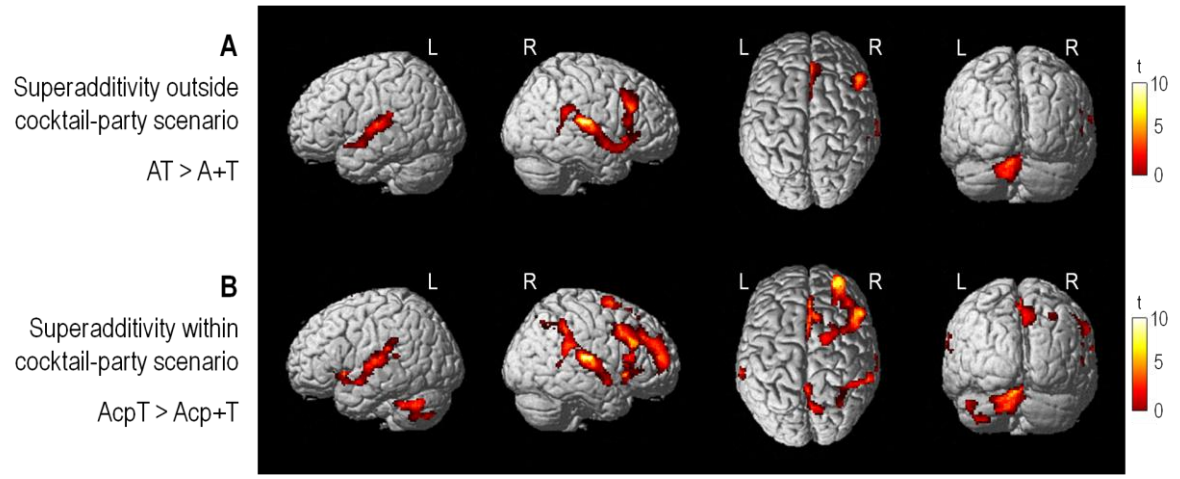

Figure 6.4: Superadditivity separately within or outside a cocktail-party scenario

Activation increases are rendered on a canonical brain ($p < 0.05$ FWE-corrected at cluster level, with auxiliary uncorrected peak-level threshold of $p < 0.001$). L: left; R: right; A: auditory; T: tactile; AT: audio-tactile; Acp: auditory cocktail-party; AcpT: audio-tactile cocktail-party.

6.3.1 Superadditivity separately within or outside a cocktail-party scenario

Brain regions that showed a superadditive integration profile outside a cocktail-party scenario are shown in Figure 6.4A and summarised in Table 6.1. In line with our hypothesis, we found strong evidence of superadditivity in early auditory areas (i.e. bilateral transverse temporal gyri). Significant superadditive integration profiles were also present bilaterally in the planum temporale and posterior insula, in the left medial posterior cerebellum and in the right inferior frontal gyrus, medial superior frontal gyrus and dorsal anterior cingulate gyrus.

Consistent results were found when testing for superadditivity within a cocktail-party scenario, as shown in Figure 6.4B and summarised in Table 6.2. Crucially, additional activations were located in the right precuneus and intraparietal sulcus, which are part of a widespread fronto-parietal network implicated in control of object-based attention (Corbetta et al., 2008; Corbetta & Shulman, 2002; Serences et al., 2004; Shomstein & Behrmann, 2006).

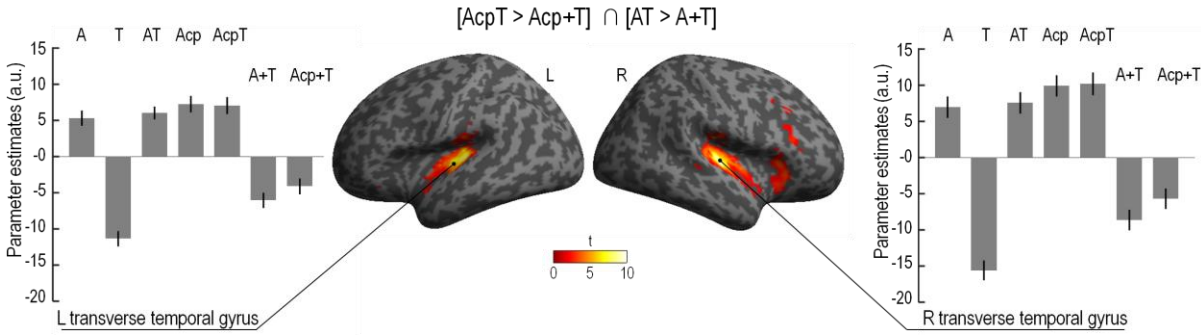

Figure 6.5: Superdditivity across auditory contexts in bilateral transverse temporal gyri

Conjunction within and outside cocktail-party scenario. Activation increases are rendered on an inflated canonical brain ($p < 0.05$ FWE-corrected at cluster level, with auxiliary uncorrected peak-level threshold of $p < 0.001$). Bar plots represent group mean (±SEM) parameter estimates in non-dimensional units (corresponding to percentage whole-brain mean). Source of anatomical labels: Duvernoy (1999). L: left; R: right; A: auditory; T: tactile; AT: audio-tactile; Acp: auditory cocktail-party; AcpT: audio-tactile cocktail-party.

6.3.2 Superadditivity jointly across auditory contexts

To better elucidate which brain regions showed a superadditive integration profile irrespective of the auditory context, we employed a logical "AND" conjunction over superadditivity within and outside a cocktail-party scenario. Results, which are summarised in Table 6.3, confirmed that superadditive effects were primarily located in early auditory areas (i.e. bilateral transverse temporal gyri). In particular, while unisensory tactile stimulation determined reduction of BOLD response relative to baseline, such effect disappeared in the case of congruent auditory and tactile stimulation (Figure 6.5). In other words, multisensory congruence eliminated cross-modal deactivation of auditory cortex by tactile stimuli, as previously reported for audio-visual pairings (Beauchamp et al., 2004; Laurienti et al., 2002). Even though the effect was most pronounced in primary auditory cortex, it also emerged in planum temporale and posterior insula, which are closely interconnected areas for temporal

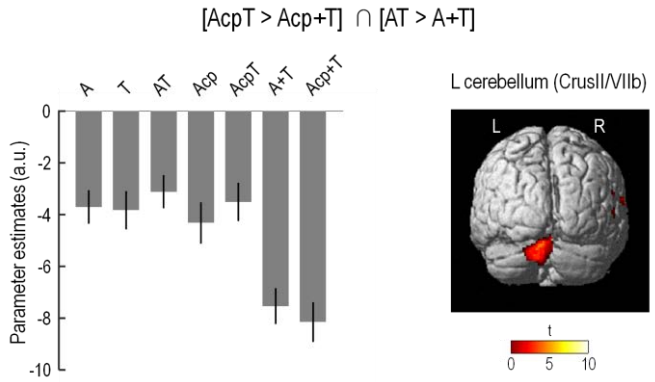

Figure 6.6: Superdditivity across auditory contexts in left medial posterior cerebellum

Conjunction within and outside cocktail-party scenario. Activation increases are rendered on a canonical brain ($p < 0.05$ FWE-corrected at cluster level, with auxiliary uncorrected peak-level threshold of $p < 0.001$). Bar plots represent group mean (±SEM) parameter estimates in non-dimensional units (corresponding to percentage whole-brain mean). Source of anatomical label: Duvernoy (1999). L: left; R: right; A: auditory; T: tactile; AT: audio-tactile; Acp: auditory cocktail-party; AcpT: audio-tactile cocktail-party.

and spectral processing in the auditory domain (Bamiou et al., 2003; Griffiths & Warren, 2002). Further brain areas including right inferior frontal gyrus, right superior frontal gyrus and left posterior cerebellum were less deactivated relative to baseline (in other words, more active) during audio-tactile stimulation relative to unisensory stimulation, with the strongest effect in the left posterior cerebellum (Figure 6.6). Importantly, these areas belong to a brain network implicated in music perception (Janata, 2015; Koelsch & Siebel, 2005; Parsons, 2001; Peretz & Zatorre, 2005). In particular, there is increasing and converging evidence that the cerebellum represents an important site of multisensory integration (Ronconi et al., 2016), especially for music information (Lee & Noppeney, 2011a; Petrini et al., 2011).

6.3.3 Superadditivity selectively within a cocktail-party scenario

To better characterise which brain regions showed a superadditive integration profile selectively within the context of competition for auditory attention, we contrasted

superadditivity within relative to outside a cocktail-party scenario. Results, which are shown in Figure 6.7 and summarised in Table 6.4, identified a widespread bilateral posterior parietal system encompassing intraparietal sulcus, superior parietal lobule and precuneus, which are central nodes of the fronto-parietal network for recruitment of selective object-based attention (Corbetta et al., 2008; Corbetta & Shulman, 2002; Serences et al., 2004; Shomstein & Behrmann, 2006).

Brain regions	**MNI coordinates (mm)**			**Cluster size (voxels)**	**z-score (peak)**	**$p_{FWE\text{-}corrected}$ (cluster)**
	x	y	z			
AT > A+T						
R transverse temporal gyrus	46	-20	10	13981	> 8	0.000
L transverse temporal gyrus	-44	-22	8		7.76	
R planum temporale	56	-26	12		6.60	
L planum temporale	-52	-22	7		5.34	
R posterior insula	44	-6	-6		5.59	
L posterior insula	-44	-8	-4		5.61	
R inferior frontal gyrus	52	26	30		5.82	
R lateral ventricle	26	-42	8		> 8	
L lateral ventricle	-24	-44	8		> 8	
L cerebellum (CrusII/VIIb)	-8	-74	-26	749	6.26	0.000
R superior frontal gyrus	4	32	40	655	4.37	0.000
R anterior cingulate gyrus	6	32	20		3.90	

Table 6.1: Superadditivity outside a cocktail-party scenario

p-values are FWE-corrected at the cluster level for multiple comparisons within the entire brain. Auxiliary uncorrected peak-level threshold of $p < 0.001$. Source of anatomical label: Duvernoy (1999). AT: audio-tactile; A: auditory; T: tactile; L: left; R: right.

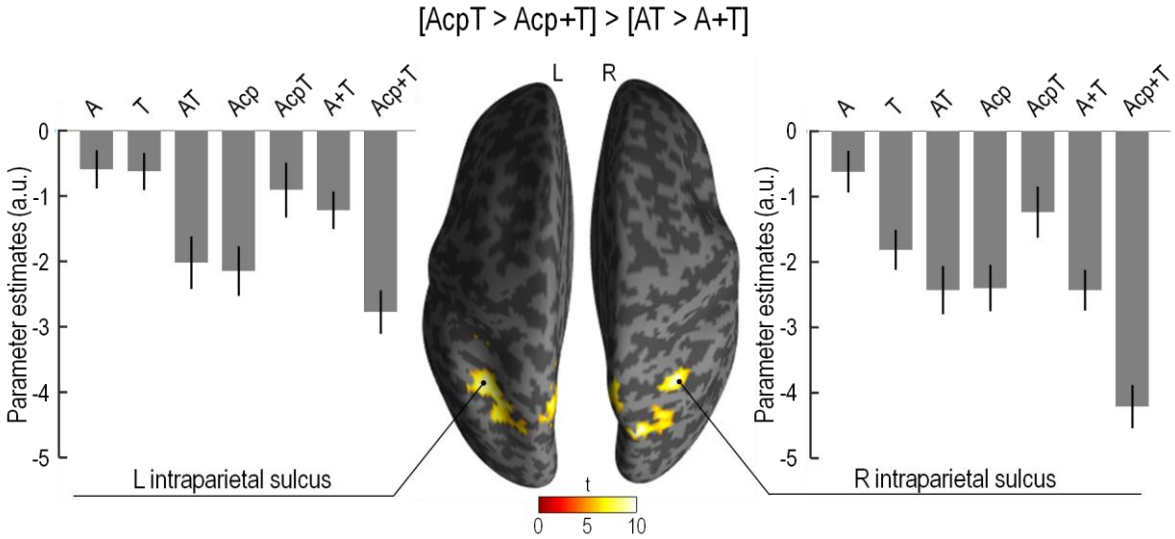

Figure 6.7: Superadditivity selectively within relative to outside a cocktail-party scenario

Activation increases are rendered on an inflated canonical brain ($p < 0.05$ FWE-corrected at cluster level, with auxiliary uncorrected peak-level threshold of $p < 0.01$). Bar plots represent group mean (±SEM) parameter estimates in non-dimensional units (corresponding to percentage whole-brain mean). Source of anatomical labels: Duvernoy (1999). L: left; R: right; A: auditory; T: tactile; AT: audio-tactile; Acp: auditory cocktail-party; AcpT: audio-tactile cocktail-party.

Brain regions	**MNI coordinates (mm)**			**Cluster size (voxels)**	**z-score (peak)**	**$p_{FWE\text{-}corrected}$ (cluster)**
	x	y	z			
AcpT > Acp+T						
R transverse temporal gyrus	46	-20	10	23143	> 8	0.000
L transverse temporal gyrus	-44	-22	8		7.35	
R planum temporale	54	-26	12		6.72	
L planum temporale	-52	-22	7		5.06	
R posterior insula	44	-6	-6		4.79	
L posterior insula	-42	-12	-4		6.20	
R middle frontal gyrus	28	52	16		5.17	
R inferior frontal gyrus	52	26	30		5.04	
R superior frontal gyrus	4	18	40		5.16	
L cerebellum (CrusII/VIIb)	-8	-72	-26		6.56	

L cerebellum (Crus I/VIIa)	-44	-58	-38		4.79	
R precuneus	10	-66	42		6.20	
R precuneus	4	-42	44		4.14	
R intraparietal sulcus	32	-50	38		4.91	
R lateral ventricle	22	-40	8		6.21	
L lateral ventricle	-24	-44	10		6.46	
R superior frontal gyrus	18	0	70	271	4.53	0.048
R superior frontal gyrus	28	8	66		3.76	

Table 6.2: Superadditivity within a cocktail-party scenario

p-values are FWE-corrected at the cluster level for multiple comparisons within the entire brain. Auxiliary uncorrected peak-level threshold of $p < 0.001$. Source of anatomical labels: Duvernoy (1999). AcpT: audio-tactile cocktail-party; Acp: auditory cocktail-party; T: tactile; L: left; R: right.

Brain regions	**MNI coordinates (mm)**			**Cluster size (voxels)**	**z-score (peak)**	**$p_{FWE\text{-}corrected}$ (cluster)**
	x	y	z			
[AT > A+T] ∩ [AcpT > Acp+T]						
R transverse temporal gyrus	46	-20	10	12522	> 8	0.000
L transverse temporal gyrus	-44	-22	8		7.35	
R planum temporale	54	-26	12		6.72	
L planum temporale	-52	-22	8		5.06	
R posterior insula	42	-12	-4		4.98	
L posterior insula	-42	-12	-4		6.08	
R inferior frontal gyrus	52	26	30		5.04	
R lateral ventricle	22	-40	8		6.21	
L lateral ventricle	-24	-44	10		6.46	
L cerebellum (CrusII/VIIb)	-8	-74	-26	731	6.24	0.001
R superior frontal gyrus	4	34	38	566	4.05	0.002

Table 6.3: Superadditivity across auditory contexts

Conjunction within and outside cocktail-party scenario. p-values are FWE-corrected at the cluster level for multiple comparisons within the entire brain. Auxiliary uncorrected peak-level threshold of $p < 0.001$. Source of anatomical labels: Duvernoy (1999). AT: audio-tactile; A: auditory; T: tactile; AcpT: audio-tactile cocktail-party; Acp: auditory cocktail-party; L: left; R: right.

Brain regions	MNI coordinates (mm)			Cluster size (voxels)	z-score (peak)	$p_{FWE\text{-}corrected}$ (cluster)
	x	y	z			
[AcpT > Acp+T] > [AT > A+T]						
R intraparietal sulcus	26	-52	38	2687	3.81	0.000
L intraparietal sulcus	-30	-54	36		3.66	
R superior parietal lobule	26	-72	46		3.65	
L superior parietal lobule	-26	-66	40		3.26	
R precuneus	2	-64	58		3.53	
L precuneus	0	-52	50		3.02	

Table 6.4: Superadditivity within relative to outside cocktail-party scenario

p-values are FWE-corrected at the cluster level for multiple comparisons within the entire brain. Auxiliary uncorrected peak-level threshold of $p < 0.01$. Source of anatomical labels: Duvernoy (1999). AcpT: audio-tactile cocktail-party; Acp: auditory cocktail-party; A: auditory; T: tactile; L: left; R: right.

6.4 Discussion

The present fMRI study evaluated the neural mechanisms underlying cross-modal binding and its relationship with attention in the human brain, during naturalistic listening free of linguistic confounds. In particular, we assessed cross-modal binding in terms of superadditive integration profiles (James & Stevenson, 2012; Noppeney, 2012), with the hypothesis of a primary involvement of early auditory areas (Atilgan et al., 2018; Bizley et al., 2016; Crosse et al., 2015, 2016; Zion Golumbic et al., 2013). Moreover, we compared superadditivity within relative to outside an auditory cocktail-party scenario to assess whether multisensory objects recruit the fronto-parietal network for control of object-based attention (Corbetta et al., 2008; Corbetta & Shulman, 2002; Serences et al., 2004; Shomstein & Behrmann, 2006) during competition for auditory attention.

In line with our hypothesis, low-level auditory areas showed the greatest effect of superadditivity, irrespective of the auditory context (i.e. within and outside an auditory cocktail-party scenario). This is in line with previous electrophysiological evidence showing modulatory effects of coherent visual stimuli in early auditory cortex for streams presented either alone (Crosse et al., 2015, 2016) or in a mixture (Atilgan et al., 2018; Zion Golumbic et al., 2013). In the present fMRI study, concurrent audio-tactile stimulation silenced the cross-modal suppression effect of unisensory tactile signals in low-level auditory areas. Similar results have previously been reported in the case of audio-visual pairings (Beauchamp et al., 2004; Laurienti et al., 2002) and are thought to represent an adaptive mechanism whereby simultaneous cross-modal signals boost activation of sensory areas (Stanford & Stein, 2007) to influence the creation of salience maps (Itti & Koch, 2000; Kayser et al., 2005) for further processing in association areas (Foxe & Schroeder, 2005; Noppeney et al., 2018). Accordingly, audio-tactile stimulation relative to unisensory stimulation was associated with decreased deactivation (in other words, more activation) of a brain network implicated in music perception (Janata, 2015; Koelsch & Siebel, 2005; Parsons, 2001; Peretz & Zatorre, 2005), which included right inferior frontal gyrus, right superior frontal gyrus (in correspondence to the supplementary motor area) and left cerebellum. Hence, we reinforce the idea that early multisensory interactions among sensory areas serve the purpose of segmenting the perceptual scene into meaningful perceptual units (Bizley et al., 2016) for subsequent elaboration of more complex task-specific representations in higher-order areas (Lewis & Noppeney, 2010; Werner & Noppeney, 2010).

Such boosting mechanism may also determine attentional capture by multisensory objects under conditions of attentional competition (Ferrari et al., in preparation). Indeed, cross-modal binding recruited brain areas implicated in orienting of object-based attention

(Corbetta et al., 2008; Corbetta & Shulman, 2002; Serences et al., 2004; Shomstein & Behrmann, 2006) in the context of a cocktail-party scenario. However, we cannot drive direct conclusions about the functional relevance of the present fMRI results, given the passive experimental design. Thus, it would now be important to assess the same questions during active listening, e.g. while performing a target detection task (Ferrari et al., in preparation). In this previous psychophysics study, we found a more liberal response criterion for targets presented redundantly across sensory modalities, both within and outside a cocktail-party scenario. Crucially, we also found changes of perceptual sensitivity as a function of multisensory congruence within a cocktail-party scenario, indexing attentional capture by multisensory objects. Hence, future fMRI studies should test the following hypotheses: on the one hand, superadditivity in auditory cortex, irrespective of the auditory scenario, may positively predict more liberal response criteria; on the other hand, superadditivity in posterior parietal areas within a cocktail-party scenario may positively predict enhanced target detectability. Importantly, we acknowledge that unisensory perceptual reliability plays a major role in shaping behavioural and neural effects. Concurrent presentation of highly degraded near-threshold unisensory stimuli boosts perceptual sensitivity (Eramudugolla et al., 2011; Frassinetti et al., 2002a; Hoefer et al., 2013; Lovelace et al., 2003; Noesselt et al., 2008, 2010) based on the principle of inverse effectiveness (Stein et al., 2009), which also drives superadditivity in low-level auditory cortex (Hoefer et al., 2013; Werner & Noppeney, 2010). In our previous psychophysics experiment (Ferrari et al., in preparation), unisensory targets were not degraded at participants' near-threshold level; accordingly, changes of perceptual sensitivity did not depend on multisensory redundancy but were instead attention-mediated within a cocktail-party scenario, suggesting the involvement of attention-related brain areas.

