The Facilitatory Crossmodal Effect of Auditory Stimuli on Visual Perception

Yi-Chuan Chen

The Queen's College, Oxford

Dissertation submitted for the degree of

Doctor of Philosophy in the University of Oxford

Trinity Term, 2011

Declaration

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

Yi-Chuan Chen

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Charles Spence, for his courage to take me on as his student and every challenge (or torture) he presented me with. I also appreciate his patient guidance and support, as well as his kind tolerance for every time I argued with him. He makes me believe I can be a scientific researcher in Psychology, and a good wonton-chef.

I would like to thank my colleagues in the lab: Mary (and the half colleague – John), Anne-Sylvie, Georgiana, Camille, Jess, Cesare, Carmel, and Vanessa who just joined in recently. I enjoy working with you at the office where we have the beautiful view on South Parks Road.

I am grateful to my friends in Oxford: Bo-Cheng, Yu-Ling, Bonnie, Simpson, Julie and Tim, and my new cooking mate – Esther. Thanks for your warming friendship and patience to listen to all my complaints and gossip. I enjoyed every night we shared meals together that totally comforted my stomach and soul.

I would like to thank my previous supervisor, Prof. Su-Ling Yeh, for her continous concern and support. I am also grateful to my friends in Taiwan: Li-Chuan, Jing-Ling, Shin-Pey, Shih-chi, Pi-Chun, Chung-Hsien, Thomas, Shao-Wei, and Cheng-Ta. I also treasure the warmest cherish from my best friends “wu-siao-fu” no matter where we are.

I would like to thank the Ministry of Education in Taiwan for providing the funds which enable me to work on this dissertation.

Finally, I deeply acknowledge my parents and my sisters for their invaluable support. I would like to dedicate this thesis to my grandmother who always makes me feel that I am blessed.

Publications and Presentations

Research from this thesis has been, or will be, published/submitted:

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

Chen, Y.-C., & Spence, C. (in press). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity *Journal of Experimental Psychology: Human Perception & Performance.*

Chen, Y.-C., & Spence, C. (resubmitted). The crossmodal facilitation of visual object representations by sound: Evidence from the backward masking paradigm. *Journal of Experimental Psychology: Human Perception & Performance.*

Chen, Y.-C., Yeh, S.-L., & Spence, C. (resubmitted). Crossmodal constraints on human visual awareness: Can auditory semantic context modulate binocular rivalry? *Journal of Experimental Psychology: Human Perception & Performance.*

Chen, Y.-C., Huang, P.-C., Yeh, S.-L., & Spence, C. (resubmitted). The presentation of a synchronous sound enhances visual sensitivity without reducing target uncertainty. *Seeing and Perceiving.*

Chen, Y.-C., & Spence, C. (submitted). Naturalistic sounds and spoken words modulate the categorization of visually-presented pictures at multiple levels of human information processing. *Acta Psychologica.*

Other relevant publications:

Hsiao, J.-Y., Chen, Y.-C., Spence, C., & Yeh, S.-L. (submitted). Assessing the effects of audiovisual semantic congruency and the interplay between covert and overt attention on the perception of a bistable figure. *Consciousness and Cognition.*

Spence, C., & Chen, Y.-C. (in press). Intramodal and crossmodal perceptual grouping. In B. E. Stein (Ed.), *The new handbook of multisensory processing.* Cambridge, MA: MIT Press.

Spence, C., Parise, C., & Chen, Y.-C. (in press). The Colavita visual dominance effect. In M. M. Murray & M. Wallace (Eds.), *Frontiers in the neural bases of multisensory processes.*

Conference presentations:

Chen, Y.-C., & Spence, C. (2010). Enhancing the detectability of a picture requires the prior presentation of a semantically-congruent sound. Poster presented at *The 11th International Multisensory Research Forum*, Liverpool, U.K. 16-19, June.

Chen, Y.-C., Yeh, S. L., & Spence, C. (2010). Crossmodal constraints on human visual awareness: Auditory semantic context modulates binocular rivalry. *Journal of Vision, 10(7),* 885. Poster presented at *Vision Science Society*, 10th Annual Meeting. Naples, FL., U.S.A. 7-12, May.

Hsiao, J.-Y., Chen, Y.-C., Spence, C. & Yeh, S.-L. (2010). Semantic congruency, attention, and fixation position modulate conscious perception when viewing a bistable figure. *Journal of Vision, 10(7),* 867. Poster presented at *Vision Science Society*, 10th Annual Meeting. Naples, FL., U.S.A. 7-12, May.