Given converging evidence that cross-modal stimuli increase the representation of congruent features in auditory cortex (Atilgan et al., 2018; Crosse et al., 2015, 2016; Zion Golumbic et al., 2013), future investigations should move beyond univariate analyses and probe changes of representational content in auditory cortex as a function of audio-tactile stimulation. Indeed, multivariate decoding approaches have demonstrated increased sensitivity in revealing recruitment of early sensory areas by cross-modal inputs (Liang et al., 2013). In particular, the employment of stimuli that could be classified in terms of musical features (such as pitch or rhythm) would enable testing for changes of multivariate pattern classification in auditory cortex. Crucially, future electrophysiological studies should investigate the timing of the cross-modal binding effect under cocktail-party conditions. In particular, we would expect an early instantiation of superadditivity in low-level sensory areas, followed by the recruitment of posterior parietal areas to orient object-based attention toward streams containing coherent multisensory information. Similarly, future connectivity analyses should characterise the network architecture supporting the current findings. It is conceivable that recurrent cortico-cortical connections among sensory areas (Musacchia & Schroeder, 2009; Schroeder et al., 2003) regulate the emergence of superadditivity at the bottom of the sensory hierarchy. Moreover, recurrent connections with posterior parietal areas may firstly arbitrate attentional capture by cross-modal binding and subsequently determine spread of object-based attention to all the features pertaining to the multisensory object (Atilgan et al., 2018; Bizley et al., 2016; Maddox et al., 2015; Shamma et al., 2011). Finally, if such correlational analyses were to support that superadditivity at the bottom of the sensory hierarchy triggers the recruitment of parietal areas for attentional orienting, intervention approaches (e.g. transcranial magnetic stimulation) should tackle the causality of this relationship.

CHAPTER 7

GENERAL DISCUSSION

To date, many investigations have concentrated on the relationship between attention and multisensory integration, producing a mixture of apparently contrasting results (Koelewijn et al., 2010): on the one hand, the integration of multisensory inputs seems to be mediated by attention; on the other hand, there is evidence of early and automatic multisensory interactions that in turn impact attentional orienting. The present thesis contributes to this debate by providing behavioural and neural evidence of a parallel framework whereby attention and multisensory integration synergistically interact at multiple levels of processing. This final chapter substantiates such claim. First, I will summarise and connect the main findings of Chapters 3-6 and I will discuss some related methodological considerations. Next, I will integrate the present results and background literature into a cohesive explanatory model. Finally, I will outline future directions of research inspired by this emerging framework.

7.1 Findings

7.1.1 The interplay between attention and multisensory integration at the behavioural level

In Chapters 3 and 4 of the present thesis I demonstrated the effect of endogenous modality-specific attention on multisensory perceptual inferences during audio-visual spatial localisation. Importantly, I took into account the critical distinction between external and internal selective attention (Chun et al., 2011), namely the selection and modulation of sensory information on the one hand and internally generated representations on the other hand. Such distinction is critical as it allowed unveiling additive effects of pre-stimulus focus (external attention) and post-stimulus response selection (internal attention) on multisensory perceptual inference. On the one hand, pre-stimulus focus increased the reliability of attended versus unattended representations. Hence, external selective attention represents an additional factor impacting multisensory perceptual inference alongside pure physical reliability (Rohe & Noppeney, 2015b). On the other hand, post-stimulus response selection biased final responses towards task-relevant perceptual representations, expanding previous investigations that confounded external and internal selective attention (Aller & Noppeney, 2019; Cao et al., 2019; Rohe et al., 2019; Rohe & Noppeney, 2016, 2018). This body of evidence is in contrast with previous claims that audio-visual spatial localisation is independent of endogenous (Bertelson et al., 2000) and exogenous (Vroomen et al., 2001) cross-modal spatial attention. However, the methodological validity of these earlier findings has been recently re-evaluated (Van der Stoep et al., 2015). In particular, it appears necessary to allow enough time (100-300 ms) for cross-modal attention to develop its strongest effect on perception (Driver & Spence, 1998). If attentional cues and stimuli are presented synchronously (Bertelson et al., 2000;

Vroomen et al., 2001), this produces a temporal misalignment between attentional shift and perceptual processing and thus invalidates the conclusion that attention cannot impact audio-visual spatial localisation. Furthermore, several recent studies do support a modulatory influence of endogenous attention on multisensory perceptual decisions (Cao et al., 2019; Donohue et al., 2015; Michail & Keil, 2018; Odegaard et al., 2016; Vercillo & Gori, 2015) and of exogenous attention on target detection (Van der Stoep et al., 2015). Collectively, it appears that attention impacts multisensory perceptual inference via modulation of sensory evidence and selection of internal task-relevant perceptual representations.

On the other hand, it is now becoming clear that attention does not impact the prior tendency to either integrate or segregate multisensory inputs, as evidenced by unaltered prior binding tendency (p_{common} parameter) in the model-based analyses of Chapters 3 and 4. Such conclusion is further strengthened by another recent investigation (Odegaard et al., 2016), which compared modality-specific valid attention (i.e. attend to auditory modality, report location of auditory stimulus) and divided attention (i.e. attend to auditory and visual modalities, report location of auditory or visual stimuli). Although it was conceivable that participants could have treated the auditory and visual inputs as coming from the same source in the case of divided attention, the authors did not find an increase of the common-source prior p_{common} relative to selective valid attention. It has not been empirically tested whether other types of attention (i.e. apart from endogenous modality-specific attention) impact observers' prior binding tendency; nevertheless, here I argue that it is unlikely the case. Since attention constitutes the selection and modulation of sensory or representational information (Chun et al., 2011), it appears that it cannot instantaneously manipulate observer's prior tendency to either integrate or segregate multisensory signals. Instead, such tendency is likely supported by observers' expectations (Chen & Spence, 2017; Spence, 2011), which in turn

depend on prior experience (Ernst, 2007; Gau & Noppeney, 2016; Jicol et al., 2018; Lee & Noppeney, 2011a; Love et al., 2012; Nahorna et al., 2012; Petrini et al., 2010; 2011), ontogenetic (Gopnik & Tenenbaum, 2007) and phylogenetic (Geisler & Diehl, 2002) factors.

Yet, it remains to be understood whether a functional connection between attention and binding tendency exists in the opposite direction, namely whether cross-modal binding influences the deployment of attentional resources. Chapter 5 directly addressed this question. In particular, given that cross-modal binding generates the perceptual experience of a unified multisensory object (Bizley et al., 2016), I evaluated its effect on cross-modal object-based attention (Fiebelkorn et al., 2012) and found that multisensory objects capture attention. Specifically, cross-modal binding enhanced selective tracking of the stream of information containing coherent multisensory features within a cocktail-party scenario. This result demonstrates that not only unisensory (Kimchi et al., 2007; 2016; Yeshurun et al., 2009) but also multisensory objects are salient entities that promote object-based attentional selection and modulation under competition for processing resources, alongside cross-modal spread of object-based attention (Maddox et al., 2015).

Overall, the present thesis provides behavioural evidence of an interactive relationship between attention and multisensory integration depending on which computational task is at hand. When observers need to construct a complex representation (e.g. spatial location) via noisy estimates, attention modulates the reliability of sensory evidence based on attentional focus and determines the selection of internal representations based on task relevance (Chapters 3-4). When observers try to detect targets within a complex perceptual scene that taxes processing resources, cross-modal binding impacts scene analysis via salience-driven attentional capture (Chapter 5). These different computational mechanisms are supported by distinct neural substrates along the sensory cortical hierarchies.

7.1.2 The interplay between attention and multisensory integration across the cortical hierarchy

In Chapter 4 of the present thesis I addressed how endogenous modality-specific attention (again, dissociating pre-stimulus focus and post-stimulus response selection) impacts the formation of neural audio-visual spatial representations along the dorsal sensory cortical hierarchies. In low-level visual areas, pre-stimulus focus biased decoded spatial representations towards attended stimuli, irrespective of response requests. In other words, attentional selection and modulation of sensory signals increased the weight of attended neural representations. As the relative weights for integration directly depend on their respective reliability (i.e. inverse of variance; Alais & Burr, 2004; Ernst & Banks, 2002; Ernst & Bülthoff, 2004), it appears that external attention impacts spatial representations' precision via sharpening of spatial tuning functions (Martinez-Trujillo & Treue, 2004) at the bottom of the sensory hierarchy, possibly via modulation of internal noise (Serences & Kastner, 2014). In higher-order areas of the dorsal visual and auditory hierarchies, post-stimulus response selection biased decoded spatial representations towards task-relevant stimuli, irrespective of pre-stimulus focus. In other words, attentional selection of internal perceptual information increased the weight of task-relevant neural representations, in accordance with the presence of priority maps that encode behavioural relevance (Bisley & Goldberg, 2010; Rohe & Noppeney, 2016; Serences & Yantis, 2006; Sprague et al., 2018). Again, this result expands previous investigations that confounded external and internal attention (Aller & Noppeney, 2019; Cao et al., 2019; Rohe et al., 2019; Rohe & Noppeney, 2015a, 2016, 2018).

As a consequence, it appears that the entire sensory cortical hierarchy is implicated in multisensory perceptual inference and its interplay with attention, but via different operations. Accordingly, it is now well accepted that multisensory interactions are pervasive in the neo-

cortex (Driver & Noesselt, 2008; Foxe & Schroeder, 2005; Ghazanfar & Schroeder, 2006; Kayser & Logothetis, 2007; Schroeder & Foxe, 2005), but they are driven by distinct computational principles (Rohe & Noppeney, 2015a, 2016). With the present work, I expand such claim by showing that so is also the attentional control over multisensory interactions. On the one hand, the bottom of the sensory hierarchy encodes modality-specific representations, which are sensitive to attentional modulation; this in turn impacts reliability-weighted integration (Alais & Burr, 2004; Ernst & Banks, 2002; Ernst & Bülthoff, 2004). On the other hand, higher-order association areas encode task-specific representations, which are sensitive to attentional selection; this in turn determines the formation of perceptual estimates in accordance with Bayesian Causal Inference (Körding et al., 2007; Shams & Beierholm, 2010). Collectively, this body of evidence underscores the need to move beyond identifying multisensory or attention-related regions, as the interplay between multisensory integration and attention is implemented in a widespread network across the cortical hierarchy. Instead, it is crucial to characterize the functional properties and behavioural relevance of such multi-stage interplay under different task contexts and attentional demands.

Accordingly, investigating the neural underpinnings of cross-modal binding and its relationship with attention (Chapter 6) revealed a hierarchy of neural computations. Sensory-specific interactions at the bottom of the cortical hierarchy appeared to promote further analysis in association areas and to trigger the recruitment of posterior parietal areas for control of object-based attention during competition for processing resources. In particular, cross-modal activations were amplified in low-level sensory areas when deriving from correspondent signals. This effect may represent an adaptive mechanism whereby cross-modal binding facilitates perceptual scene analysis (Laurienti et al., 2002; Stanford & Stein, 2007; Stein & Stanford, 2008). In particular, cross-modal binding may influence the creation

of salience maps (Itti & Koch, 2000; Kayser et al., 2005), which direct further domain-specific processing in association areas (Foxe & Schroeder, 2005; Lewis & Noppeney, 2010; Werner & Noppeney, 2010) and recruit executive control areas (Corbetta et al., 2008; Corbetta & Shulman, 2002) during attentional competition (Desimone & Duncan, 1995). Since in Chapter 6 participants were passively experiencing the stimulation, it appears that these effects can arise independently of specific task instructions, in line with complementary evidence in awake and anesthetised ferrets (Atilgan et al., 2018), in anesthetised monkeys (Kayser et al., 2005; 2008) and in humans (Werner & Noppeney, 2010). Instead, a causal role may be played by the characteristics of the stimulation itself: when this carries evidence of congruence (e.g. spatial, temporal), it may automatically trigger a cascade of neural events that impact sensory processing and attentional selection. In agreement with this hypothesis, it has recently been demonstrated that multisensory congruence effects arise independent of the locus of selective modality-specific attention (Misselhorn et al., 2016), again reflecting the automatic analysis of cross-modal correspondent features in the sensory scene. Crucially, the behavioural results of Chapter 5 nicely dovetail with the conclusion of automatic attentional capture by multisensory objects and sustain the functional relevance of the neural effects found in Chapter 6.

7.1.3 Methodological considerations

In the following, I will address a few methodological considerations regarding the present empirical work, which could inform follow-up investigations targeting similar experimental questions.

Chapters 3 and 4 offer clear evidence that cognitive control in the form of endogenous modality-specific attention impacts multisensory perceptual inference; however, multiple

methodological trade-offs had to be addressed in order to optimise the sensitivity of the experiments. Specifically, it was necessary to solve a multiple-constraint problem regarding stimuli arrangement along the azimuth, their spatial reliability and the consequent impact on spatial localisation difficulty. Positioning stimuli at relatively close spatial locations and near 0˚ azimuth reflected spatial constraints of the fMRI experimental equipment. This likely impacted the ability to reliably decode spatial disparity effects in the fMRI study of Chapter 4; moreover, this spatial arrangement amplified the imbalance of spatial localisation difficulty between vision and audition, alongside ongoing scanner noise. On the other hand, the use of high spatial eccentricities could have determined variations of visual reliability across the azimuth (Charbonneau et al., 2013), thus confounding the effect of attention. Overall, it appears highly challenging to optimally account for all these methodological factors. Therefore, future studies may seek to use different types of tasks, for example object categorisation (Cao et al., 2019). A further methodological point deserves consideration. In order to optimise the experimental procedure to fMRI analyses, each trial presented a single pairing of audio-visual stimuli and directly asked for a localisation response. Future psychophysics studies targeting the role of attention in multisensory perceptual inference may employ different procedures that better control for decision strategies (e.g. two-interval forced-choice design, Petrini et al., 2015).

Chapters 5 and 6 sustain the idea that multisensory congruence determines attentional capture. Importantly, to fully demonstrate that congruency itself (instead of mere co-stimulation) is the key factor driving attentional recruitment, it would be necessary to compare the present findings with conditions of multisensory incongruence within and outside attentional competition. A second point of consideration regards the use of music to account for linguistic confounds when studying the interplay between cross-modal binding

and attention during naturalistic listening. Although music stimuli indeed remove semantic confounds that may impact speech intelligibility (Broderick et al., 2019), syntactic confounds and the associated temporal expectations are still present in the case of music (Koelsch et al., 2019; Noppeney & Lee, 2018; Pearce, 2018; Tillmann, 2012). Thus, future studies may seek to employ intermediately artificial stimuli consisting of continuous and dynamic noise envelopes (Atilgan et al., 2018; Maddox et al., 2015). On the other hand, expectations are intertwined with naturalistic perception (Kaiser et al., 2019; Soto-Faraco et al., 2019) and attention (Peelen & Kastner, 2014), thus it becomes arguable to attempt to fully account for their impact without distorting the nature itself of the process under investigation.

7.2 Towards a cohesive model

The empirical work presented in this thesis provides complementary evidence of the mutual multi-stage interplay between attention and multisensory integration. Yet, it is still necessary to integrate the present findings and background literature into a cohesive explanatory model that allows interactions at multiple levels of processing, from the segmentation of the sensory scene for signals detection (Chapters 5 and 6) to the formation of more complex task-specific representations (Chapters 3 and 4). Crucially, the present thesis is in line with the hypothesis of a tight functional connection between multisensory processing, attention and prior knowledge (Talsma, 2015). Accordingly, the framework of Bayesian Causal Inference (Körding et al., 2007; Shams & Beierholm, 2010) offers an appropriate structure for integrating findings into a cohesive model (Figure 7.1), alongside the characterisation of the respective neural implementation (Figure 7.2).

First of all, Chapter 5 offers evidence that cross-modal binding (Figure 7.1a) and attention impact scene analysis for signal detection via salience-driven attentional capture

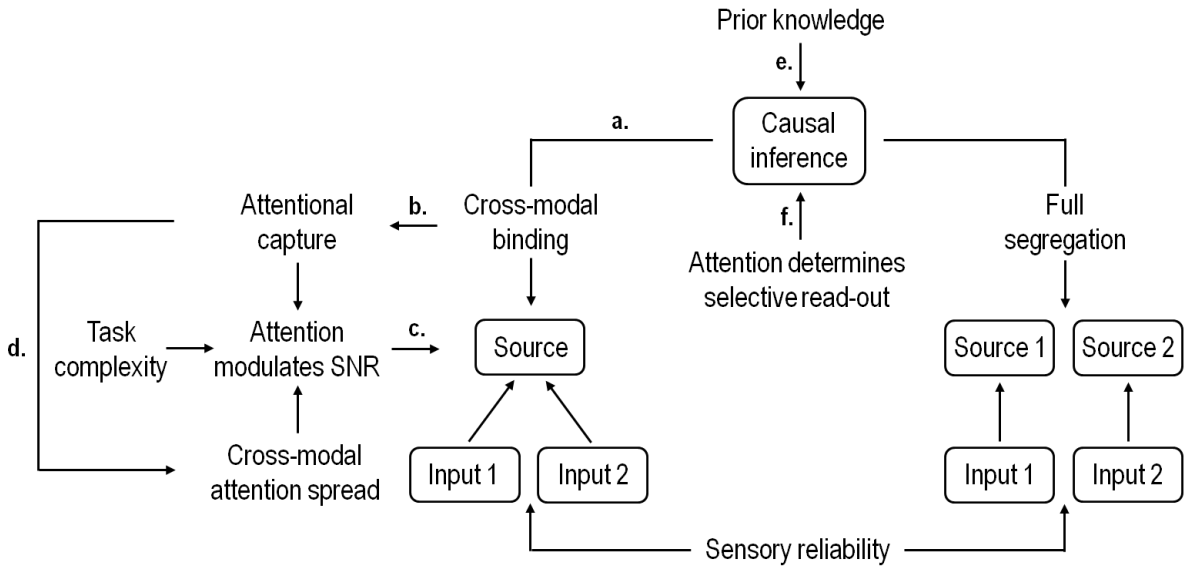

Figure 7.1: Schematic representation of the interplay between attention and multisensory integration
Letters refer to more extensive explanation in the text.

(Figure 7.1b). This is in line with abundant evidence that multisensory correspondences are highly salient and can thus orient attention during competition for processing resources (Ho et al., 2009; Mastroberardino et al., 2015; Matusz & Eimer, 2011; Santangelo et al., 2008; Santangelo & Spence, 2007; Van der Burg et al., 2008; Van der Burg et al., 2011). The SNR of attended representations is consequently enhanced (Figure 7.1c) and facilitates the extraction of information from background clutter (Macaluso et al., 2016). Chapter 6 elucidates the underlying neural mechanisms by showing the emergence of cross-modal binding in low-level sensory areas (Figure 7.2a), which in turn recruits posterior parietal areas for attentional orienting (Figure 7.2b). This, in turn, is known to modulate SNR in sensory cortices (Figure 7.2c) via signal enhancement and external noise suppression (Serences & Kastner, 2014). Importantly, facilitation effects due to multisensory congruence are larger when attention is divided between sensory modalities compared to when it is focused on one modality (Göschl et al., 2014; Mozolic et al., 2008a; Talsma et al., 2007). Thus, while

modality-specific stimulation (Chapter 6; Laurienti et al., 2002) and attention (Ciaramitaro et al., 2007; Johnson & Zatorre, 2005, 2006; Mozolic et al., 2008b) deactivate cross-modal sensory areas to boost modality-specific processing, multisensory stimuli (Chapter 6; Laurienti et al., 2002) and cross-modal divided attention (Johnson & Zatorre, 2006) counteract this effect and co-determine the segmentation of the multisensory scene. Notably, such relationship is governed by the inverse effectiveness principle (Van der Stoep et al., 2015), which underscores the flexible cooperation of cross-modal binding and attention in boosting signal-to-noise ratio. Furthermore, cross-modal spread of attention (Busse et al., 2005; Molholm et al., 2007) represents a complementary mechanism impacting perceptual scene analysis (Figure 7.1d). Importantly, such mechanism arises once attention is already put in place either endogenously (Maddox et al., 2015; van Ee et al., 2009) or potentially exogenously (as in Chapters 5-6 of the present thesis; see Tang et al., 2016 for discussion) and thus implies feedback signals from fronto-parietal areas to low-level sensory areas (Figure 7.2d; Fiebelkorn, 2012; van Ee et al., 2009; Zimmer et al., 2010a). Crucially, Bayesian Causal Inference postulates that prior knowledge modulates the strength of cross-modal binding (Figure 7.1e; Chen & Vroomen, 2013; Chen & Spence, 2017; Parise, 2015; Spence, 2011). Following predictive coding (Friston, 2010), the influence of prior knowledge may be implemented via signals from higher-order association areas (in particular, medial prefrontal cortex, Summerfield et al., 2006) to low-level sensory areas (Figure 7.2e). Accordingly, implicit associations due to prior exposure to cross-modal stimuli co-activate the correspondent early sensory cortices (Zangenehpour & Zatorre, 2010). Prior knowledge may then indirectly influence the synergistic interplay of cross-modal binding and attention. In line with this conjecture, the strength of perceptual grouping modulates both the strength of unisensory object-based attentional capture (Kimchi et al., 2016) and the strength of cross-

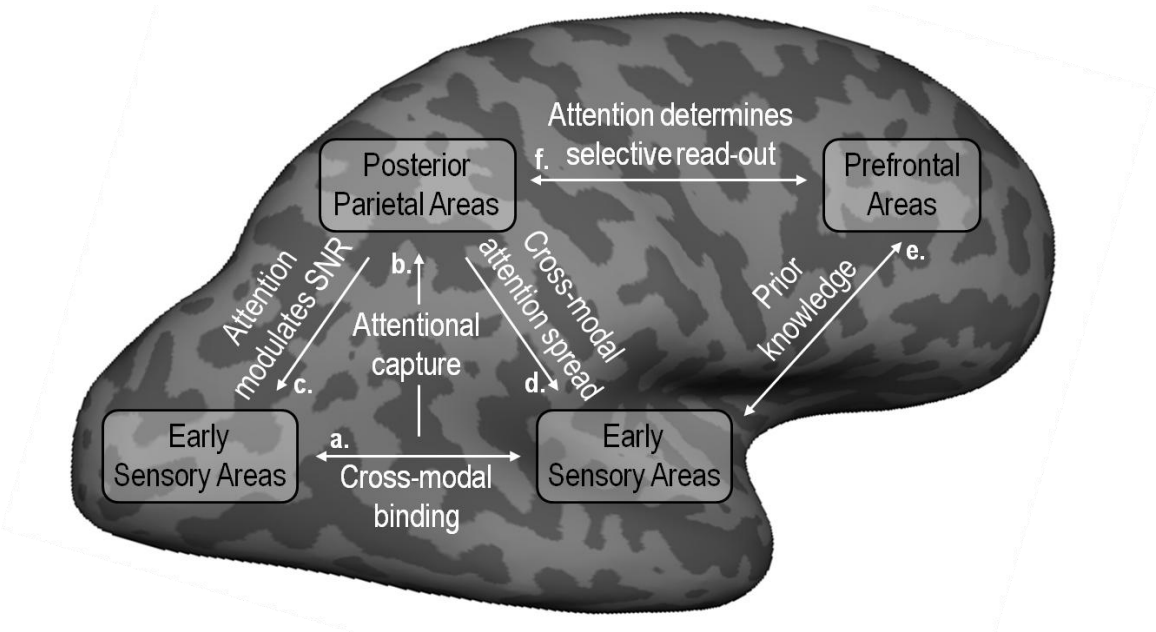

Figure 7.2: Neural implementation of the interplay between attention and multisensory integration
Letters refer to more extensive explanation in the text.

modal object-based spread of attention (Donohue et al., 2011; Fiebelkorn et al., 2010; Zimmer et al., 2010b).

A second level of computation whereby selective attention impacts multisensory integration concerns the modulation of sensory reliabilities (Figure 7.1c; Chapters 3 and 4 of the present thesis; Odegaard et al., 2016; Vercillo & Gori, 2015). This arises from internal noise reduction in low-level sensory areas (Serences & Kastner, 2014), again reflecting a top-down effect from fronto-parietal areas (Figure 7.2c; Corbetta et al., 2008; Corbetta & Shulman, 2002; Santangelo et al., 2010; Shomstein & Behrmann, 2006; Shomstein & Yantis, 2004). Notably, the relationship between attention and multisensory integration adheres to an inverse effectiveness principle also at this level of processing: the lower stimulus physical reliability, the higher the impact of attention on the final percept (Oruc et al., 2008). This underscores the flexible cooperation of attention and sensory reliability (Rohe & Noppeney,

2015b) in forming complex representations (e.g. spatial location; Macaluso et al., 2016) via reliability-weighed integration (Alais & Burr, 2004; Ernst & Banks, 2002; Ernst & Bülthoff, 2004). Sensory reliability also directly impacts the level of complexity or difficulty of the task at hand: the lower stimulus reliability, the higher task difficulty (e.g. auditory localisation is more difficult that visual localisation because auditory stimuli are far less reliable than visual stimuli in the spatial domain, Freides, 1974). Overall, the lower stimulus physical reliability, the higher task difficulty, the higher the impact of attention on the SNR (Oruc et al., 2008).

Interestingly, the critical role of attention in determining the precision of sensory representations is in line with the definition of attention itself within the framework of predictive coding (Feldman & Friston, 2010). According to this account, attention increases the precision of error signals that derive from the mismatch between prior knowledge and current sensory evidence. In other words, attention boosts the reliability (and therefore the weight) of error signals and consequently determines how strongly sensory evidence influences the revision of current expectations. In the context of multisensory perceptual inference, attention may therefore impact the development and revision of common source priors based on experience (Chen & Spence, 2017; Spence, 2011). Furthermore, attention determines the flexible selection of internal representations for production of a final response in accordance with task relevance (Figure 7.1f). Such selective read-out mechanism (Serences & Kastner, 2014) arises at the top of the sensory processing hierarchy (Figure 7.2f; Chapter 4), in line with the idea that it represents a late-stage decisional mechanism independent from sensory gain modulations in low-level sensory areas (Pestilli et al., 2011). However, it remains to be understood how observers behave when not engaged in a task with specific response requirements. In naturalistic contexts, it is conceivable that prior knowledge instructs selective read-out (Figure 7.2f). In particular, the learnt efficacy of the different sensory

modalities in relation to specific tasks (modality-appropriateness hypothesis, Welch & Warren, 1980; see also Chen & Vroomen, 2013) may determine biased representations towards a specific sensory modality based on the task at hand.

Collectively, the cohesive integration of the present findings with previous literature (Figure 7.1 and Figure 7.2) fosters the thesis that attention and multisensory integration synergistically interact at multiple levels of processing to serve a common computational goal: to promote scene analysis flexibly adjusting for environmental conditions (competition for processing resources, sensory noise) and task demands (detection, discrimination) and to ultimately guide adaptive behaviour in our complex world. Underscoring the flexible cooperation of attention and multisensory integration, given environmental conditions and task demands, reconciles the artificial dichotomy between early and late integration frameworks (Calvert & Thesen, 2004; Noppeney et al., 2018; see Section 1.3). Multisensory integration determines attentional capture in the presence of competing streams of information; attention modulates sensory uncertainty and determines selective read-out of internal task-relevant representations. Prior knowledge instructs such mutual interplay within a Bayesian framework (Talsma, 2015), which also allows iterative revision of prior knowledge itself in accordance with predictive coding (Friston, 2010). Hence, future work should embrace a parallel integration framework (Calvert & Thesen, 2004; Noppeney et al., 2018) with the aim to better characterise the computational mechanisms that allow detection of information and construction of complex representations in our complex multisensory world.

7.3 Future directions

Given the emerging evidence of a tight functional connection between multisensory integration, attention and predictive processes, a promising area of future investigations concerns the characterisation of their interplay in real-life scenarios. In fact, naturalistic scenes are intrinsically multisensory (Soto-Faraco et al., 2019), they tax our attentional resources with a constant influx of sensory inputs (Peelen & Kastner, 2014) and they are loaded with statistical regularities (Kaiser et al., 2019).

Multisensory coherence may play a critical role in promoting attentional selection of meaningful objects in cluttered environments (e.g. Chapter 5; Maddox et al., 2015). On the other hand, high perceptual load (Lavie, 2005) may weaken the ability of multisensory congruence to orient attention (Alsius & Soto-Faraco, 2011), unless highly salient (e.g. abrupt, loud) cues are able to win the competition for processing resources (Desimone & Duncan, 1995; Talsma et al., 2010). Hence, future investigations should systematically address the extent to which the perceptual complexity of real-life environments impacts multisensory-mediated attentional orienting. Importantly, this line of research has implications for the design of warning signals during the execution of demanding tasks (e.g. driving in traffic, Ho et al., 2005).

Crucially, naturalistic scenes are also highly structured and thus predictable (Kaiser et al., 2019). Predictive mechanisms exploiting everyday statistical regularities trigger the use of search templates to optimise attentional allocation (Peelen & Kastner, 2014; Torralba et al., 2006; Wolfe et al., 2011). Moreover, structured objects dominate single features for attentional selection during naturalistic scenes analysis (Stoll et al., 2015). Hence, multisensory objects (guided by the related prior knowledge, Parise, 2015) may represent

search templates (Mast et al., 2015; Matusz & Eimer, 2013) for efficient attentional orienting during naturalistic perception. Future research could explore the impact of various scales of multisensory statistical regularities (temporal, spatial, semantic etc.) on naturalistic visual search and listening and seek to unveil the underlying neural architecture. Since semantic correspondences appear to take over temporal correspondences during emotion perception with music stimuli (Petrini et al., 2010), it would also be important to put various scales of multisensory statistical regularities into conflict and quantify which ones are given more weight and therefore bias behaviour to a greater extent.

Another key feature of real-world scenes is that multiple objects are always present at the same time; hence Bayesian Causal Inference becomes critical for arbitration between integration and segregation of multisensory inputs in order to construct a coherent perceptual scene. To this end, it is crucial not only to understand which sensory inputs belong to the same object but also what is the relation between different objects across sensory modalities (e.g. relative position and size, direction and speed of motion). Hence, a future challenge will be to characterise Bayesian Causal Inference (and its relationship with attention) when multiple objects are at stake (Deroy & Spence, 2016).

Furthermore, everyday scenarios are dynamic and ever-changing. Consequently, it is critical to flexibly update the interpretation of the causal structure underlying multisensory stimulation. By boosting the weight of error signals (Feldman & Friston, 2010), attention may highlight the need for revision of the current perceptual interpretation. Future research could explore how much discrepancy (i.e. how much attentional focus) is needed to trigger the revision process and which neural networks are responsible for such effect.

Finally, it will be fundamental to establish not only how multisensory causal inference flexibly adjusts for changing environmental conditions in adulthood, but also how such

remarkable ability evolves across the lifespan. Emerging evidence shows unaltered capacities in healthy ageing, despite decreased speed of processing (Jones et al., 2019); on the other hand, the development of multisensory causal inference still remains unknown (Petrini et al., 2015). Future studies should target such unexplored issue and characterise the interactions between the ontogenesis of multisensory causal inference, attention and predictive processes. Crucially, different developmental trajectories may instruct on the causal relationships between these processes, which are deeply intertwined in adulthood. In this respect, it is worth noticing that intersensory redundancies generated by the same object support the development of selective attention in human infants (Bahrick et al., 2004; Bahrick & Lickliter, 2000).

Collectively, the emerging evidence of a tight functional interconnection between multisensory integration, attention and predictive processes provides a promising framework wherein to characterise the development and flexible adjustment of adaptive behaviour in our complex and dynamic world.

CHAPTER 8

SUPPLEMENTARY MATERIALS

8.1 Chapter 3

8.1.1 Response times

Analysis of response times with a focus on invalidity effects (i.e. Attention × Report interaction) was performed at the individual level as inclusion criterion in the study (see Section 3.2.5). Group-level analysis was performed for included participants as follow-up investigation. For every participant, median response times of each experimental condition were averaged across all combinations of AV locations at a particular level of AV spatial disparity and entered into a 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 4 (AV spatial disparity: 0°, 6°, 12° or 18° visual angle, i.e. zero, low, mid or high disparity) repeated measures ANOVA.

Results are shown in Figure 8.1A and summarised in Table 8.1. As a consequence of the inclusion criterion applied at the individual level (see Section 3.2.5), we found a significant Attention × Report interaction ($F_{1,29} = 150.330$, $p < 0.001$, $\eta p^2 = 0.838$): response times were faster for valid versus invalid trials. Post-hoc t-tests (Bonferroni-corrected $\alpha = 0.025$) confirmed that the validity effect was significant both under auditory reports (valid vs

invalid: t_{29} = -13.298, $p < 0.001$, Cohen's d_{AV} = 1.168, Wilcoxon signed-ranks z = -4.782, $p < 0.001$, r = 0.617) and visual reports (valid vs invalid: t_{29} = -11.135, $p < 0.001$, Cohen's d_{AV} = 1.632, Wilcoxon signed-ranks z = -4.782, $p < 0.001$, r = 0.617). In addition, we found a significant main effect of AV spatial disparity ($F_{1,29}$ = 13.604, $p < 0.001$, ηp^2 = 0.319). Post-hoc t-tests (Bonferroni-corrected α = 0.008) revealed that response times decreased for spatially congruent trials relative to any other AV disparity levels (zero vs low: t_{29} = -6.156, $p < 0.001$, Cohen's d_{AV} = 0.130, Wilcoxon signed-ranks z = -4.227, $p < 0.001$, r = 0.546; zero vs mid: t_{29} = -7.123, $p < 0.001$, Cohen's d_{AV} = 0.180, Wilcoxon signed-ranks z = -4.515, $p < 0.001$, r = 0.583; zero vs high: t_{29} = -4.358, $p < 0.001$, Cohen's d_{AV} = 0.160, Wilcoxon signed-ranks z = -3.671, $p < 0.001$, r = 0.474). As recently suggested (Jones et al., 2019), AV spatially congruent relative to incongruent trials involve less ambiguity in resolving signals' causal structure; thus, they enable faster computation of the final spatial estimate for localisation, which is reflected in a decrease of response times. In summary, faster response times for valid versus invalid trials show participants' appropriate attentional focus based on pre- and post-cues; faster response times for spatially congruent versus incongruent trials indicate participants' active processing of spatial information to solve spatial localisation.

8.1.2 Response errors

As follow-up inspection, we also evaluated the effect of modality-specific attention, modality-specific report and AV spatial disparity on response errors by entering each participant's mean proportion of missed and wrong responses (i.e. no answer and use of wrong keypad respectively) separately into a 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 4 (AV spatial disparity: 0°, 6°, 12° or 18° visual angle, i.e. zero, low, mid or high disparity) repeated measures ANOVA.

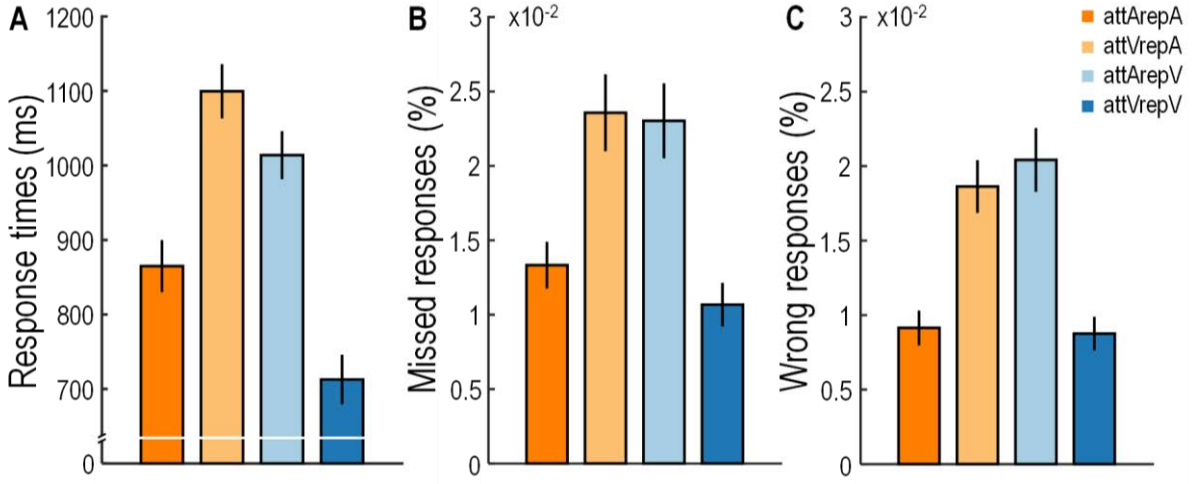

Figure 8.1: Response times and errors

Group mean (±SEM) of A) response times in milliseconds, B) missed responses (i.e. no key press within response time window) and C) wrong responses (i.e. use of wrong keypad) as a function of Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

Results are shown in Figure 8.1B-C and summarised in Table 8.2 as a function of Attention and Report (we pulled over AV spatial disparity as it did not show any significant effects). We found a significant Attention × Report interaction for missed responses ($F_{1,29}$ = 18.231, $p < 0.001$, $\eta p^2 = 0.386$), which decreased for valid versus invalid trials. Post-hoc t-tests (Bonferroni-corrected $\alpha = 0.025$) confirmed that the validity effect was significant both under auditory reports (valid vs invalid: $t_{29} = -3.284$, $p = 0.003$, Cohen's $d_{AV} = 0.520$, Wilcoxon signed-ranks $z = -2.870$, $p = 0.004$, $r = 0.370$) and visual reports (valid vs invalid: $t_{29} = -3.834$, $p = 0.001$, Cohen's $d_{AV} = 0.656$, Wilcoxon signed-ranks $z = -3.652$, $p < 0.001$, $r = 0.471$). Similarly, the 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 4 (AV spatial disparity: zero/low/mid/high) repeated measures ANOVA with wrong responses as dependent variable showed a significant Attention × Report interaction ($F_{1,29}$ = 26.266, $p < 0.001$, $\eta p^2 = 0.475$): response errors in terms of wrong responses decreased for valid versus

invalid trials. Post-hoc t-tests (Bonferroni-corrected $\alpha = 0.025$) confirmed that the validity effect was significant both under auditory reports (valid vs invalid: $t_{29} = -4.289$, $p < 0.001$, Cohen's $d_{AV} = 0.657$, Wilcoxon signed-ranks $z = -3.741$, $p < 0.001$, $r = 0.483$) and visual reports (valid vs invalid: $t_{29} = -4.016$, $p < 0.001$, Cohen's $d_{AV} = 0.703$, Wilcoxon signed-ranks $z = -4.120$, $p < 0.001$, $r = 0.532$). As a whole, response errors decreased when locating attended relative to unattended stimuli. Thus, in accordance with response times, this result provides evidence that participants appropriately focused their attention based on pre- and post-cues in the ventriloquist paradigm.

RT (ms) mean (±SEM)	**attArepA**	**attVrepA**	**attArepV**	**attVrepV**
Zero disparity	849.892 (±33.548)	1092.926 (±37.527)	986.311 (±32.220)	694.506 (±32.936)
Low disparity	869.106 (±35.962)	1102.698 (±37.556)	1028.325 (±32.769)	715.953 (±33.379)
Mid disparity	886.993 (±36.065)	1113.527 (±36.534)	1028.876 (±31.437)	720.684 (±33.245)
High disparity	881.065 (±35.921)	1096.842 (±36.781)	1014.964 (±35.748)	745.731 (±35.555)

Table 8.1: Response times (RT)

Group mean (±SEM) as a function of AV spatial disparity (Zero/Low/Mid/High: 0°/6°/12°/18° visual angle), Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

Proportion mean (±SEM)	**attArepA**	**attVrepA**	**attArepV**	**attVrepV**
Missed responses	0.013 (±0.003)	0.024 (±0.005)	0.023 (±0.005)	0.011 (±0.002)
Wrong responses	0.009 (±0.002)	0.019 (±0.003)	0.020 (±0.004)	0.009 (±0.002)

Table 8.2: Response errors

Group mean (±SEM) proportion of missed and wrong responses as a function of Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

8.2 Chapter 4

8.2.1 Response times

Analysis of response times with a focus on validity effects (i.e. Attention × Report interaction) was performed at the individual level as inclusion criterion in the study (see Section 4.2.6). Group-level analysis was performed for included participants as follow-up investigation. For each participant, median response times of each experimental condition were averaged across all combinations of AV locations at a particular level of AV spatial disparity and entered into a 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 3 (AV spatial disparity: 0°, 9° or 18° visual angle, i.e. zero, low or high disparity) repeated measures ANOVA.