Chen, Y.-C., & Spence, C. (2009). Crossmodal facilitation of visual target identification at the level of object representation by the presentation of a concomitant sound. *Perception, 38 (ECVP Abstract Supplement),* 145. Poster presented at The 32th *European Conference on Visual Perception*. Regensburg, Germany. 24-28, August.

Chen, Y.-C., & Spence, C. (2009). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. Poster presented at *The 10th International Multisensory Research Forum*, New York, U.S.A. 29, June-2, July.

Spence, C., Parise, C., & Chen, Y.-C. (2009). Explaining the Colavita visual dominance effect. Talk presented at *The 10th International Multisensory Research Forum*, New York, U.S.A. 29, June-2, July.

Table of Contents

Table of Contents **1**

Short Abstract **7**

Extended Abstract **8**

Chapter 1

Introduction **17**

1.1 Why does audiovisual interaction occur? 19

1.2 When (in which conditions) do audiovisual interactions occur? 22

1.2.1 Structural factors: Spatial and temporal coincidence 23

1.2.2 Cognitive factor: Semantic congruency 25

1.2.3 The roles of spatial, temporal, and semantic factors in audiovisual interactions 27

1.3 What are the possible outcomes of audiovisual interactions? 29

1.3.1 Perceptual enhancement by redundant multisensory signals 30

1.3.2 Perceptual dominance by nonredundant multisensory signals 32

1.3.3 Perceptual modulation by nonredundant multisensory signals 34

1.3.4 Perceptual emergence by nonredundant multisensory signals 40

1.3.5 The inverse effectiveness rule in neurophysiological and behavioural studies 42

1.4 How do visual and auditory information interact? 46

1.4.1 Modality appropriateness, modality preciseness, and the directed-attention hypothesis 46

1.4.2 Unitary or multiple mechanisms of multisensory interactions 52

1.4.3 Direct and indirect audiovisual interactions 54

1.5 Summary 59

Chapter 2

Aims and Scope **61**

2.1 Temporal coincidence in audiovisual interactions 61

2.1.1 Is temporal coincidence a sufficient condition for crossmodal facilitatory effects to occur? 62

2.1.2 How does the presentation of a simultaneous sound enhance visual perception? 65

2.1.3 Outline of the studies on temporal coincidence 69

2.2 Does semantic congruency influence audiovisual interactions? 71

2.2.1 When do audiovisual semantic interactions occur? 71
2.2.2 How do pictures and meaningful sounds interact? 73
2.2.3 Outline of the studies on audiovisual semantic congruency 76

Chapter 3

The presentation of a synchronous sound enhances visual sensitivity without reducing target uncertainty 78
3.1 Introduction 78
3.2 Methods 82
3.2.1 Participants 82
3.2.2 Apparatus and stimuli 83
3.2.3 Design and procedure 84
3.3 Results 87
3.3.1 Results of the grouped data 88
3.3.2 The psychometric function at the intermediate (-22 dB) noise level 90
3.4 Discussion 93

Chapter 4

The crossmodal facilitation of visual object representations by sound: Evidence from the backward masking paradigm 99
4.1 Introduction 99
4.2 Experiment 4.1 108
4.2.1 Methods 109
4.2.2 Results 113
4.2.3 Discussion 117
4.3 Experiment 4.2 120
4.3.1 Methods 121
4.3.2 Results and discussion 121
4.4 Experiment 4.3 126
4.4.1 Methods 126
4.4.2 Results and discussion 127
4.5 Experiment 4.4A 131
4.5.1 Methods 131
4.5.2 Results and discussion 132
4.6 Experiment 4.4B 134
4.6.1 Methods 134
4.6.2 Results and discussion 135
4.7 Experiment 4.5 137
4.7.1 Methods 138

4.7.2 Results and discussion 139
4.8 General discussion 142
4.8.1 Bottom-up and top-down factors modulating audiovisual integration 144
4.8.2 How does the presentation of the sound enhance visual perception?.. 146
4.8.3 Audiovisual integration at the level of object representation 150
4.8.4 Linking behavioral and physiological evidence 154
4.9 Conclusions 157