Results of psychophysics and fMRI experiments are shown in Figure 8.2 and summarised in Table 8.3. For both experiments, we found a significant Attention × Report interaction (psychophysics: $F_{1,26} = 247.330$, $p < 0.001$, $\eta p^2 = 0.905$; fMRI: $F_{1,11} = 128.590$, $p < 0.001$, $\eta p^2 = 0.921$): response times decreased for valid versus invalid trials. Post-hoc t-tests (Bonferroni-corrected $\alpha = 0.025$) confirmed that the validity effect was significant both under auditory reports (psychophysics: $t_{26} = -15.295$, $p < 0.001$, Cohen's $d_{AV} = 1.645$, Wilcoxon signed-ranks $z = -4.541$, $p < 0.001$, $r = 0.618$; fMRI: $t_{11} = -10.087$, $p < 0.001$, Cohen's $d_{AV} = 0.967$, Wilcoxon signed-ranks $z = -3.059$, $p < 0.001$, $r = 0.624$) and visual reports (psychophysics: $t_{26} = -11.746$, $p < 0.001$, Cohen's $d_{AV} = 2.030$, Wilcoxon signed-ranks $z = -4.541$, $p < 0.001$, $r = 0.618$; fMRI: $t_{11} = -9.198$, $p < 0.001$, Cohen's $d_{AV} = 1.392$, Wilcoxon signed-ranks $z = -3.059$, $p < 0.001$, $r = 0.624$). In addition, we found a significant main effect of AV spatial disparity (psychophysics: $F_{1,26} = 67.195$, $p < 0.001$, $\eta p^2 = 0.721$; fMRI: $F_{1,11} = 59.209$, $p < 0.001$, $\eta p^2 = 0.843$). Post-hoc t-tests (Bonferroni-corrected $\alpha = 0.017$) revealed

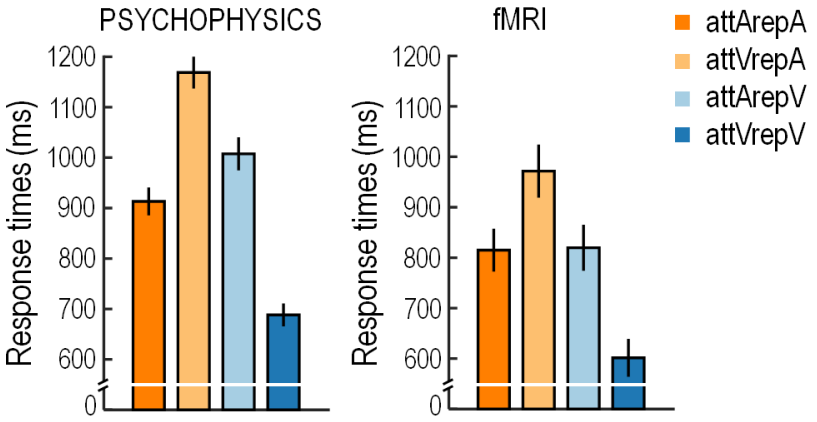

Figure 8.2: Response times in the psychophysics and fMRI experiments

Group mean (±SEM) response times in milliseconds as a function of Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

that in the psychophysics experiment response times decreased for lower than higher AV disparities (zero vs low: $t_{26} = -6.306$, $p < 0.001$, Cohen's $d_{AV} = 0.388$, Wilcoxon signed-ranks $z = -4.252$, $p < 0.001$, $r = 0.579$; zero vs high: $t_{26} = -11.505$, $p < 0.001$, Cohen's $d_{AV} = 0.772$, Wilcoxon signed-ranks $z = -4.541$, $p < 0.001$, $r = 0.618$; low vs high: $t_{26} = -5.474$, $p < 0.001$, Cohen's $d_{AV} = 0.402$, Wilcoxon signed-ranks $z = -4.228$, $p < 0.001$, $r = 0.575$). Similarly, in the fMRI experiment response times decreased for spatially congruent trials (i.e. AV disparity equal to zero) relative to higher AV disparities (zero vs low: $t_{11} = -6.969$, $p < 0.001$, Cohen's $d_{AV} = 0.414$, Wilcoxon signed-ranks $z = -3.059$, $p = 0.002$, $r = 0.624$; zero vs high: $t_{11} = -14.177$, $p < 0.001$, Cohen's $d_{AV} = 0.578$, Wilcoxon signed-ranks $z = -3.059$, $p = 0.002$, $r = 0.624$). As recently suggested (Jones et al., 2019), AV spatially congruent relative to incongruent trials involve easier determination of signals' causal structure; thus, they enable faster computation of the final localisation response, which is reflected in faster response times. Collectively, response time benefits due to attention validity and AV spatial congruence corroborate participants' appropriate task engagement.

RT (ms) mean (±SEM)	attArepA	attVrepA	attArepV	attVrepV
Psychophysics				
Zero disparity	813.576 (±29.475)	1090.314 (±31.575)	924.470 (±29.713.)	672.741 (±25.696)
Low disparity	886.226 (±29.255)	1158.215 (±31.743)	989.133 (±30.273)	673.381 (±23.498)
High disparity	980.776 (±32.297)	1196.172 (±33.812)	1034.613 (±36.368)	714.720 (±23.017)
fMRI				
Zero disparity	699.306 (±43.987)	886.360 (±43.670)	774.940 (±44.098)	574.800 (±36.620)
Low disparity	810.199 (±47.281)	973.045 (±52.790)	812.414 (±48.054)	589.126 (±36.040)
High disparity	821.971 (±36.734)	972.205 (±48.258)	833.904 (±43.746)	636.261 (±41.780)

Table 8.3: Response times (RT) in the psychophysics and fMRI experiments

Group mean (±SEM) as a function of AV spatial disparity (Zero/Low/High: 0°/9°/18° visual angle), Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

8.2.2 Response errors

As follow-up inspection we also evaluated the effect of modality-specific attention, modality-specific report and AV spatial disparity on response errors by entering each participant's mean proportion of missed and wrong responses (i.e. no answer and use of wrong keypad respectively) separately into a 2 (Attention: Auditory/Visual) × 2 (Report: Auditory/Visual) × 3 (AV spatial disparity: 0°, 9° or 18° visual angle, i.e. zero, low or high disparity) repeated measures ANOVA.

Results of psychophysics and fMRI experiments are shown in Figure 8.3 and summarised in Table 8.4 as a function of Attention and Report (we pulled over AV spatial disparity as it did not show any significant effects). For both experiments, we found a significant Attention × Report interaction for missed responses (psychophysics: $F_{1,26} = 9.054$, $p = 0.006$, $\eta p^2 = 0.258$; fMRI: $F_{1,11} = 9.243$, $p = 0.011$, $\eta p^2 = 0.457$), indicating that they decreased for valid versus invalid trials. Post-hoc t-tests (Bonferroni-corrected $\alpha = 0.025$) confirmed that the validity effect was significant under auditory reports (psychophysics: $t_{26} = -3.312$, $p = 0.003$, Cohen's $d_{AV} = 0.560$, Wilcoxon signed-ranks $z = -3.549$, $p < 0.001$, $r = 0.483$; fMRI: $t_{11} = -3.971$, $p = 0.002$, Cohen's $d_{AV} = 0.244$, Wilcoxon signed-ranks $z = -2.669$, $p = 0.008$, $r = 0.551$) and under visual reports for psychophysics ($t_{26} = -2.561$, $p = 0.017$, Cohen's $d_{AV} = 0.433$, Wilcoxon signed-ranks $z = -3.038$, $p = 0.002$, $r = 0.413$). Instead, only a small trend was present for fMRI ($t_{11} = -1.886$, $p = 0.086$, Cohen's $d_{AV} = 0.201$, Wilcoxon signed-ranks $z = -1.601$, $p = 0.109$, $r = 0.327$).

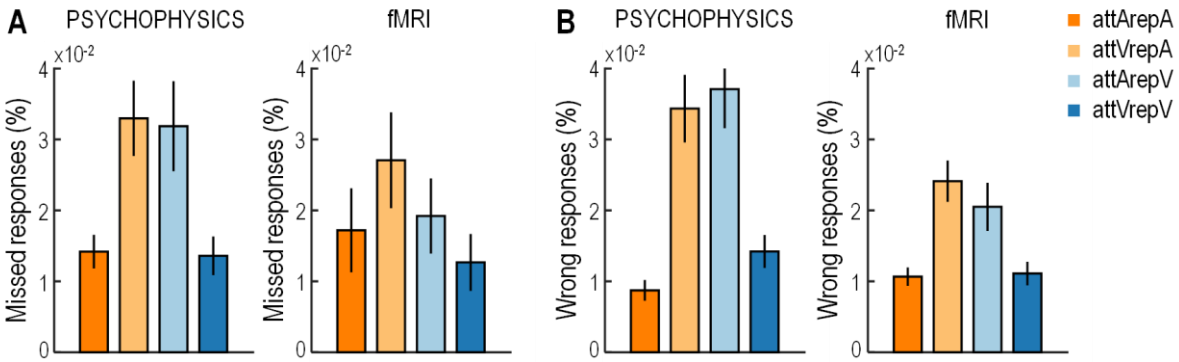

Figure 8.3: Response errors in the psychophysics and fMRI experiments

Group mean percentage (±SEM) of A) missed responses (i.e. no key press within response time window) and B) wrong responses (i.e. use of wrong keypad) as a function of Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

Similarly to missed responses, the repeated measures ANOVA with wrong responses as dependent variable showed a significant Attention × Report interaction (psychophysics: $F_{1,26} = 13.096$, $p = 0.001$, $\eta p^2 = 0.335$; fMRI: $F_{1,11} = 12.061$, $p = 0.005$, $\eta p^2 = 0.523$), indicating that wrong responses decreased for valid versus invalid trials. Post-hoc t-tests (Bonferroni-corrected $\alpha = 0.025$) confirmed that the validity effect was significant both under auditory reports (psychophysics: $t_{26} = -3.473$, $p = 0.002$, Cohen's $d_{AV} = 0.866$, Wilcoxon signed-ranks $z = -3.930$, $p < 0.001$, $r = 0.535$; fMRI: $t_{11} = -3.855$, $p = 0.003$, Cohen's $d_{AV} = 1.001$, Wilcoxon signed-ranks $z = -2.936$, $p = 0.003$, $r = 0.599$) and visual reports (psychophysics: $t_{26} = -3.199$, $p = 0.004$, Cohen's $d_{AV} = 0.653$, Wilcoxon signed-ranks $z = -3.271$, $p = 0.001$, $r = 0.445$; fMRI: $t_{11} = 2.590$, $p = 0.025$, Cohen's $d_{AV} = 0.654$, Wilcoxon signed-ranks $z = -2.825$, $p = 0.005$, $r = 0.577$). Collectively, and alongside response times results, decreases of response errors due to attention validity corroborate participants' appropriate focus based on attentional cues in the ventriloquist paradigm.

Proportion mean (±SEM)	attArepA	attVrepA	attArepV	attVrepV
Psychophysics				
Missed responses	0.014 (±0.003)	0.033 (±0.008)	0.032 (±0.010)	0.014 (±0.004)
Wrong responses	0.009 (±0.001)	0.034 (±0.008)	0.037 (±0.009)	0.014 (±0.003)
fMRI				
Missed responses	0.017 (±0.010)	0.027 (±0.012)	0.019 (±0.009)	0.013 (±0.007)
Wrong responses	0.011 (±0.002)	0.024 (±0.005)	0.021 (±0.005)	0.011 (±0.002)

Table 8.4: Response errors in the psychophysics and fMRI experiments

Group mean (±SEM) proportion of missed and wrong responses as a function of Attention (attA: auditory; attV: visual) and Report (repA: auditory; repV: visual).

8.2.3 Model-based analysis and results: Bayesian Causal Inference

To unveil the computational principles underlying behaviour, we fitted three computational models to participants' localisation responses: (i) Full Segregation model; (ii) Forced Fusion model; (iii) Bayesian Causal Inference model (for details of each model and fitting procedure, please refer to Section 2.2.1; for details of the generative model, see Körding et al., 2007). We first checked whether the BCI model outperformed the two alternative models in predicting participants' behaviour. In this way, we verified the presence of response selection based on task-relevance, as BCI is the only model that explicitly accounts for it. We compared the three models using the Bayesian Information Criterion (BIC) as an approximation to model evidence (Raftery, 1995). For analysis at the group level, we applied both a fixed-effects approach (i.e. sum of individual BICs across subjects) and a random-effects approach (i.e.

Bayesian model selection via SPM's *spm_BMS* function). Finally, we evaluated the effect of modality-specific attention on the sensory variance parameters of the winning model. After rejection of normality (Kolmogorov-Smirnov Test), non-parametric two-tailed Wilcoxon signed-ranks tests assessed pair-wise changes of auditory standard deviation σ_A and visual standard deviation σ_V under auditory versus visual attention. We accounted for multiple comparisons via Bonferroni correction (α= 0.025).

The fitted BCI model comprised the following parameters: common-source prior p_{common} (binding tendency); spatial prior standard deviation σ_P (spread of the central bias); auditory standard deviation σ_A under auditory and visual attention; visual standard deviation σ_V under auditory and visual attention. The fitted Full Segregation and Forced Fusion models did not comprise the common-source prior (which were set to 0 and 1 respectively). Group summary statistics for each model and parameter (mean ± SEM) are reported in Table 8.5 for the psychophysics and fMRI experiments.

First of all, Bayesian model comparison corroborated previous results (Körding et al., 2007) by revealing that the BCI model outperformed the Forced Fusion and Full Segregation models in predicting participants' localisation responses in both the psychophysics and fMRI experiments. This was verified via fixed-effects analysis (highest sum of individual BICs across subjects) and random-effects analysis (highest protected exceedance probability, i.e. probability that a model is more likely than any other model, beyond differences due to chance). This result confirmed the influence of task-relevance on response selection. Consequently, we evaluated the effect of modality-specific attention on the sensory variance parameters of the BCI model. The two-tailed Wilcoxon signed-ranks test contrasting auditory versus visual attention revealed that σ_A significantly decreased under auditory relative to visual attention (psychophysics: σ_A: z = -2.931, p = 0.003, r = 0.399; fMRI: σ_A: z = -2.510, p =

0.012, r = 0.512) and σ_V significantly decreased under visual relative to auditory attention (psychophysics: σ_V: z = -2.138, p = 0.032, r = 0.291; fMRI: σ_V: z = -2.824, p = 0.005, r = 0.576). In other words, sensory reliability increased for signals in the attended versus unattended sensory modality. In a follow-up investigation, we also verified the absence of attention-dependent changes pertaining to common-source prior p_{common} and spatial prior standard deviation σ_P, while the remaining results were virtually the same (and thus are not reported).

Model	**p_{common}**	**σ_P**	**σ_A (attA)**	**σ_A (attV)**	**σ_V (attA)**	**σ_V (attV)**	**relBIC**	**pxp**
Psychophysics								
Bayesian Causal Inference	0.616 (±0.034)	22.087 (±1.846)	6.889 (±0.489)	8.703 (±0.723)	1.942 (±0.111)	1.571 (±0.125)	0	0.999
Forced Fusion	n/a	28.325 (±0.835)	11.442 (±0.918)	12.529 (±1.244)	6.368 (±0.209)	5.676 (±0.267)	5228.351	1.775 x10^{-4}
Full segregation	n/a	19.012 (±1.398)	11.697 (±1.510)	13.552 (±1.479)	2.128 (±0.138)	1.465 (±0.153)	1827.193	1.775 x10^{-4}
fMRI								
Bayesian Causal Inference	0.535 (±0.051)	19.146 (±2.639)	6.091 (±0.554)	7.433 (±0.733)	2.153 (±0.176)	1.694 (±0.282)	0	0.999
Forced Fusion	n/a	29.716 (±0.139)	9.326 (±0.357)	11.390 (±1.736)	6.765 (±0.389)	5.914 (±0.419)	1238.033	1.125 x10^{-4}
Full segregation	n/a	18.150 (±1.856)	8.027 (±0.884)	10.844 (±2.025)	2.200 (±0.179)	1.843 (±0.241)	3047.853	1.355 x10^{-4}

Table 8.5: Model-based results in the psychophysics and fMRI experiments

p_{common}, prior common-source probability; σ_P, spatial prior standard deviation (° visual angle); σ_A, auditory likelihood standard deviation (° visual angle); σ_V, visual likelihood standard deviation (° visual angle); relBIC, Bayesian information criterion of a model summed over subjects (BIC = LL − 0.5 × P × ln(N), LL = log-likelihood, P = number of parameters, N = number of data points) relative to the BCI ("model averaging") model (a model with smaller relBIC provides better data explanation); pxp, protected exceedance probability (probability that a model is more likely than the other models, beyond differences due to chance). attA: auditory attention; attV: visual attention.

8.2.4 Attention invalidity separately for A and V report

Brain regions	MNI coordinates (mm)			Cluster size (voxels)	z-score (peak)	$p_{FWE\text{-}corrected}$ (peak)
	x	y	z			
Attention invalidity (A report)						
L superior frontal gyrus	-22	-4	58	722	> 8	0.000
R superior frontal gyrus	24	-4	48	173	5.72	0.000
L superior frontal gyrus	-4	12	48	423	7.65	0.000
L superior parietal lobule	-20	-66	46	573	6.26	0.000
L precuneus	-6	-64	50		5.75	0.000
L intraparietal sulcus	-28	-54	46		5.73	0.000
R superior parietal lobule	24	-64	44	3	4.77	0.031
L middle frontal gyrus	-28	50	10	103	5.66	0.000
L supramarginal gyrus	-42	-36	46	56	5.48	0.001
R fusiform gyrus	32	-52	-20	27	5.42	0.001
L fusiform gyrus	-32	-50	-20	12	4.91	0.017
Attention invalidity (V report)						
L superior frontal gyrus	-4	8	52	1647	> 8	0.000
L superior frontal gyrus	-28	-6	58		> 8	0.000
R superior frontal gyrus	30	-6	52	495	6.32	0.000
L intraparietal sulcus	-32	-50	46	884	6.85	0.000
L supramarginal gyrus	-48	-36	48		5.78	0.000
L precuneus	-6	-58	46	268	6.67	0.000
R precuneus	6	-60	48		5.82	0.000
L middle frontal gyrus	-28	46	14	40	5.41	0.002
L inferior frontal gyrus (pars opercularis)	-46	2	32	705	7.29	0.000

Table 8.6: fMRI results: Attention invalidity separately for A and V report

Effect of attention invalidity separately for A report (attVrepA > attArepA) and V report (attArepV > attVrepV). p-values are FWE-corrected at the peak level for multiple comparisons within the entire brain. Source of anatomical labels: Duvernoy (1999). attA: auditory attention; attV: visual attention; repA: auditory report; repV: visual report; L: left; R: right.

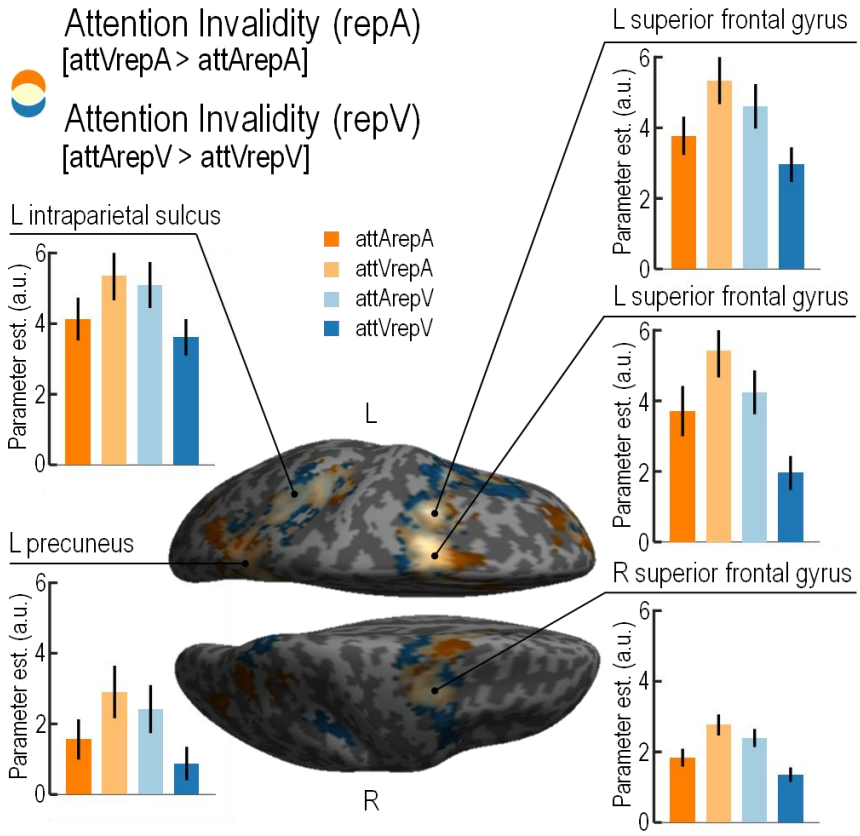

Figure 8.4: fMRI results: Attention invalidity separately for A and V report

Increases of BOLD response associated with attention invalidity separately for auditory report (orange) and visual report (blue). Activation increases are rendered on an inflated canonical brain ($p < 0.001$ uncorrected at peak level for visualisation purposes, extent threshold $k > 0$ voxels). Bar plots represent group mean (±SEM) parameter estimates in non-dimensional units (corresponding to percentage whole-brain mean). Source of anatomical labels: Duvernoy (1999). L: left; R: right; A: auditory; V: visual; attA: auditory attention; attV: visual attention; repA: auditory report; repV: visual report.

8.2.5 Auditory localisation within MR scanner

Aside the main experiment, we tested for successful unisensory auditory localisation inside the scanner despite MR scanner noise and we predicted recruitment of the posterior superior temporal gyrus (in particular, planum temporale) for auditory space perception (Ahveninen et

al., 2014; Barrett & Hall, 2006; Battal et al., 2019; Rauschecker & Tian, 2000; Shapleske et al., 1999).

8.2.5.1 Experimental design and procedure

We used the same auditory stimuli as in the ventriloquist paradigm. Signals were sampled from three positions along the azimuth (-9°, 0° or 9° visual angle). To increase design efficiency, auditory spatial positions were presented in a pseudo-randomized fashion, creating mini-blocks of 3, 2 or 1 trials with the same auditory location. In each trial, a 750 ms inter-trial interval was followed by a 50 ms auditory signal (in one of the three azimuthal positions) and a 2 seconds response interval, during which participants reported as accurately as possible their perceived auditory location using a keypad (Figure 8.5A). Throughout the experiment, participants maintained their gaze on a fixation cross (1° diameter) in the centre of the screen. They completed 2 scanning runs (3 conditions × 42 trails / condition / run × 2 runs = 252 trails in total), with each run divided into 7 task blocks (18 trials / block) and 7 fixation blocks presented in an interleaved fashion. Participants were familiarized with stimuli and procedure via one practice run before entering the scanner room.

8.2.5.2 Experimental setup

The experiment was presented via Psychtoolbox version 3.0.11 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) running under MATLAB R2011b (MathWorks Inc.) on a MacBook Pro (Mac OSX 10.6.8). Auditory stimuli were played using MR-compatible headphones (MR Confon HP-VS03). Participants gave responses via one MR-compatible keypad (NATA LXPAD 1×5-10M, NATAtech.com) with their right hand.

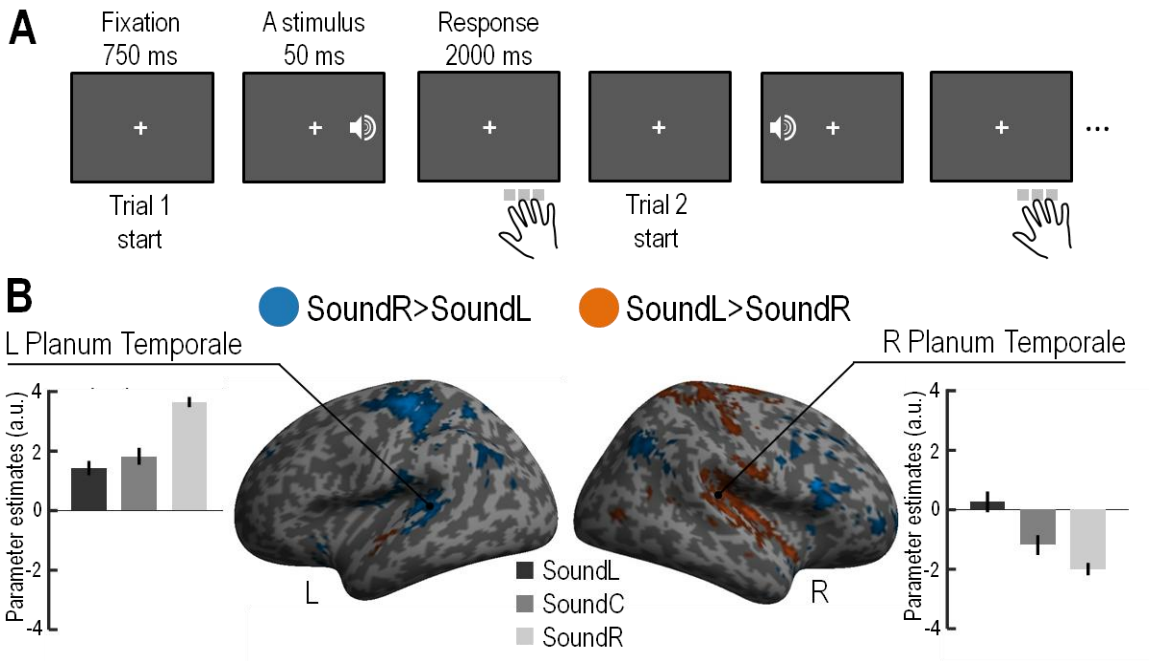

Figure 8.5: fMRI results: Auditory localisation within MR scanner

A) Experimental procedure: after each sound presentation, participants reported their perceived auditory location via button press with the correspondent key. B) Increases of BOLD response for lateralised right versus left sounds (blue) and vice-versa (orange) are rendered on an inflated canonical brain ($p < 0.001$ uncorrected at peak level for visualisation purposes, extent threshold $k > 0$ voxels). Bar plots represent group mean (±SEM) parameter estimates in non-dimensional units (corresponding to percentage whole-brain mean). Source of anatomical labels: Duvernoy (1999). L: left; C: centre; R: right.

8.2.5.3 MRI data acquisition

We acquired T2*-weighted axial echoplanar images (EPI) with blood-oxygenation-level-dependent contrast (gradient echo, SENSE factor of 2, TR = 2800 ms, TE = 40 ms, flip angle = 90°, FOV = 192 × 192 × 114 mm^2, 38 axial slices acquired in sequential ascending direction, voxel size = 2.5 × 2.5 × 2.5 mm^3 + 0.5 mm interslice gap). A total of 128 volumes times 2 runs were acquired, after discarding the first four volumes of each run to allow for T1 equilibration effects. Data acquisition was performed during one scanning day.

8.2.5.4 Behavioural data analysis and results

Participants' spatial localization reliability was quantified by computing the root-mean-square error (RMSE) between participants' reported location and signal's true location, pulling over all spatial locations. Every participant showed RMSE < 5.5°, resulting in group mean RMSE (± SEM) = 3.268° (± 0.290°)[1]. In addition, participants' reported spatial locations were strongly correlated with the true auditory signal locations (across-participants mean ±SEM Fisher-z transformed Pearson correlation coefficient $z = 1.563$ (± 0.080), $p < 0.001$ for two-tailed one-sample Wilcoxon signed-ranks test against zero, after Fisher-z transformation of individual correlation coefficients).

8.2.5.5 fMRI data analysis and results

MRI data were analysed using SPM12 (Wellcome Department of Imaging Neuroscience, London; www.fil.ion.ucl.ac.uk/spm; Friston et al., 1994a). We applied the same pre-processing pipeline as in the univariate analysis of the main experiment (see Section 4.2.7.2). In an event-related design, unit impulses representing stimuli onsets were convolved with a canonical hemodynamic response function and its first temporal derivative. The three experimental conditions (i.e. three auditory locations) were included as regressors in the design matrix. Realignment parameters were also added as nuisance covariates to account for noise due to residual head motion artefacts. The voxel-wise magnitude of the BOLD signal in response to the audio-visual onsets was defined by the parameter estimates pertaining to the canonical hemodynamic response function. Following a hierarchical summary statistics approach, subject-specific images were entered into a first-level general linear model and contrasts (each experimental condition versus baseline summed over the two runs) were

[1] Threshold was defined as two standards deviations above the group mean RMSE in a preliminary psychophysics pilot study with 8 participants.

passed to a second-level ANOVA, where contrasts of interest were defined. Following random effect analysis, inferences were made at the second level (Friston et al., 1994a), where we checked for the effect of unisensory auditory localisation collapsing across spatial locations (Task > Baseline) and separately for left versus right lateralised sounds (SoundL > SoundR; SoundR > SoundL).

Whole-brain activations are reported at $p < 0.05$ (Family-Wise Error corrected) at the peak level (Friston et al., 1994b) in Table 8.7. In line with our predictions (Ahveninen et al., 2014; Barrett & Hall, 2006; Battal et al., 2019; Rauschecker & Tian, 2000; Shapleske et al., 1999), auditory localisation (i.e. Task > Baseline) increased activations in bilateral planum temporale; in addition, we found increased activations in bilateral superior frontal gyrus, in right parietal operculum and in a motor network encompassing the left pre- and post-central sulcus and the right cerebellum (mapping the hand area), which reflected motor response execution. As shown in Figure 8.5B, lateralised sounds activated contralateral planum temporale, in accordance with emerging theories of opponent channel coding both in humans (Derey et al., 2016) and non-human primates (Ortiz-Rios et al., 2017). Overall, behavioural and fMRI data provide converging evidence that participants successfully processed auditory spatial locations in the scanning environment despite MR scanner noise.

Brain regions	MNI coordinates (mm)			Cluster size (voxels)	z-score (peak)	$p_{FWE\text{-}corrected}$ (peak)
	x	y	z			
Task > Baseline						
R cerebellum	18	-52	-24	618	7.47	0.000
L planum temporale	-50	-30	10	489	7.29	0.000
R planum temporale	64	-34	16	114	6.42	0.000
R parietal operculum	64	-26	22		5.55	0.000
L post-central gyrus	-36	-28	50	795	6.72	0.000
L pre-central gyrus	-36	-12	64		6.42	0.000
R superior frontal gyrus	26	-16	46	79	6.25	0.000
L superior frontal gyrus	-22	-16	52	120	6.22	0.000
SoundL > SoundR						
R pre-central gyrus	26	-16	74	54	6.34	0.000
R transverse temporal gyrus	34	-22	6	61	5.95	0.000
R parietal operculum	52	-24	22	29	5.75	0.001
R planum temporale	50	-32	18		4.89	0.031
R planum temporale	52	-22	4	40	5.74	0.001
R superior parietal lobule	22	-52	72	76	5.72	0.001
SoundR > SoundL						
R inferior frontal gyrus (pars opercularis)	52	20	28	288	7.13	0.000
L post-central gyrus	-36	-28	52	654	7.03	0.000
L planum temporale	-50	-32	8	143	6.30	0.000
L parietal operculum	-60	-36	14		5.97	0.000
L precuneus	-8	-74	38	48	5.79	0.000
L superior parietal lobule	-38	-54	58	35	5.69	0.001
R superior parietal lobule	38	-68	48	65	5.69	0.001

Table 8.7: fMRI results: Auditory localisation within MR scanner

Effect of unisensory auditory localisation collapsing across spatial locations (Task > Baseline) and separately for left versus right lateralised sounds (soundL > soundR; soundR > soundL). p-values are FWE-corrected at the peak level for multiple comparisons within the entire brain. Source of anatomical labels: Duvernoy (1999). L: left; R: right.

8.3 Chapter 6

8.3.1 Visual oddball task: fMRI results

To evaluate brain activations associated with the visual oddball task, the regressor specifying flashes onsets in the GLM design matrix (see Section 6.2.6) was contrasted against baseline (i.e. Task > Baseline). As shown in Figure 8.6 and summarised in Table 8.8, we found increased BOLD response in a widespread network for motor control, preparation and execution (Hardwick et al., 2018), which most notably encompassed left primary motor cortex mapping the foot area, left secondary motor cortices, bilateral posterior parietal cortex and right medial anterior cerebellum (again, mapping the foot area). Such results clearly descend from participants' motor responses via the right foot. In addition, we found activations in bilateral anterior insula and right dorsal anterior cingulate gyrus. These areas belong to the so-called salience network implicated in detection of salient events and cognitive control (Menon & Uddin, 2010). Finally, we found increased activations in bilateral calcarine cortex, in line with perceptual processing of the visual targets. As a whole, the visual oddball task recruited a widespread network responsible for targets detection, motor preparation and response, in accordance with proper task execution.

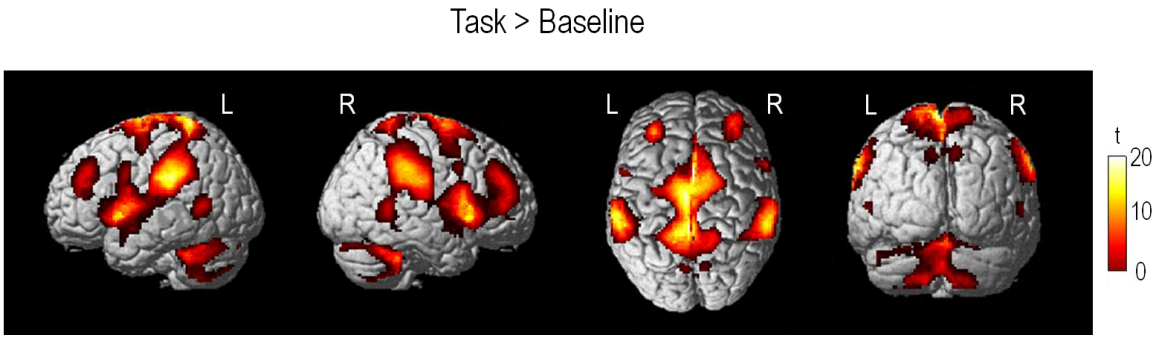

Figure 8.6: fMRI results: Visual oddball task

Activation increases are rendered on a canonical brain ($p < 0.05$ FWE-corrected at cluster level, with auxiliary uncorrected peak-level threshold of $p < 0.001$). L: left; R: right.

Brain regions	**MNI coordinates (mm)**			**Cluster size (voxels)**	**z-score (peak)**	**$p_{FWE\text{-}corrected}$ (cluster)**
	x	y	z			
Task > Baseline						
L precentral gyrus	-6	2	46	71784	> 8	0.000
L central lobule	-6	-16	74		> 8	
L frontal operculum	-44	0	10		> 8	
L precuneus	-12	-44	76		> 8	
L anterior insula	-42	4	8		> 8	
R anterior insula	36	14	8		> 8	
R middle cingulate gyrus	4	10	44		> 8	
R anterior cingulate gyrus	2	16	36		> 8	
L supramarginal gyrus	-56	-30	28		> 8	
R supramarginal gyrus	56	-32	28		> 8	
L putamen	-28	0	10		> 8	
L thalamus	-12	-18	12		> 8	
R cerebellum (Vermis I-IV)	2	-48	-4		> 8	
R calcarine cortex	28	-62	6		4.73	
L calcarine cortex	-26	-66	4		4.66	

Table 8.8: fMRI results: Visual oddball task

p-values are FWE-corrected at the cluster level for multiple comparisons within the entire brain. Auxiliary uncorrected peak-level threshold of $p < 0.001$. Source of anatomical labels: Duvernoy (1999). L: left; R: right.

LIST OF REFERENCES

Adam, R., & Noppeney, U. (2010). Prior auditory information shapes visual category-selectivity in ventral occipito-temporal cortex. *NeuroImage*, *52*(4), 1592–1602. http://doi.org/10.1016/j.neuroimage.2010.05.002

Ahveninen, J., Kopčo, N., & Jääskeläinen, I. P. (2014). Psychophysics and neuronal bases of sound localization in humans. *Hearing Research*, *307*(2), 86–97. http://doi.org/10.1016/j.heares.2013.07.008

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*(3), 257–262. http://doi.org/10.1016/j.cub.2004.01.029

Alais, D., Newell, F., & Mamassian, P. (2010). Multisensory processing in review: from physiology to behaviour. *Seeing and Perceiving*, *23*(1), 3–38. http://doi.org/10.1163/187847510X488603

Aller, M., & Noppeney, U. (2019). To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian causal inference. *PLOS Biology*, *17*(4), e3000210. http://doi.org/10.1371/journal.pbio.3000210

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., & Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, *59*(4), 3677–3689. http://doi.org/10.1016/j.neuroimage.2011.11.019

Alluri, V., Toiviainen, P., Lund, T. E., Wallentin, M., Vuust, P., Nandi, A. K., … Brattico, E. (2013). From vivaldi to beatles and back: Predicting lateralized brain responses to music. *NeuroImage*, *83*, 627–636. http://doi.org/10.1016/j.neuroimage.2013.06.064

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, *15*(9), 839–843. http://doi.org/10.1016/j.cub.2005.03.046

Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research*, *183*(3), 399–404. http://doi.org/10.1007/s00221-007-1110-1

Alsius, A., & Soto-Faraco, S. (2011). Searching for audiovisual correspondence in multiple speaker scenarios. *Experimental Brain Research*, *213*(2-3), 175–183. http://doi.org/10.1007/s00221-011-2624-0

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences*, *108*(25), 10367–10371. http://doi.org/10.1073/pnas.1104047108

Anton-Erxleben, K., & Carrasco, M. (2013). Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. *Nature Reviews Neuroscience*, *14*(3), 188–200. http://doi.org/10.1038/nrn3443

Ashburner, J., Barnes, G., Chen, C.-C., Daunizeau, J., Flandin, G., Friston, K., … Phillips, C. (2015). SPM12 Manual.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851. http://doi.org/10.1016/j.neuroimage.2005.02.018

Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K. C., & Bizley, J. K. (2018). Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. *Neuron*, *97*(3), 640–655.e4. http://doi.org/10.1016/j.neuron.2017.12.034

Avillac, M., Ben Hamed, S., & Duhamel, J.-R. (2007). Multisensory integration in the ventral intraparietal area of the macaque monkey. *Journal of Neuroscience*, *27*(8), 1922–1932. http://doi.org/10.1523/JNEUROSCI.2646-06.2007

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443. http://doi.org/10.1016/j.tics.2012.06.010

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, *36*(2), 190–201. http://doi.org/10.1037/0012-1649.36.2.190

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science*, *13*(3), 99–102. http://doi.org/10.1111/j.0963-7214.2004.00283.x

Bamiou, D.-E., Musiek, F. E., & Luxon, L. M. (2003). The insula (Island of Reil) and its role in auditory processing. *Brain Research Reviews*, *42*(2), 143–154. http://doi.org/10.1016/S0165-0173(03)00172-3

Barraclough, N. E., Xiao, D., Baker, C. I., Oram, M. W., & Perrett, D. I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience*, *17*(3), 377–391. http://doi.org/10.1162/0898929053279586

Barrett, D. J. K., & Hall, D. A. (2006). Response preferences for “what” and “where” in human non-primary auditory cortex. *NeuroImage*, *32*(2), 968–977. http://doi.org/10.1016/j.neuroimage.2006.03.050

Battal, C., Rezk, M., Mattioni, S., Vadlamudi, J., & Collignon, O. (2019). Representation of auditory motion directions and sound source locations in the human planum temporale. *The Journal of Neuroscience*, *39*(12), 2208–2220. http://doi.org/10.1523/JNEUROSCI.2289-18.2018

Beauchamp, M. S., Lee, K., Argall, B., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*, 809–823. http://doi.org/10.1016/S0896-6273(04)00070-4

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*, 5–18. http://doi.org/10.1016/j.specom.2004.10.011

Bertelson, P., Vroomen, J., De Gelder, B., & Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics*, *62*(2), 321–332. http://doi.org/10.3758/BF03205552

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M.-H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. *Journal of Neuroscience*, *28*(52), 14301–14310. http://doi.org/10.1523/JNEUROSCI.2875-08.2008

Bishop, C. W., & Miller, L. M. (2011). Speech cues contribute to audiovisual spatial integration. *PLoS ONE*, *6*(8). http://doi.org/10.1371/journal.pone.0024016