Chapter 5

Crossmodal constraints on human visual awareness: Can auditory semantic context modulate binocular rivalry? 158
5.1 Introduction 158
5.2 Experiment 5.1 161
5.2.1 Methods 162
5.2.2 Results and discussion 165
5.3 Experiment 5.2 170
5.3.1 Methods 170
5.3.2 Results and discussion 171
5.4 Experiment 5.3 177
5.4.1 Methods 178
5.4.2 Results and discussion 179
5.5 Experiment 5.4 181
5.5.1 Methods 182
5.5.2 Results and discussion 184
5.6 General discussion 188
5.6.1 Possible mechanisms underlying the modulation of visual awareness by auditory semantic context 189
5.6.2 Human perceptual awareness: Unisensory or multisensory? 195
5.7 Conclusions 196

Chapter 6

Naturalistic sounds and spoken words modulate the categorization of visually-presented pictures at multiple levels of human information processing. 198
6.1 Introduction 198
6.2 Methods 202
6.2.1 Participants 202
6.2.2 Apparatus and stimuli 202
6.2.3 Design 205

6.2.4 Procedure 206
6.3 Results 207
6.3.1 Analysis of the one-syllable spoken words 212
6.4 Discussion 214
6.4.1 Audiovisual semantic interactions at both perceptual and decisional stages 215
6.4.2 Differences between naturalistic sounds and spoken words 217

Chapter 7

Enhancing visual sensitivity by means of crossmodal semantic priming 220
7.1 Introduction 220
7.2 Experiment 7.1 226
7.2.1 Methods 226
7.2.2 Results and discussion 230
7.3 Experiment 7.2 233
7.3.1 Experiment 7.2A 234
7.3.2 Experiment 7.2B 236
7.3.3 Discussion 237
7.4 Experiment 7.3 238
7.4.1 Experiment 7.3A 239
7.4.2 Experiment 7.3B 241
7.4.3 Discussion 242
7.5 Experiment 7.4 245
7.5.1 Experiment 7.4A 245
7.5.2 Experiment 7.4B 248
7.5.3 Discussion 252
7.6 General discussion 253
7.6.1 The timecourse of the auditory modulation of visual perception 254
7.6.2 The effect of naturalistic sounds vs. spoken words on the priming of visual pictures 257
7.6.3 Human brain networks for auditory processing and audiovisual semantic interactions 262
7.7 Conclusions 263

Chapter 8

When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures 265
8.1 Introduction 265
8.1.2 Previous research on audiovisual semantic congruency effects 268

8.2 Experiment 8.1 275
8.2.1 Methods 275
8.2.2 Results 280
8.2.3 Discussion 283
8.3 Experiment 8.2 286
8.3.1 Methods 286
8.3.2 Results 287
8.3.3 Discussion 288
8.4 Experiment 8.3 290
8.4.1 Methods 290
8.4.2 Results and discussion 291
8.5 Experiment 8.4 295
8.5.1 Methods 295
8.5.2 Results and discussion 295
8.6 Experiment 8.5 297
8.6.1 Methods 298
8.6.2 Results 299
8.6.3 Discussion 302
8.7 General discussion 303
8.7.1 A possible framework for multisensory integration 309
8.8 Conclusions 312

Chapter 9

General discussion 313
9.1 The role of temporal coincidence in audiovisual interactions 317
9.1.1 Temporal coincidence and the unity assumption 318
9.1.2 The role of spatial coincidence in the temporal coincidence situation 319
9.1.3 The mechanism(s) underlying crossmodal facilitation following the presentation of a temporally coincident sound 320
9.2 The role of semantic congruency in audiovisual interactions 323
9.2.1 Audiovisual semantic interactions are asymmetrical 323
9.2.2 Audiovisual semantic interactions and Bayesian priors 324
9.2.3 The mechanism(s) underlying crossmodal facilitation by semantic congruency 325
9.3 The type-token binding framework 329
9.3.1 Separable type and token processing in audiovisual interactions 332
9.3.2 The role of attention in multisensory type-token binding 333
9.3.3 Binding type and token information to construct a multisensory

object representation 337
9.4 Directions for future research 339
9.5 Conclusions 341