Bisley, J. W., & Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annual Review of Neuroscience*, *33*(1), 1–21. http://doi.org/10.1146/annurev-neuro-060909-152823

Bizley, J. K., & King, A. J. (2009). Visual influences on ferret auditory cortex. *Hearing Research*, *258*(1-2), 55–63. http://doi.org/10.1016/j.heares.2009.06.017

Bizley, J. K., Maddox, R. K., & Lee, A. K. C. (2016). Defining auditory-visual objects: behavioral tests and physiological mechanisms. *Trends in Neurosciences*, *39*(2), 74–85. http://doi.org/10.1016/j.tins.2015.12.007

Bizley, J. K., Nodal, F. R., Bajo, V. M., Nelken, I., & King, A. J. (2007). Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cerebral Cortex*, *17*(9), 2172–2189. http://doi.org/10.1093/cercor/bhl128

Blignaut, P. (2009). Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception, & Psychophysics*, *71*(4), 881–895. http://doi.org/10.3758/APP.71.4.881

Boulter, L. R. (1977). Attention and reaction times to signals of uncertain modality. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(3), 379–388. http://doi.org/10.1037/0096-1523.3.3.379

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. http://doi.org/10.1163/156856897X00357

Bremmer, F., Schlack, A., Shah, N. J., Zafiris, O., Kubischik, M., Hoffmann, K.-P., … Fink, G. R. (2001). Polymodal motion processing in posterior parietal and premotor cortex. *Neuron*, *29*(1), 287–296. http://doi.org/10.1016/s0896-6273(01)00198-2

Broderick, M. P., Anderson, A. J., & Lalor, E. C. (2019). Semantic context enhances the early auditory encoding of natural speech. *The Journal of Neuroscience*, 0584–19. http://doi.org/10.1523/JNEUROSCI.0584-19.2019

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, and Psychophysics*, *77*(5), 1465–1487. http://doi.org/10.3758/s13414-015-0882-9

Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology*, *46*(2), 369–384. http://doi.org/10.1152/jn.1981.46.2.369

Buschman, T. J., & Kastner, S. (2015). From Behavior to Neural Dynamics: An Integrated Theory of Attention. *Neuron*, *88*(1), 127–144. http://doi.org/10.1016/j.neuron.2015.09.017

Busse, L., Roberts, K. C., Crist, R. E., Weissman, D. H., & Woldorff, M. G. (2005). The spread of attention across modalities and space in a multisensory object. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(51), 18751–6. http://doi.org/10.1073/pnas.0507704102

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*(11), 649–657. http://doi.org/10.1016/S0960-9822(00)00513-3

Calvert, G. a., & Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology-Paris*, *98*(1-3), 191–205. http://doi.org/10.1016/j.jphysparis.2004.03.018

Cao, Y., Summerfield, C., Park, H., Giordano, B. L., & Kayser, C. (2019). Causal inference in the multisensory brain. *Neuron*, *102*, 1–12. http://doi.org/10.1016/j.neuron.2019.03.043

Cappe, C., & Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *European Journal of Neuroscience*, *22*(11), 2886–2902. http://doi.org/10.1111/j.1460-9568.2005.04462.x

Carrasco, M. (2011). Visual attention : The past 25 years. *Vision Res.*, *51*(13), 1484–1525. http://doi.org/10.1016/j.visres.2011.04.012.Visual

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. http://doi.org/10.1145/1961189.1961199

Charbonneau, G., Veronneau, M., Boudrias-Fournier, C., Lepore, F., & Collignon, O. (2013). The ventriloquist in periphery: Impact of eccentricity-related reliability on audio-visual localization. *Journal of Vision*, *13*(12), 20–20. http://doi.org/10.1167/13.12.20

Chen, L., & Vroomen, J. (2013). Intersensory binding across space and time: A tutorial review. *Attention, Perception, & Psychophysics*, *75*(5), 790–811. http://doi.org/10.3758/s13414-013-0475-4

Chen, Y.-C., & Spence, C. (2017). Assessing the role of the “Unity Assumption” on multisensory integration: a review. *Frontiers in Psychology*, *8*, 1–22. http://doi.org/10.3389/fpsyg.2017.00445

Chen, Z. (2012). Object-based attention: A tutorial review. *Attention, Perception, and Psychophysics*, *74*(5), 784–802. http://doi.org/10.3758/s13414-012-0322-z

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979.

http://doi.org/10.1121/1.1907229

Chin, T., & Rickard, N. S. (2012). The Music USE (MUSE) questionnaire: An instrument to measure engagement in music. *Music Perception: An Interdisciplinary Journal*, *29*(4), 429–446. http://doi.org/10.1525/mp.2012.29.4.429

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*(1), 73–101. http://doi.org/10.1146/annurev.psych.093008.100427

Ciaramitaro, V. M., Buracas, G. T., & Boynton, G. M. (2007). Spatial and cross-modal attention alter responses to unattended sensory information in early visual and auditory human cortex. *Journal of Neurophysiology*, *98*, 2399–2413. http://doi.org/10.1152/jn.00580.2007

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, *58*(3), 306–324. http://doi.org/10.1016/j.neuron.2008.04.017

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. http://doi.org/10.1038/nrn755

Coull, J. T., & Nobre, A. C. (1998). Where and when to pay attention: The neural systems for directing attention to spatial locations and to time intervals as revealed by both PET and fMRI. *Journal of Neuroscience*, *18*(18), 7426–7435.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, *35*(42), 14195–14204. http://doi.org/10.1523/JNEUROSCI.1829-15.2015

Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *Journal of Neuroscience*, *36*(38), 9888–9895. http://doi.org/10.1523/JNEUROSCI.1396-16.2016

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1-2), 132–147. http://doi.org/10.1016/j.heares.2007.01.014

de la Mothe, L. A., Blumell, S., Kajikawa, Y., & Hackett, T. A. (2006a). Cortical connections of the auditory cortex in marmoset monkeys: Core and medial belt regions. *The Journal of Comparative Neurology*, *496*(1), 27–71. http://doi.org/10.1002/cne.20923

de la Mothe, L. A., Blumell, S., Kajikawa, Y., & Hackett, T. A. (2006b). Thalamic connections of the auditory cortex in marmoset monkeys: Core and medial belt regions. *The Journal of Comparative Neurology*, *496*(1), 72–96. http://doi.org/10.1002/cne.20924

Derey, K., Valente, G., De Gelder, B., & Formisano, E. (2016). Opponent coding of sound location (azimuth) in planum temporale is robust to sound-level variations. *Cerebral Cortex*, *26*(1), 450–464. http://doi.org/10.1093/cercor/bhv269

Deroy, O., & Spence, C. (2016). Crossmodal correspondences: four challenges. *Multisensory Research*, *29*(1-3), 29–48. http://doi.org/10.1163/22134808-00002488

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222. http://doi.org/10.1146/annurev-psych-122414-033400

Desimone, R., & Ungerleider, L. G. (1986). Multiple visual areas in the caudal superior temporal sulcus of the macaque. *The Journal of Comparative Neurology*, *248*(2), 164–189. http://doi.org/10.1002/cne.902480203

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. http://doi.org/10.1016/j.neuroimage.2010.06.010

Disbergen, N. R., Valente, G., Formisano, E., & Zatorre, R. J. (2018). Assessing top-down and bottom-up contributions to auditory stream segregation and integration with polyphonic music. *Frontiers in Neuroscience*, *12*(121). http://doi.org/10.3389/fnins.2018.00121

Donohue, S. E., Green, J. J., & Woldorff, M. G. (2015). The effects of attention on the temporal integration of multisensory stimuli. *Frontiers in Integrative Neuroscience*, *9*, 1–14. http://doi.org/10.3389/fnint.2015.00032

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. *Journal of Neuroscience*, *31*(22), 7982–7990. http://doi.org/10.1523/JNEUROSCI.5298-10.2011

Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on "sensory-specific" brain regions, neural responses, and judgments. *Neuron*, *57*(1), 11–23. http://doi.org/10.1016/j.neuron.2007.12.013

Driver, J., & Spence, C. (1998). Cross-modal links in spatial attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *353*(1373), 1319–1331.

Drullman, R., & Bronkhorst, A. W. (2004). Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *116*(5), 3090–3098. http://doi.org/10.1121/1.1802535

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. http://doi.org/10.1016/j.tics.2010.01.004

Duvernoy, H. M., & Guyot, J. (1999). *Human brain stem vessels: including the pineal gland and information on brain stem infarction*. Springer Science & Business Media.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–1335. http://doi.org/10.1016/j.neuroimage.2004.12.034

Eramudugolla, R., Henderson, R., & Mattingley, J. B. (2011). Effects of audio-visual integration on the detection of masked speech and non-speech sounds. *Brain and Cognition*, *75*(1), 60–66. http://doi.org/10.1016/j.bandc.2010.09.005

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, *7*(5), 7. http://doi.org/10.1167/7.5.7

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. http://doi.org/10.1038/415429a

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–169. http://doi.org/10.1016/j.tics.2004.02.002

Fairhall, S. L., & Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *European Journal of Neuroscience*, *29*(6), 1247–1257. http://doi.org/10.1111/j.1460-9568.2009.06688.x

Faul F, Erdfelder ER, Lang AG, & Buchner A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior. *Behavior Research Methods*, *39*(2), 175–191.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. http://doi.org/10.3758/BRM.41.4.1149

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, *110*(41), 16616–16621. http://doi.org/10.1073/pnas.1315235110

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 1–23. http://doi.org/10.3389/fnhum.2010.00215

Ferrari, A., Degano, G. & Noppeney, U. (in preparation). Cross-modal binding captures attention within a cocktail-party scenario.

Ferrari, A. & Noppeney, U. (in preparation). Attention modulates sensory reliability and impacts response selection during multisensory perceptual inference.

Fiebelkorn, I. C., Foxe, J. J., & Molholm, S. (2010). Dual mechanisms for the cross-sensory

spread of attention: How much do learned associations matter? *Cerebral Cortex*, *20*(1), 109–120. http://doi.org/10.1093/cercor/bhp083

Fiebelkorn, I. C., Foxe, J. J., & Molholm, S. (2012). Attention and Multisensory Feature Integration (pp. 383-394). In *The new handbook of multisensory processing*. Cambridge, MA, USA: MIT Press.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. http://doi.org/10.1016/j.neuroimage.2012.01.021

Fletcher, M. D., Mills, S. R., & Goehring, T. (2018). Vibro-tactile enhancement of speech intelligibility in multi-talker noise for simulated cochlear implant listening. *Trends in Hearing*, *22*, 1–11. http://doi.org/10.1177/2331216518797838

Foxe, J. J., Morocz, I. A., Murray, M. M., Higgins, B. A., Javitt, D. C., & Schroeder, C. E. (2000). Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Cognitive Brain Research*, *10*(1-2), 77–83. http://doi.org/10.1016/S0926-6410(00)00024-0

Foxe, J. J., & Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing. *NeuroReport*, *16*(5), 419–423. http://doi.org/10.1097/00001756-200504040-00001

Foxe, J. J., Simpson, G. V, & Ahlfors, S. P. (1998). Parieto-occipital ~10Hz activity reflects anticipatory state of visual attention mechanisms. *NeuroReport*, *9*(17), 3929–3933. http://doi.org/10.1097/00001756-199812010-00030

Foxe, J. J., Wylie, G. R., Martinez, A., Schroeder, C. E., Javitt, D. C., Guilfoyle, D., … Murray, M. M. (2002). Auditory-somatosensory multisensory processing in auditory association cortex: an fMRI study. *Journal of Neurophysiology*, *88*(1), 540–543. http://doi.org/DOI 10.1152/jn.00694.2001

Frassinetti, F., Bolognini, N., & Làdavas, E. (2002a). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research*, *147*(3), 332–343. http://doi.org/10.1007/s00221-002-1262-y

Frassinetti, F., Pavani, F., & Ladavas, E. (2002b). Acoustical vision of neglected stimuli: Interaction among spatially converging audiovisual inputs in neglect patients. *Journal of Cognitive Neuroscience*, *14*(1), 62–69. http://doi.org/10.1162/089892902317205320

Freides, D. (1974). Human information processing and sensory modality: cross-modal functions, information complexity, memory, and deficit. *Psychological Bulletin*, *81*(5), 284–310. http://doi.org/10.1037/h0036331

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. http://doi.org/10.1038/nrn2787

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., & Turner, R. (1998). Event-related fMRI: Characterizing differential responses. *NeuroImage*, *7*(1), 30–40. http://doi.org/10.1006/nimg.1997.0306

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994a). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210. http://doi.org/10.1002/hbm.460020402

Friston, K. J., Price, C. J., Fletcher, P., Moore, C., Frackowiak, R. S. J., & Dolan, R. J. (1996). The trouble with cognitive subtraction. *NeuroImage*, *4*(2), 97–104. http://doi.org/10.1006/nimg.1996.0033

Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C., & Evans, A. C. (1994b). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, *1*(3), 210–220. http://doi.org/10.1002/hbm.460010306

Fu, K. M., Foxe, J. J., Murray, M. M., Higgins, B. A., Javitt, D. C., & Schroeder, C. E. (2001). Attention-dependent suppression of distracter visual input can be cross-modally cued as indexed by anticipatory parieto-occipital alpha-band oscillations. *Brain Research. Cognitive Brain Research*, *12*(1), 145–52.

Fu, K.-M. G., Johnston, T. A., Shah, A. S., Arnold, L., Smiley, J., Hackett, T. A., … Schroeder, C. E. (2003). Auditory cortical neurons respond to somatosensory stimulation. *Journal of Neuroscience*, *23*(20), 7510–5.

Gardner, W. G., & Martin, K. D. (1995). HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America*, *97*(6), 3907–3908. http://doi.org/10.1121/1.412407

Gau, R., & Noppeney, U. (2016). How prior expectations shape multisensory perception. *NeuroImage*, *124*, 876–886. http://doi.org/10.1016/j.neuroimage.2015.09.045

Geisler, W. S., & Diehl, R. L. (2002). Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *357*(1420), 419–448. http://doi.org/10.1098/rstb.2001.1055

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, *10*(6), 278–285. http://doi.org/10.1016/j.tics.2006.04.008

Giessing, C., Thiel, C. M., Stephan, K. E., Rösler, F., & Fink, G. R. (2004). Visuospatial attention: How to measure effects of infrequent, unattended events in a blocked stimulus design. *NeuroImage*, *23*(4), 1370–1381. http://doi.org/10.1016/j.neuroimage.2004.08.008

Gillmeister, H., & Eimer, M. (2007). Tactile enhancement of auditory detection and perceived loudness. *Brain Research*, *1160*(1), 58–68. http://doi.org/10.1016/j.brainres.2007.03.041

Gleiss, S., & Kayser, C. (2013). Eccentricity dependent auditory enhancement of visual stimulus detection but not discrimination. *Frontiers in Integrative Neuroscience*, *7*, 1–8. http://doi.org/10.3389/fnint.2013.00052

Gomez-Ramirez, M., Kelly, S. P., Molholm, S., Sehatpour, P., Schwartz, T. H., & Foxe, J. J. (2011). Oscillatory sensory selection mechanisms during intersensory attention to rhythmic auditory and visual inputs: A human electrocorticographic investigation. *Journal of Neuroscience*, *31*(50), 18556–18567. http://doi.org/10.1523/jneurosci.2164-11.2011

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–5. http://doi.org/10.1093/litthe/11.1.80

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive

development. *Developmental Science*, *10*(3), 281–287. http://doi.org/10.1111/j.1467-7687.2007.00584.x

Göschl, F., Engel, A. K., & Friese, U. (2014). Attention modulates visual-tactile interaction in spatial pattern matching. *PLoS ONE*, *9*(9), e106896. http://doi.org/10.1371/journal.pone.0106896

Grant, K. W., & Seitz, P.-F. (2002). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197. http://doi.org/10.1121/1.1288668

Griffiths, T. D., & Warren, J. D. (2002). The planum temporale as a computational hub. *Trends in Neurosciences*, *25*(7), 348–353. http://doi.org/10.1016/S0166-2236(02)02191-4

Hackett, T. A., Smiley, J. F., Ulbert, I., Karmos, G., Lakatos, P., de la Mothe, L. A., & Schroeder, C. E. (2007). Sources of somatosensory input to the caudal belt areas of auditory cortex. *Perception*, *36*(10), 1419–1430. http://doi.org/10.1068/p5841

Hannemann, R., Obleser, J., & Eulitz, C. (2007). Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain Research*, *1153*(1), 134–143. http://doi.org/10.1016/j.brainres.2007.03.069

Hardwick, R. M., Caspers, S., Eickhoff, S. B., & Swinnen, S. P. (2018). Neural correlates of action: Comparing meta-analyses of imagery, observation, and execution. *Neuroscience & Biobehavioral Reviews*, *94*(August), 31–44. http://doi.org/10.1016/j.neubiorev.2018.08.003

Haynes, J. D. (2015). A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. *Neuron*, *87*(2), 257–270. http://doi.org/10.1016/j.neuron.2015.05.025

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *NeuroImage*, *180*, 4–18. http://doi.org/10.1016/j.neuroimage.2017.08.005

Hebart, M. N., Görgen, K., Haynes, J.-D., & Dubois, J. (2015). The Decoding Toolbox

(TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*, 1–18. http://doi.org/10.3389/fninf.2014.00088

Helbig, H. B., & Ernst, M. O. (2007). Knowledge about a common source can promote visual-haptic integration. *Perception*, *36*(10), 1523–1533. http://doi.org/10.1068/p5851

Helbig, H. B., & Ernst, M. O. (2008). Visual-haptic cue weighting is independent of modality-specific attention. *Journal of Vision*, *8*(1), 1–16. http://doi.org/10.1167/8.1.21

Helfer, K. S., & Freyman, R. L. (2005). The role of visual speech cues in reducing energetic and informational masking. *The Journal of the Acoustical Society of America*, *117*(2), 842–849. http://doi.org/10.1121/1.1836832

Henson, R. (2006). Efficient experimental design for fMRI. In *Statistical parametric mapping: The analysis of functional brain images* (pp. 193–210).

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision*, *4*(12), 967–992. http://doi.org/10.1167/4.12.1

Ho, C., Santangelo, V., & Spence, C. (2009). Multisensory warning signals: When spatial correspondence matters. *Experimental Brain Research*, *195*(2), 261–272. http://doi.org/10.1007/s00221-009-1778-5

Ho, C., Tan, H. Z., & Spence, C. (2005). Using spatial vibrotactile cues to direct visual attention in driving scenes. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(6), 397–412. http://doi.org/10.1016/j.trf.2005.05.002

Hoefer, M., Tyll, S., Kanowski, M., Brosch, M., Schoenfeld, M. A., Heinze, H.-J., & Noesselt, T. (2013). Tactile stimulation and hemispheric asymmetries modulate auditory perception and neural responses in primary auditory cortex. *NeuroImage*, *79*, 371–382. http://doi.org/10.1016/j.neuroimage.2013.04.119

Hoefle, S., Engel, A., Basilio, R., Alluri, V., Toiviainen, P., Cagy, M., & Moll, J. (2018). Identifying musical pieces from fMRI data using encoding and decoding models.

Scientific Reports, *8*(1), 1–13. http://doi.org/10.1038/s41598-018-20732-3

Huang, J., Gamble, D., Sarnlertsophon, K., Wang, X., & Hsiao, S. (2012). Feeling music: Integration of auditory and tactile inputs in musical meter perception. *PLoS ONE*, *7*(10), e48496. http://doi.org/10.1371/journal.pone.0048496

Huang, J., Sheffield, B., Lin, P., & Zeng, F. G. (2017). Electro-tactile stimulation enhances cochlear implant speech recognition in noise. *Scientific Reports*, *7*(1), 1–5. http://doi.org/10.1038/s41598-017-02429-1

Huettel, S. A., Song, A. W., & McCarthy, G. (2004). *Functional Magnetic Resonance Imaging* (Vol. 1). Sunderland, MA: Sinauer Associates.

Humphreys, G. W., & Riddoch, M. J. (2003). From what to where: Neuropsychological Evidence for Implicit Interactions between Object- and Space-based Attention. *Psychological Science*, *14*(5), 487–492. http://doi.org/10.1111/1467-9280.02457

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489–1506. http://doi.org/10.1016/S0042-6989(99)00163-7

Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the “ventriloquism” effect. *Perceptual and Motor Skills*, *37*, 967–979. http://doi.org/10.2466/pms.1973.37.3.967

James, T., & Stevenson, R. (2012). The use of fMRI to assess multisensory integration. In M. M. Murray & M. T. Wallace (Eds.), *The Neural Bases of Multisensory Processes*. Boca Raton (FL): CRC Press. http://doi.org/10.1201/b11092-11

Janata, P. (2015). Neural basis of music perception. In G. G. Celesia & G. Hickok (Eds.), *Handbook of Clinical Neurology* (Vol. 129, pp. 187–205). Elsevier B.V. http://doi.org/10.1016/B978-0-444-62630-1.00011-1

Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Frontiers in Human Neuroscience*, *4*, 186.

http://doi.org/10.3389/fnhum.2010.00186

Jicol, C., Proulx, M. J., Pollick, F. E., & Petrini, K. (2018). Long-term music training modulates the recalibration of audiovisual simultaneity. *Experimental Brain Research*, *236*(7), 1869–1880. http://doi.org/10.1007/s00221-018-5269-4

Johnson, J. A., & Zatorre, R. J. (2005). Attention to simultaneous unrelated auditory and visual events: Behavioral and neural correlates. *Cerebral Cortex*, *15*(10), 1609–1620. http://doi.org/10.1093/cercor/bhi039

Johnson, J. A., & Zatorre, R. J. (2006). Neural substrates for dividing and focusing attention between simultaneous auditory and visual events. *NeuroImage*, *31*(4), 1673–81. http://doi.org/10.1016/j.neuroimage.2006.02.026

Jones, S. A., Beierholm, U., Meijer, D., & Noppeney, U. (2019). Older adults sacrifice response speed to preserve multisensory integration performance. *Neurobiology of Aging*, *84*, 148–157. http://doi.org/10.1016/j.neurobiolaging.2019.08.017

Juan, C. H., Shorter-Jacobi, S. M., & Schall, J. D. (2004). Dissociation of spatial attention and saccade preparation. *Proceedings of the National Academy of Sciences of the USA*, *101*(43), 15541–15544. http://doi.org/10.1073/pnas.0403507101

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences*, 1–14. http://doi.org/10.1016/j.tics.2019.04.013

Kanaya, S., & Yokosawa, K. (2011). Perceptual congruency of audio-visual speech affects ventriloquism with bilateral visual stimuli. *Psychonomic Bulletin and Review*, *18*(1), 123–128. http://doi.org/10.3758/s13423-010-0027-z

Kastner, S., Pinsk, M. a., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, *22*(4), 751–761. http://doi.org/10.1016/S0896-6273(00)80734-5

Kayser, C., & Logothetis, N. K. (2007). Do early sensory cortices integrate cross-modal information? *Brain Structure and Function*, *212*(2), 121–132.

http://doi.org/10.1007/s00429-007-0154-0

Kayser, C., Logothetis, N. K., & Panzeri, S. (2010). Visual enhancement of the information representation in auditory cortex. *Current Biology*, *20*(1), 19–24. http://doi.org/10.1016/j.cub.2009.10.068

Kayser, C., Petkov, C. I., Augath, M., & Logothetis, N. K. (2005). Integration of touch and sound in auditory cortex. *Neuron*, *48*(2), 373–384. http://doi.org/10.1016/j.neuron.2005.09.018

Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, *15*(21), 1943–1947. http://doi.org/10.1016/j.cub.2005.09.040

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex*, *18*(7), 1560–1574. http://doi.org/10.1093/cercor/bhm187

Kimchi, R., Yeshurun, Y., & Cohen-Savransky, A. (2007). Automatic, stimulus-driven attentional capture by objecthood. *Psychonomic Bulletin & Review*, *14*(1), 166–72.

Kimchi, R., Yeshurun, Y., Spehar, B., & Pirkner, Y. (2016). Perceptual organization, visual attention, and objecthood. *Vision Research*, *126*, 34–51. http://doi.org/10.1016/j.visres.2015.07.008

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). What's new in Psychtoobox-3? *Perception*, *36*, 14.

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica*, *134*(3), 372–384. http://doi.org/10.1016/j.actpsy.2010.03.010

Koelsch, S., & Siebel, W. A. (2005). Towards a neural basis of music perception. *Trends in Cognitive Sciences*, *9*(12), 578–584. http://doi.org/10.1016/j.tics.2005.10.001

Koelsch, S., Vuust, P., & Friston, K. (2019). Predictive processes and the peculiar case of

music. *Trends in Cognitive Sciences*, *23*(1), 63–77. http://doi.org/10.1016/j.tics.2018.10.006

Körding, K. P., Beierholm, U. R., Ma, W. J., Quartz, S. R., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS One*, *2*(9), e943. http://doi.org/10.1371/journal.pone.0000943

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, *38*(4), 649–662. http://doi.org/10.1016/j.neuroimage.2007.02.022

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. http://doi.org/10.1073/pnas.0600244103

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. http://doi.org/10.1080/23273798.2015.1102299

Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron*, *53*(2), 279–292. http://doi.org/10.1016/j.neuron.2006.12.011

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*(3), 420–429. http://doi.org/10.1162/089892902317361930

Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T., & Stein, B. E. (2005). On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Experimental Brain Research*, *166*(3-4), 289–297. http://doi.org/10.1007/s00221-005-2370-2

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences*, *9*(2), 75–82. http://doi.org/10.1016/j.tics.2004.12.004

Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current Directions in Psychological Science*, *19*(3), 143–148. http://doi.org/10.1177/0963721410370295

Lee, H., & Noppeney, U. (2011a). Long-term music training tunes how the brain temporally binds signals from multiple senses. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(51), E1441–50. http://doi.org/10.1073/pnas.1115267108

Lee, H., & Noppeney, U. (2011b). Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension. *Journal of Neuroscience*, *31*(31), 11338–11350. http://doi.org/10.1523/JNEUROSCI.6510-10.2011

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.

Lewald, J., & Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Research*, *16*(3), 468–478. http://doi.org/10.1016/S0926-6410(03)00074-0

Lewis, R., & Noppeney, U. (2010). Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. *Journal of Neuroscience*, *30*(37), 12329–12339. http://doi.org/10.1523/JNEUROSCI.5745-09.2010

Liang, M., Mouraux, A., Hu, L., & Iannetti, G. D. (2013). Primary sensory cortices contain distinguishable spatial patterns of activity for each sense. *Nature Communications*, *4*(1), 1979. http://doi.org/10.1038/ncomms2979

Lickliter, R., & Bahrick, L. E. (2013). The concept of homology as a basis for evaluating developmental mechanisms: Exploring selective attention across the life-span. *Developmental Psychobiology*, *55*(1), 76–83. http://doi.org/10.1002/dev.21037

Linden, J. F., Grunewald, A., & Andersen, R. A. (1999). Responses to auditory stimuli in macaque lateral intraparietal area II. Behavioral modulation. *Journal of Neurophysiology*, *82*(1), 343–358. http://doi.org/10.1152/jn.1999.82.1.343

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, *412*(6843), 150–157. http://doi.org/10.1038/35084005

Love, S. A., Petrini, K., Pernet, C. R., Latinus, M., & Pollick, F. E. (2018). Overlapping but divergent neural correlates underpinning audiovisual synchrony and temporal order judgments. *Frontiers in Human Neuroscience*, *12*, 1–11. http://doi.org/10.3389/fnhum.2018.00274

Love, S. A., Pollick, F. E., & Petrini, K. (2012). Effects of Experience, Training and Expertise on Multisensory Perception: Investigating the Link between Brain and Behavior. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7403, pp. 304–320). http://doi.org/10.1007/978-3-642-34584-5_27

Lovelace, C. T., Stein, B. E., & Wallace, M. T. (2003). An irrelevant light enhances auditory detection in humans: A psychophysical analysis of multisensory integration in stimulus detection. *Cognitive Brain Research*, *17*(2), 447–453. http://doi.org/10.1016/S0926-6410(03)00160-5

Macaluso, E., & Doricchi, F. (2013). Attention and predictions: control of spatial attention beyond the endogenous-exogenous dichotomy. *Frontiers in Human Neuroscience*, *7*, 75–80. http://doi.org/10.3389/fnhum.2013.00685

Macaluso, E., Driver, J., & Frith, C. D. (2003). Multimodal spatial representations engaged in human parietal cortex during both saccadic and manual spatial orienting. *Current Biology*, *13*(12), 990–999. http://doi.org/10.1016/S0960-9822(03)00377-4

Macaluso, E., Noppeney, U., Talsma, D., Vercillo, T., Hartcher-O'Brien, J., & Adam, R. (2016). The curious incident of attention in multisensory integration: Bottom-up vs. top-down. *Multisensory Research*, *29*(6-7), 557–583. http://doi.org/10.1163/22134808-00002528

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. C. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human

listeners. *eLife*, *4*, 1–11. http://doi.org/10.7554/eLife.04995

Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. *Frontiers in Psychology*, *4*(NOV), 1–10. http://doi.org/10.3389/fpsyg.2013.00854

Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 245–256. http://doi.org/10.1037/a0019952

Marinato, G., & Baldauf, D. (2019). Object-based attention in complex, naturalistic auditory streams. *Scientific Reports*, *9*(1), 2854. http://doi.org/10.1038/s41598-019-39166-6

Marks, L. E., Ben-Artzi, E., & Lakatos, S. (2003). Cross-modal interactions in auditory and visual discrimination. *International Journal of Psychophysiology*, *50*(1-2), 125–145. http://doi.org/10.1016/S0167-8760(03)00129-6

Martinez-Trujillo, J. C., & Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, *14*(9), 744–751. http://doi.org/10.1016/j.cub.2004.04.028

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. http://doi.org/10.1093/cercor/bhl077

Mast, F., Frings, C., & Spence, C. (2015). Multisensory top-down sets: Evidence for contingent crossmodal capture. *Attention, Perception, & Psychophysics*, *77*, 1970–1985. http://doi.org/10.3758/s13414-015-0915-4

Mastroberardino, S., Santangelo, V., & Macaluso, E. (2015). Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. *Frontiers in Integrative Neuroscience*, *9*. http://doi.org/10.3389/fnint.2015.00045

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953–978. http://doi.org/10.1080/01690965.2012.705006

Matusz, P. J., & Eimer, M. (2011). Multisensory enhancement of attentional capture in visual search. *Psychonomic Bulletin and Review*, *18*(5), 904–909. http://doi.org/10.3758/s13423-011-0131-8

Matusz, P. J., & Eimer, M. (2013). Top-down control of audiovisual search by bimodal search templates. *Psychophysiology*, *50*(10), 996–1009. http://doi.org/10.1111/psyp.12086

Mazaheri, A., van Schouwenburg, M. R., Dimitrijevic, A., Denys, D., Cools, R., & Jensen, O. (2014). Region-specific modulations in oscillatory alpha activity serve to facilitate processing in the visual and auditory modalities. *NeuroImage*, *87*, 356–362. http://doi.org/10.1016/j.neuroimage.2013.10.052

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain Structure and Function*, *214*(5-6), 655–667. http://doi.org/10.1007/s00429-010-0262-0

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of Neuroscience*, *7*(10), 3215–29. http://doi.org/citeulike-article-id:409430

Meredith, M. A., & Stein, B. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science*, *221*(4608), 389–391. http://doi.org/10.1126/science.6867718

Meredith, M. A., & Stein, B. E. (1986). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Research*, *365*(2), 350–354. http://doi.org/10.1016/0006-8993(86)91648-3

Michail, G., & Keil, J. (2018). High cognitive load enhances the susceptibility to non-speech audiovisual illusions. *Scientific Reports*, *8*(1), 11530. http://doi.org/10.1038/s41598-018-30007-6

Misselhorn, J., Daume, J., Engel, A. K., & Friese, U. (2016). A matter of attention: Crossmodal congruence enhances and impairs performance in a novel trimodal matching paradigm. *Neuropsychologia*, *88*, 113–122. http://doi.org/10.1016/j.neuropsychologia.2015.07.022

Molholm, S., Martinez, A., Shpaner, M., & Foxe, J. J. (2007). Object-based attention is multisensory: Co-activation of an object's representations in ignored sensory modalities. *European Journal of Neuroscience*, *26*(2), 499–509. http://doi.org/10.1111/j.1460-9568.2007.05668.x

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex*, *14*(4), 452–465. http://doi.org/10.1093/cercor/bhh007

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research*, *14*(1), 115–128. http://doi.org/10.1016/S0926-6410(02)00066-6

Morís Fernández, L., Visser, M., Ventura-Campos, N., Ávila, C., & Soto-Faraco, S. (2015). Top-down attention regulates the neural expression of audiovisual integration. *NeuroImage*, *119*, 272–285. http://doi.org/10.1016/j.neuroimage.2015.06.052

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., & Zilles, K. (2001). Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. *Neuroimage*, *13*(4), 684–701. http://doi.org/10.1006/nimg.2000.0715

Mozolic, J. L., Hugenschmidt, C. E., Peiffer, A. M., & Laurienti, P. J. (2008a). Modality-specific selective attention attenuates multisensory integration. *Experimental Brain Research*, *184*(1), 39–52. http://doi.org/10.1007/s00221-007-1080-3

Mozolic, J. L., Joyner, D., Hugenschmidt, C. E., Peiffer, A. M., Kraft, R. a, Maldjian, J. a, & Laurienti, P. J. (2008b). Cross-modal deactivations during modality-specific selective attention. *BMC Neurology*, *8*, 35. http://doi.org/10.1186/1471-2377-8-35

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, *58*(3), 351–362. http://doi.org/10.3758/BF03206811

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*(1), 101–109. http://doi.org/10.1093/scan/nsn044

Musacchia, G., & Schroeder, C. E. (2009). Neuronal mechanisms, response dynamics and perceptual functions of multisensory interactions in auditory cortex. *Hearing Research*, *258*(1-2), 72–79. http://doi.org/10.1016/j.heares.2009.06.018

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *The Journal of the Acoustical Society of America*, *132*(2), 1061–1077. http://doi.org/10.1121/1.4728187

Natale, E., Marzi, C. A., & Macaluso, E. (2010). Right temporal-parietal junction engagement during spatial reorienting does not depend on strategic attention control. *Neuropsychologia*, *48*(4), 1160–1164. http://doi.org/10.1016/j.neuropsychologia.2009.11.012

Nobre, A. C., & Kastner, S. (2014). Attention: time capsule 2013. In *The Oxford Handbook of Attention* (pp. 1201–1222). Oxford: Oxford University Press.

Nobre, A. C., & Van Ede, F. (2018). Anticipated moments: Temporal structure in attention. *Nature Reviews Neuroscience*, *19*(1), 34–48. http://doi.org/10.1038/nrn.2017.141

Noesselt, T., Bergmann, D., Hake, M., Heinze, H. J., & Fendrich, R. (2008). Sound increases the saliency of visual events. *Brain Research*, *1220*, 157–163. http://doi.org/10.1016/j.brainres.2007.12.060

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H.-J., & Driver, J. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *Journal of Neuroscience*, *27*(42), 11431–11441. http://doi.org/10.1523/JNEUROSCI.2252-07.2007

Noesselt, T., Tyll, S., Boehler, C. N., Budinger, E., Heinze, H. J., & Driver, J. (2010). Sound-induced enhancement of low-intensity vision: Multisensory influences on human sensory-specific cortices and thalamic bodies relate to perceptual enhancement of visual detection sensitivity. *Journal of Neuroscience*, *30*(41), 13609–13623. http://doi.org/10.1523/JNEUROSCI.4524-09.2010

Noppeney, U. (2012). Characterization of multisensory integration with fMRI. In M. M. Murray & M. T. Wallace (Eds.), *The Neural Bases of Multisensory Processes*. Boca Raton (FL): CRC Press.

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. http://doi.org/10.1515/nf-2017-A066

Noppeney, U., & Lee, H. L. (2018). Causal inference and temporal predictions in audiovisual perception of speech and music. *Annals of the New York Academy of Sciences*, 1–15. http://doi.org/10.1111/nyas.13615

Noppeney, U., Ostwald, D., & Werner, S. (2010). Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *Journal of Neuroscience*, *30*(21), 7434–7446. http://doi.org/10.1523/JNEUROSCI.0455-10.2010

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. http://doi.org/10.1016/j.tics.2006.07.005

Odegaard, B., Wozny, D. R., & Shams, L. (2015). Biases in visual, auditory, and audiovisual perception of space. *PLOS Computational Biology*, *11*(12), e1004649. http://doi.org/10.1371/journal.pcbi.1004649

Odegaard, B., Wozny, D. R., & Shams, L. (2016). The effects of selective and divided attention on sensory precision and integration. *Neuroscience Letters*, *614*, 24–28. http://doi.org/10.1016/j.neulet.2015.12.039

Odgaard, E. C., Arieh, Y., & Marks, L. E. (2003). Cross-modal enhancement of perceived

brightness: sensory interaction versus response bias. *Perception & Psychophysics*, *65*(1), 123–32. http://doi.org/10.3758/BF03194804

Odgaard, E. C., Arieh, Y., & Marks, L. E. (2004). Brighter noise: Sensory enhancement of perceived loudness by concurrent visual stimulation. *Cognitive, Affective and Behavioral Neuroscience*, *4*(2), 127–132. http://doi.org/10.3758/CABN.4.2.127

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. http://doi.org/10.1016/0028-3932(71)90067-4

Ortiz-Rios, M., Azevedo, F. A. C., Kusmierek, P., Balla, D. Z., Munk, M. H., Keliris, G. A., … Rauschecker, J. P. (2017). Widespread and opponent fMRI signals represent sound location in macaque auditory cortex. *Neuron*, *93*(4), 971–983. http://doi.org/10.1016/j.neuron.2017.01.013

Oruc, I., Sinnett, S., Bischof, W. F., Soto-Faraco, S., Lock, K., & Kingstone, A. (2008). The effect of attention on the illusory capture of motion in bimodal stimuli. *Brain Research*, *1242*, 200–208. http://doi.org/10.1016/j.brainres.2008.04.014

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). http://doi.org/10.1371/journal.pone.0005664

Parise, C. V. (2015). Crossmodal correspondences: Standing issues and experimental guidelines. *Multisensory Research*, *29*(July), 7–28. http://doi.org/10.1163/22134808-00002502

Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nature Communications*, *7*(12), 11543. http://doi.org/10.1038/ncomms11543

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Current Biology*, *22*(1), 46–49. http://doi.org/10.1016/j.cub.2011.11.039

Parsons, L. M. (2001). Exploring the functional neuroanatomy of music performance, perception, and comprehension. *Annals of the New York Academy of Sciences*, *930*(1), 211–231. http://doi.org/10.1111/j.1749-6632.2001.tb05735.x

Pashler, H. (1998). *The Psychology of Attention*. Cambridge, MA, USA: MIT Press. http://doi.org/10.1016/s1364-6613(98)01249-2

Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, *1423*(1), 378–395. http://doi.org/10.1111/nyas.13654

Peelen, M. V., & Kastner, S. (2014). Attention in the real world: Toward understanding its neural basis. *Trends in Cognitive Sciences*, *18*(5), 242–250. http://doi.org/10.1016/j.tics.2014.02.004

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

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, *45*, S199–S209. http://doi.org/10.1016/j.neuroimage.2008.11.007

Peretz, I., & Zatorre, R. J. (2005). Brain organization for music processing. *Annual Review of Psychology*, *56*(1), 89–114. http://doi.org/10.1146/annurev.psych.56.091103.070225

Pestilli, F., Carrasco, M., Heeger, D. J., & Gardner, J. L. (2011). Attentional enhancement via selection and pooling of early sensory responses in human visual cortex. *Neuron*, *72*(5), 832–846. http://doi.org/10.1016/j.neuron.2011.09.025

Petrini, K., Crabbe, F., Sheridan, C., & Pollick, F. E. (2011). The music of your emotions: Neural substrates involved in detection of emotional correspondence between auditory and visual music actions. *PLoS ONE*, *6*(4), e19165. http://doi.org/10.1371/journal.pone.0019165

Petrini, K., Dahl, S., Rocchesso, D., Waadeland, C. H., Avanzini, F., Puce, A., & Pollick, F.