References 343

Short Abstract

The aim of the experiments reported in this thesis was to investigate the multisensory interactions taking place between vision and audition. The focus is on the modulatory role of the temporal coincidence and semantic congruency of pairs of auditory and visual stimuli. With regards to the temporal coincidence factor, whether, and how, the presentation of a simultaneous sound facilitates visual target perception was tested using the equivalent noise paradigm (Chapter 3) and the backward masking paradigm (Chapter 4). The results demonstrate that crossmodal facilitation can be observed in both visual detection and identification tasks. Importantly, however, the results also reveal that the sound not only had to be presented simultaneously, but also reliably, with the visual target. The suggestion is made that the reliable co-occurrence of the auditory and visual stimuli provides observers with the statistical regularity needed to assume that the visual and auditory stimuli likely originate from the same perceptual event (i.e., that they in some sense 'belong together'). The experiments reported in Chapters 5 through 8 were designed to investigate the role of semantic congruency on audiovisual interactions. The results of the experiments reported in Chapter 5 revealed that the semantic context provided by the soundtrack that a person happens to be listening to can modulate his/her visual conscious perception in the binocular rivalry situation. In Chapters 6-8, the timecourse of audiovisual semantic interactions were investigated using categorization, detection, and identification tasks on visual pictures. The results suggested that when the presentation of the sound leads the presentation of a picture by more than 240 ms, it induces a crossmodal semantic priming effect. In addition, when the presentation of the sound lags a semantically-congruent picture by about 300 ms, it enhances performance, presumably by helping to maintain the visual representation in short-term memory. The results indicate that audiovisual semantic interactions constitute a heterogeneous group of phenomena. A crossmodal type-token binding framework is proposed to account for the parallel processing of the spatiotemporal and semantic interactions of multisensory inputs. The suggestion is that the congruent information in the type and token representation systems would integrate, and they finally bind into a unified multisensory object representation.

Extended Abstract

Human sensory and perceptual systems serve a critical function in acquiring information from the outside world and representing it as veridically as possible in order for us to understand the continuously changing environments in which we live. A perceptual event, in most cases, delivers information to multiple sensory modalities simultaneously. One problem faced by human perceptual systems is to determine which inputs from different sensory modalities belong to the same event (and hence can be integrated) while others are not (so they should be kept segregated). This multisensory interaction process is modulated not only by the nature of the sensory inputs, but also by some heuristics or forms of knowledge that individuals have in order to help them to infer the original stimulus. Ideally, multisensory interactions should provide the most accurate representation corresponding to the likely external stimulation, such that an observer can generate adequate behavioural responses.

In the majority of natural situations, vision and audition constitute the two most informative modalities for humans. The property of each sensory stimulus can be summarized in terms of three key aspects: the *space* and the *time* that the stimulus occurs, and the *identity* (or semantic) of that stimulus. The congruency between visual and auditory stimuli in these three areas therefore constitutes possible modulatory factors for audiovisual interactions. In the literature, researchers often demonstrate the

modulation of vision by audition in terms of its temporal properties, while less frequently showing such effect in terms of its spatial properties or identity. Regarding the mechanisms underlying these audiovisual interaction phenomena, the perceptual outcome may be determined according to the source modality of the stimulus, by the preciseness of the information of each sensory input, or by the observer's attentional modulation. All of the factors that modulate audiovisual interactions may be summed up to in a unitary overarching mechanism, or may be separately computed by several individual mechanisms concerning space, time, or identity. And finally, audiovisual interactions may occur via direct connections, during feedforward processing, or may be modulated by feedback control.

The aims of the experiments reported in this thesis were to provide further evidence regarding how a putatively task-irrelevant auditory stimulus modulate human performance on visual target, and to examine, or propose, possible underlying mechanisms in order to account for the results reported here. The influence of two main factors, temporal coincidence and the semantic congruency between the visual target and the accessory sound are investigated in some detail.

In Chapter 3, the visual detection and discrimination sensitivity of participants was estimated using psychophysical methods under the conditions in which a simultaneous sound was either presented or else was not. In each trial, a tilted Gabor patch was presented in either the first or second of two sequences which constitutes random dot

noise with one of seven possible contrast levels. A dual two-alternative forced-choice task was adopted in which the participant had to report which of the two sequences contained the Gabor and in which orientation was it tilted. The results revealed that the sensitivity of participant' visual detection and discrimination performance could both be enhanced by the presentation of a simultaneous sound, though only when the noise contrast was at an intermediate level. A further analysis of the psychometric function at the intermediate noise level revealed that the increase in sensitivity could not be attributed to any reduction of participants' uncertainty regarding the onset time of the visual target. Instead, the crossmodal facilitatory effects may be accounted for by the idea that the visual signal is enhanced, as well as by the visual noise being reduced, by the presentation of a sound.