E. (2009). Multisensory integration of drumming actions: Musical expertise affects perceived audiovisual asynchrony. *Experimental Brain Research*, *198*(2-3), 339–352. http://doi.org/10.1007/s00221-009-1817-2

Petrini, K., Jones, P. R., Smith, L., & Nardini, M. (2015). Hearing Where the Eyes See: Children Use an Irrelevant Visual Cue When Localizing Sounds. *Child Development*, *86*(5), 1449–1457. http://doi.org/10.1111/cdev.12397

Petrini, K., McAleer, P., & Pollick, F. (2010). Audiovisual integration of emotional signals from music improvisation does not depend on temporal correspondence. *Brain Research*, *1323*, 139–148. http://doi.org/10.1016/j.brainres.2010.02.012

Petrini, K., Pollick, F. E., Dahl, S., McAleer, P., McKay, L., Rocchesso, D., … Puce, A. (2011). Action expertise reduces brain activity for audiovisual matching actions: An fMRI study with expert drummers. *NeuroImage*, *56*(3), 1480–1492. http://doi.org/10.1016/j.neuroimage.2011.03.009

Petrini, K., Remark, A., Smith, L., & Nardini, M. (2014). When vision is not an option: Children's integration of auditory and haptic information is suboptimal. *Developmental Science*, *17*(3), 376–387. http://doi.org/10.1111/desc.12127

Poldrack, R. A., Nichols, T., & Mumford, J. (2011). *Handbook of Functional MRI Data Analysis*. *Handbook of Functional MRI Data Analysis*. Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9780511895029

Pooley, R. A. (2005). Fundamental physics of MR imaging. *Radiographics*, *25*(4), 1087–1099. http://doi.org/10.1148/rg.254055027

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. *Attention and Performance: Control of Language Processes*, *32*, 531–556. http://doi.org/10.1162/jocn.1991.3.4.335

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*(2), 160–174. http://doi.org/10.1037/0096-3445.109.2.160

Price, C. J., Moore, C. J., & Friston, K. J. (1997). Subtractions, conjunctions, and interactions in experimental design of activation studies. *Human Brain Mapping*, *5*(4), 264–272. http://doi.org/10.1002/(SICI)1097-0193(1997)5:4<264::AID-HBM11>3.0.CO;2-E

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. http://doi.org/10.2307/271063

Raichle, M. E. (2000). A default mode of brain function. *Proceedings of the National Academy of Sciences*, *98*(2), 676–682. http://doi.org/10.1002/jso.20675

Raichle, M. E. (2015). The brain's default mode network. *Annual Review of Neuroscience*, *38*(1), 433–447. http://doi.org/10.1146/annurev-neuro-071013-014030

Rauschecker, J. P. (2018). Where, when, and how: Are they all sensorimotor? Towards a unified view of the dorsal pathway in vision and audition. *Cortex*, *98*, 262–268. http://doi.org/10.1016/j.cortex.2017.10.020

Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proceedings of the National Academy of Sciences*, *97*(22), 11800–11806. http://doi.org/10.1073/pnas.97.22.11800

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185. http://doi.org/10.1016/j.neuron.2009.01.002

Riecke, L., Snipes, S., van Bree, S., Kaas, A., & Hausfeld, L. (2019). Audio-tactile enhancement of cortical speech-envelope tracking. *NeuroImage*, *202*(August), 116134. http://doi.org/10.1016/j.neuroimage.2019.116134

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - Revisited. *NeuroImage*, *84*, 971–985. http://doi.org/10.1016/j.neuroimage.2013.08.065

Rohe, T., Ehlis, A.-C., & Noppeney, U. (2019). The neural dynamics of hierarchical Bayesian causal inference in multisensory perception. *Nature Communications*, *10*(1), 1907. http://doi.org/10.1038/s41467-019-09664-2

Rohe, T., & Noppeney, U. (2015a). Cortical hierarchies perform Bayesian Causal Inference in multisensory perception. *PLOS Biology*, *13*(2), e1002073. http://doi.org/10.1371/journal.pbio.1002073

Rohe, T., & Noppeney, U. (2015b). Sensory reliability shapes Bayesian Causal Inference in perception via two mechanisms. *Journal of Vision*, *15*(5), 1–16. http://doi.org/10.1167/15.5.22.doi

Rohe, T., & Noppeney, U. (2016). Distinct computational principles govern multisensory integration in primary sensory and association cortices. *Current Biology*, *26*(4), 509–514. http://doi.org/10.1016/j.cub.2015.12.056

Rohe, T., & Noppeney, U. (2018). Reliability-weighted integration of audiovisual signals can be modulated by top-down control. *eNeuro*, *5*(1), e0315–17. http://doi.org/10.1523/ENEURO.0315-17.2018

Ronconi, L., Casartelli, L., Carna, S., Molteni, M., Arrigoni, F., & Borgatti, R. (2016). When one is Enough: Impaired Multisensory Integration in Cerebellar Agenesis. *Cerebral Cortex*, *27*(3), 2041–2051. http://doi.org/10.1093/cercor/bhw049

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147–1153. http://doi.org/10.1093/cercor/bhl024

Sadaghiani, S., Maier, J. X., & Noppeney, U. (2009). Natural, metaphoric, and linguistic auditory direction signals have distinct influences on visual motion processing. *Journal of Neuroscience*, *29*(20), 6490–6499. http://doi.org/10.1523/JNEUROSCI.5437-08.2009

Sankaran, N., Thompson, W. F., Carlile, S., & Carlson, T. A. (2018). Decoding the dynamic representation of musical pitch from human brain activity. *Scientific Reports*, *8*, 839. http://doi.org/10.1038/s41598-018-19222-3

Santangelo, V. (2018). Large-scale brain networks supporting divided attention across spatial locations and sensory modalities. *Frontiers in Integrative Neuroscience*, *12*, 1–11. http://doi.org/10.3389/fnint.2018.00008

Santangelo, V., Fagioli, S., & Macaluso, E. (2010). The costs of monitoring simultaneously two sensory modalities decrease when dividing attention in space. *NeuroImage*, *49*(3), 2717–2727. http://doi.org/10.1016/j.neuroimage.2009.10.061

Santangelo, V., Ho, C., & Spence, C. (2008). Capturing spatial attention with multisensory cues. *Psychonomic Bulletin & Review*, *15*(2), 398–403. http://doi.org/10.3758/PBR.15.2.398

Santangelo, V., Olivetti Belardinelli, M., Spence, C., & Macaluso, E. (2009). Interactions between voluntary and stimulus-driven spatial attention mechanisms across sensory modalities. *Journal of Cognitive Neuroscience*, *21*, 2384–2397. http://doi.org/10.1162/jocn.2008.21178

Santangelo, V., & Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1311–1321. http://doi.org/10.1037/0096-1523.33.6.1311

Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., & Formisano, E. (2017). Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proceedings of the National Academy of Sciences*, *114*(18), 4799–4804. http://doi.org/10.1073/pnas.1617622114

Schroeder, C. E., & Foxe, J. J. (2002). The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Cognitive Brain Research*, *14*(1), 187–198. http://doi.org/10.1016/S0926-6410(02)00073-3

Schroeder, C. E., & Foxe, J. J. (2005). Multisensory contributions to low-level, “unisensory” processing. *Current Opinion in Neurobiology*, *15*(4), 454–458. http://doi.org/10.1016/j.conb.2005.06.008

Schroeder, C. E., Lindsley, R. W., Specht, C., Marcovici, A., Smiley, J. F., & Javitt, D. C. (2001). Somatosensory input to auditory association cortex in the macaque monkey. *Journal of Neurophysiology*, *85*(3), 1322–1327.

Schroeder, C. E., Smiley, J., Fu, K. G., McGinnis, T., O’Connell, M. N., & Hackett, T. A.

(2003). Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. *International Journal of Psychophysiology*, *50*(1-2), 5–17. http://doi.org/10.1016/S0167-8760(03)00120-X

Schürmann, M., Caetano, G., Hlushchuk, Y., Jousmäki, V., & Hari, R. (2006). Touch activates human auditory cortex. *NeuroImage*, *30*(4), 1325–1331. http://doi.org/10.1016/j.neuroimage.2005.11.020

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. http://doi.org/10.1214/aos/1176344136

Senkowski, D., Schneider, T. R., Foxe, J. J., & Engel, A. K. (2008). Crossmodal binding through neural coherence: implications for multisensory processing. *Trends in Neurosciences*, *31*(8), 401–409. http://doi.org/10.1016/j.tins.2008.05.002

Serences, J. T., & Kastner, S. (2014). A multi-level account of selective attention. In *The Oxford Handbook of Attention* (pp. 76–104). Oxford: Oxford University Press.

Serences, J. T., Schwarzbach, J., Courtney, S. M., Golay, X., & Yantis, S. (2004). Control of object-based attention in human cortex. *Cerebral Cortex*, *14*(12), 1346–1357. http://doi.org/10.1093/cercor/bhh095

Serences, J. T., & Yantis, S. (2006). Selective visual attention and perceptual coherence. *Trends in Cognitive Sciences*, *10*(1), 38–45. http://doi.org/10.1016/j.tics.2005.11.008

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123. http://doi.org/10.1016/j.tins.2010.11.002

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences*, *14*(9), 425–432. http://doi.org/10.1016/j.tics.2010.07.001

Shapleske, J., Rossell, S. ., Woodruff, P. W. ., & David, A. . (1999). The planum temporale: a systematic, quantitative review of its structural, functional and clinical significance. *Brain Research Reviews*, *29*(1), 26–49. http://doi.org/10.1016/S0165-0173(98)00047-2

Shomstein, S., & Behrmann, M. (2006). Cortical systems mediating visual attention to both objects and spatial locations. *Proceedings of the National Academy of Sciences*, *103*(30), 11387–11392. http://doi.org/10.1073/pnas.0601813103

Shomstein, S., & Yantis, S. (2004). Control of attention shifts between vision and audition in human cortex. *Journal of Neuroscience*, *24*(47), 10702–10706. http://doi.org/10.1523/JNEUROSCI.2939-04.2004

Shomstein, S., & Yantis, S. (2006). Parietal cortex mediates voluntary control of spatial and nonspatial auditory attention. *Journal of Neuroscience*, *26*(2), 435–9. http://doi.org/10.1523/JNEUROSCI.4408-05.2006

Slutsky, D. a, & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport*, *12*(1), 7–10. http://doi.org/10.1097/00001756-200101220-00009

Smiley, J. F., Hackett, T. A., Ulbert, I., Karmas, G., Lakatos, P., Javitt, D. C., & Schroeder, C. E. (2007). Multisensory convergence in auditory cortex, I. Cortical connections of the caudal superior temporal plane in macaque monkeys. *The Journal of Comparative Neurology*, *502*(6), 894–923. http://doi.org/10.1002/cne.21325

Soto-Faraco, S., & Deco, G. (2009). Multisensory contributions to the perception of vibrotactile events. *Behavioural Brain Research*, *196*(2), 145–154. http://doi.org/10.1016/j.bbr.2008.09.018

Soto-Faraco, S., Kvasova, D., Biau, E., Ikumi, N., Ruzzoli, M., Morís-Fernández, L., & Torralba, M. (2019). Multisensory interactions in the real world. In *Elements in perception*. Cambridge, UK: Cambridge University Press. http://doi.org/10.1017/9781108578738

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995. http://doi.org/10.3758/s13414-010-0073-7

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. *Annals of the New York Academy of Sciences*, *1296*(1), 31–

49. http://doi.org/10.1111/nyas.12121

Spence, C., & Driver, J. (1997). On measuring selective attention to an expected sensory modality. *Perception & Psychophysics*, *59*(3), 389–403. http://doi.org/10.3758/BF03211906

Spence, C., Nicholls, M. E., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics*, *63*(2), 330–336. http://doi.org/10.3758/BF03194473

Sprague, T. C., Itthipuripat, S., Vo, V. A., & Serences, J. T. (2018). Dissociable signatures of visual salience and behavioral relevance across attentional priority maps in human cortex. *Journal of Neurophysiology*, *119*(6), 2153–2165. http://doi.org/10.1152/jn.00059.2018

Stanford, T. R., & Stein, B. E. (2007). Superadditivity in multisensory integration: Putting the computation in context. *NeuroReport*, *18*(8), 787–792. http://doi.org/10.1097/WNR.0b013e3280c1e315

Stanley, B. M., Chen, Y. C., Lewis, T. L., Maurer, D., & Shore, D. I. (2019). Developmental changes in the perception of audiotactile simultaneity. *Journal of Experimental Child Psychology*, *183*, 208–221. http://doi.org/10.1016/j.jecp.2019.02.006

Stein, B. E. (Ed.). (2012). *The new handbook of multisensory processing*. Cambridge, MA, USA: MIT Press.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, *9*(4), 255–266. http://doi.org/10.1038/nrn2331

Stein, B. E., Stanford, T. R., Ramachandran, R., Perrault, T. J., & Rowland, B. A. (2009). Challenges in quantifying multisensory integration: alternative criteria, models, and inverse effectiveness. *Experimental Brain Research*, *198*(2-3), 113–126. http://doi.org/10.1007/s00221-009-1880-8

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. http://doi.org/10.1016/j.neuroimage.2009.03.025

Stoll, J., Thrun, M., Nuthmann, A., & Einhäuser, W. (2015). Overt attention in natural scenes: Objects dominate features. *Vision Research*, *107*, 36–48. http://doi.org/10.1016/j.visres.2014.11.006

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. http://doi.org/10.1121/1.1907309

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. *Science*, *314*(5803), 1311–1314. http://doi.org/10.1126/science.1132028

Swisher, J. D., Halko, M. A., Merabet, L. B., McMains, S. A., & Somers, D. C. (2007). Visual topography of human intraparietal sulcus. *Journal of Neuroscience*, *27*(20), 5326–5337. http://doi.org/10.1523/JNEUROSCI.0991-07.2007

Talsma, D. (2015). Predictive coding and multisensory integration: an attentional account of the multisensory mind. *Frontiers in Integrative Neuroscience*, *9*(19), 1–13. http://doi.org/10.3389/fnint.2015.00019

Talsma, D., Doty, T. J., & Woldorff, M. G. (2007). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cerebral Cortex*, *17*(3), 679–690. http://doi.org/10.1093/cercor/bhk016

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, *14*(9), 400–410. http://doi.org/10.1016/j.tics.2010.06.008

Talsma, D., & Woldorff, M. G. (2005). Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience*, *17*(7), 1098–114. http://doi.org/10.1162/0898929054475172

Tang, X., Wu, J., & Shen, Y. (2016). The interactions of multisensory integration with endogenous and exogenous attention. *Neuroscience and Biobehavioral Reviews*, *61*, 208–224. http://doi.org/10.1016/j.neubiorev.2015.11.002

Theeuwes, J. (2018). Visual selection: usually fast and automatic; seldom slow and volitional. *Journal of Cognition*, *1*(1), 1–15. http://doi.org/10.5334/joc.13

Theeuwes, J., & Chen, C. Y. D. (2005). Attentional capture and inhibition (of return): The effect on perceptual sensitivity. *Perception and Psychophysics*, *67*(8), 1305–1312. http://doi.org/10.3758/BF03193636

Theeuwes, J., Kramer, A. F., & Kingstone, A. (2004). Attentional capture modulates perceptual sensitivity. *Psychonomic Bulletin and Review*, *11*(3), 551–554. http://doi.org/10.3758/BF03196609

Theeuwes, J., & van der Burg, E. (2011). On the limits of top-down control of visual selection. *Attention, Perception, and Psychophysics*, *73*(7), 2092–2103. http://doi.org/10.3758/s13414-011-0176-9

Tillmann, B. (2012). Music and Language Perception: Expectations, Structural Integration, and Cognitive Sequencing. *Topics in Cognitive Science*, *4*(4), 568–584. http://doi.org/10.1111/j.1756-8765.2012.01209.x

Toiviainen, P., Alluri, V., Brattico, E., Wallentin, M., & Vuust, P. (2014). Capturing the musical brain with Lasso: Dynamic decoding of musical features from fMRI data. *NeuroImage*, *88*, 170–180. http://doi.org/10.1016/j.neuroimage.2013.11.017

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786. http://doi.org/10.1037/0033-295X.113.4.766

Tranchant, P., Shiell, M. M., Giordano, M., Nadeau, A., Peretz, I., & Zatorre, R. J. (2017). Feeling the beat: Bouncing synchronization to vibrotactile music in hearing and early deaf people. *Frontiers in Neuroscience*, *11*, 1–8. http://doi.org/10.3389/fnins.2017.00507

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. http://doi.org/10.1016/0010-0285(80)90005-5

Van Bergen, R. S., Ji Ma, W., Pratte, M. S., & Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, *18*(12), 1728–1730. http://doi.org/10.1038/nn.4150

van de Rijt, L. P. H., Roye, A., Mylanus, E. A. M., van Opstal, A. J., & van Wanrooij, M. M. (2019). The Principle of Inverse Effectiveness in Audiovisual Speech Perception. *Frontiers in Human Neuroscience*, *13*, 1–15. http://doi.org/10.3389/fnhum.2019.00335

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and Pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1053–1065. http://doi.org/10.1037/0096-1523.34.5.1053

Van der Burg, E., Talsma, D., Olivers, C. N. L., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *NeuroImage*, *55*(3), 1208–1218. http://doi.org/10.1016/j.neuroimage.2010.12.068

Van der Stoep, N., Van der Stigchel, S., & Nijboer, T. C. W. (2015). Exogenous spatial attention decreases audiovisual integration. *Attention, Perception, and Psychophysics*, *77*(2), 464–482. http://doi.org/10.3758/s13414-014-0785-1

van Ee, R., van Boxtel, J. J. A., Parker, A. L., & Alais, D. (2009). Multisensory congruency as a mechanism for attentional control over perceptual selection. *Journal of Neuroscience*, *29*(37), 11641–11649. http://doi.org/10.1523/JNEUROSCI.0873-09.2009

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. http://doi.org/10.1016/j.neuropsychologia.2006.01.001

Vatakis, A., & Spence, C. (2007). Crossmodal binding: evaluating the “unity assumption” using audiovisual speech stimuli. *Perception & Psychophysics*, *69*(5), 744–756. http://doi.org/10.3758/BF03193776

Vercillo, T., & Gori, M. (2015). Attention to sound improves auditory reliability in audio-tactile spatial optimal integration. *Frontiers in Integrative Neuroscience*, *9*, 34. http://doi.org/10.3389/fnint.2015.00034

Vroomen, J., Bertelson, P., & de Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics*, *63*(4), 651–659. http://doi.org/10.3758/BF03194427

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. http://doi.org/10.1093/cercor/bhu277

Watanabe, M. (1992). Frontal units of the monkey coding the associative significance of visual and auditory stimuli. *Experimental Brain Research*, *89*(2), 233–247. http://doi.org/10.1007/BF00228241

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, *88*(3), 638–667. http://doi.org/10.1037/0033-2909.88.3.638

Werner, S., & Noppeney, U. (2010). Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *Journal of Neuroscience*, *30*(7), 2662–2675. http://doi.org/10.1523/JNEUROSCI.5091-09.2010

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford University Press, USA.

Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*(2), 77–84. http://doi.org/10.1016/j.tics.2010.12.001

Yeshurun, Y., Kimchi, R., Sha'shoua, G., & Carmel, T. (2009). Perceptual objects capture attention. *Vision Research*, *49*(10), 1329–1335. http://doi.org/10.1016/j.visres.2008.01.014

Zangenehpour, S., & Zatorre, R. J. (2010). Crossmodal recruitment of primary visual cortex

following brief exposure to bimodal audiovisual stimuli. *Neuropsychologia*, *48*(2), 591–600. http://doi.org/10.1016/j.neuropsychologia.2009.10.022

Zimmer, U., Itthipanyanan, S., Grent-'t-Jong, T., & Woldorff, M. G. (2010a). The electrophysiological time course of the interaction of stimulus conflict and the multisensory spread of attention. *European Journal of Neuroscience*, *31*(10), 1744–1754. http://doi.org/10.1111/j.1460-9568.2010.07229.x

Zimmer, U., Roberts, K. C., Harshbarger, T. B., & Woldorff, M. G. (2010b). Multisensory conflict modulates the spread of visual attention across a multisensory object. *NeuroImage*, *52*(2), 606–616. http://doi.org/10.1016/j.neuroimage.2010.04.245

Zion Golumbic, E. M., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party." *Journal of Neuroscience*, *33*(4), 1417–1426. http://doi.org/10.1523/JNEUROSCI.3675-12.2013

Zuanazzi, A., & Noppeney, U. (2018). Additive and interactive effects of spatial attention and expectation on perceptual decisions. *Scientific Reports*, *8*(1), 6732. http://doi.org/10.1038/s41598-018-24703-6