In Chapter 4, a series of experiments was designed to examine underlying mechanisms of the crossmodal facilitation elicited by the presentation of a sound on the identification of a concomitantly-presented visual target letter. The well-established backward masking paradigm was utilized: Two visual letters, serving as the target and its mask, were presented successively at various interstimulus intervals (ISIs). The results demonstrated that the crossmodal facilitation of participants' identification performance elicited by a sound which was presented simultaneously with the visual target occurred over only a very narrow range of ISIs. This critical time-window was just beyond the interval needed for participants to differentiate the target and mask as

constituting two distinct events (Experiment 4.1) and was dissociated from any facilitation elicited by making the visual target physically brighter (Experiment 4.2). When the sound was presented at the same time as the mask, a facilitatory, rather than an inhibitory, effect on participants' visual target identification performance was still observed (Experiment 4.3). Further experiments demonstrated that the crossmodal facilitation of the visual target by the sound depends on the sound and visual target co-occurring reliably (Experiment 4.4); however, by contrast, spatial coincidence was not necessary in order to demonstrate this effect (Experiment 4.5). These results suggest that when visual and auditory stimuli are reliably presented in synchrony, they will likely result in the construction of a better-consolidated object representation (than that resulting from unimodal visual stimulation). This result is explained in terms of a crossmodal type-token binding framework (e.g., Bowman & Wyble, 2007).

In Chapter 5, a study designed to investigate whether crossmodal semantic context modulates visual awareness by utilizing the binocular rivalry paradigm is reported. Binocular rivalry refers to the phenomenon whereby when different figures are presented to the corresponding location in each eye, observers may perceive each figure as being dominant in alternation over time. The results revealed that participants reported a particular percept as being dominant for more of the time when listening to a semantically-congruent auditory soundtrack, while no such pattern of results was observed when no sound, or else when a semantically-irrelevant soundtrack, was

presented instead (Experiment 5.1). Furthermore, the crossmodal modulation of binocular rivalry by auditory semantic context can be dissociated from two visual factors: Participants' attentional control over the dichoptic figures and the relative luminance contrast between the figures (see Experiments 5.2 & 5.3). Critically, the crossmodal modulation of visual awareness by auditory semantic context was still observed under the conditions where the meaning of the soundtrack cannot directly bias participants' reporting of the dominance of a given percept (Experiment 5.4). Hence, the crossmodal semantic effects reported in Chapter 5 cannot simply be attributed to the meaning of the soundtrack guiding participants' attention, or biasing their responses. Instead, the auditory semantic cues themselves can enhance the likelihood that any semantically-congruent stimuli will win the competition for awareness under conditions of binocular rivalry. The results therefore suggest that crossmodal semantic congruency can serve as a constraint on resolving perceptual ambiguity in the visual system. In turn, multisensory stimulation provides a novel means of probing the contextual constraints on human visual awareness.

In Chapter 6, crossmodal semantic interactions between visually-presented pictures and both naturalistic sounds (e.g., a dog's "woofing") and spoken words (e.g., /dɔg/) were compared using a picture categorization task (Experiment 6.1). The results demonstrated that both types of auditory stimulus induced semantic congruency effects (i.e., crossmodal facilitation by semantically-congruent sounds and inhibition by

semantically-incongruent sounds, relative to a baseline noise condition) when the onset of the sound led that of the picture by 240 ms or more. Both types of sound also gave rise to inhibition irrespective of their semantic congruency when they were presented within approximately 100 ms of the onset of picture, with the peak of inhibition occurring somewhat earlier for spoken words than for naturalistic sounds. These results are taken to suggest that audiovisual semantic interactions occur at multiple levels of human information processing, including semantic access and decision making. The different timecourses of late inhibition observed for naturalistic sounds and spoken words are attributed to their different processing pathways to access semantic representations.

In Chapter 7, four experiments designed to investigate whether the semantic congruency of a meaningful sound can enhance visual sensitivity are reported. Two main issues are addressed in this chapter: First, at which stimulus onset asynchronies (SOAs) can the auditory stimulus effectively prime visual sensitivity (rather than merely triggering a visual response)? And, second, do naturalistic sounds and spoken words elicit similar semantic priming effects? Here, the participants' sensitivity and response criterion were estimated using signal detection theory in a picture detection task. The results demonstrated that the onset of the naturalistic sound had to lead the onset of the picture by 346 ms in order for semantic priming to enhance participants' visual sensitivity (Experiments 7.1-7.3A); at the same SOA, however, spoken words did not

induce semantic priming effects on visual sensitivity in the picture detection task (Experiments 7.3B & 7.4A). When using a dual picture detection/identification task, both kinds of auditory stimulus induced a similar magnitude of semantic priming effect. These results therefore suggest that the semantic priming from audition to vision relies on: 1) There being sufficient processing time for the auditory stimulus to access its associated meaning; and 2) The representations of the visual picture that are required for participants to respond should be enhanced by the processing of earlier-presented sound. A multisensory framework based on Glaser and Glaser's (1989) general reading-naming interference model is proposed in order to account for the processing of naturalistic sounds and spoken words and how they interact with visual picture processing.

In Chapter 8, a series of experiments designed to assess the effect of a simultaneously- or subsequently-presented sound on the identification of visually-presented pictures in terms of their semantic congruency is reported. Participants made unspeeded identification responses concerning a series of briefly-presented, and then rapidly-masked, pictures. A naturalistic sound was sometimes presented together with the picture at an SOA that varied between 0 and 533 ms (auditory lagging). The sound could be semantically congruent, semantically incongruent, or else neutral (white noise) with respect to the target picture. The results of these experiments revealed that when the onset of the picture and sound occurred simultaneously, a semantically congruent sound improved, whereas a semantically

incongruent sound impaired, participants' picture identification performance, as compared to performance in the white-noise control condition. A significant crossmodal facilitatory effect was also observed at SOAs of around 300 ms, whereas no such semantic congruency effects were observed at the longest interval (533 ms). These results therefore suggest that the representations associated with visual and auditory stimuli can interact in a shared semantic system. Furthermore, this crossmodal semantic interaction is not constrained by the need for the strict temporal coincidence of the constituent auditory and visual stimuli. The results reported in Chapter 8 therefore suggest that audiovisual semantic interactions likely occur in a short-term buffer which rapidly accesses, and temporarily retains, the semantic representations of multisensory stimuli in order to form coherent multisensory object representations. These results are explained in terms of Potter's (1993) notion of conceptual short-term memory.

In summary, the audiovisual interactions in the studies reported in this thesis occurred in terms of temporal coincidence of stimuli (Chapters 3 & 4) and/or in terms of their semantic congruency (Chapters 5-8). A crossmodal type-token binding framework is proposed in order to provide a comprehensive model to account for these results: The three basic properties of a multisensory perceptual event (space, time, and identity) can be categorized into spatiotemporal factors ('when' and 'where', or 'token') versus identity factors ('what' or 'type'). Audiovisual interactions in terms of stimulus onset time likely occur when constructing token representations (e.g., one or two tokens

should open, corresponding to a multisensory event or two unimodal events), while those in terms of semantic congruency likely occur in the type activation network (such as the crossmodal general reading-naming interference model proposed in Chapter 7). The type and token information may finally be integrated in order to construct a complete object representation, and this binding process likely needs the aid of attention. The type-token binding process provides a framework with which to conceptualize how internal representations are constructed to infer the outside world.

Chapter 1

Introduction

In everyday life, the majority of perceptual events provide information to multiple sensory modalities simultaneously. So, for example, when watching a ball bounce, we normally also hear the sound emanating from the location of impact when it hits the ground. Additionally, the timing of each impact sound will be synchronized with the ball's rebounding from the floor. The quality of the sound may also inform the viewer about the material properties of the objects that collided, not to mention how forceful the collision was (e.g., Grassi & Casco, 2010; see Spence & Zampini, 2006, for a review). This example of the bouncing ball highlights three basic properties of multisensory perceptual events (Kubovy & van Valkenburg, 2001; Werner & Noppeney, 2010a): Two of them involve the spatial and temporal information about the event (i.e., 'where' and 'when' it occurred), while the remaining one refers to the content of the event (i.e., the identity, 'what'). When multisensory information originates from a common environmental event, sensory inputs to each modality ought to be congruent in

terms of their space, time, and semantics (as shown in the above example with the ball). Accordingly, the congruency of these three properties constitutes potentially important factors contributing to the final perception of the environmental stimuli.

Researchers have been exploring whether and how inputs from different sensory modalities interact and merge together in human perceptual processing for more than 50 years now (see Calvert, Spence, & Stein, 2004; London, 1954, for reviews). For example, spatial ventriloquism is a classic phenomenon in which people incorrectly locate the source of an auditory stimulus toward a simultaneously-presented visual stimulus. It has been used to demonstrate how visual and auditory inputs interact in the spatial domain (see Howard & Templeton, 1966, for an early review). Similarly, the ventriloquism effect has now been demonstrated in the temporal domain as well: That is, the presentation of a sound close to a visual stimulus will pull the perceived time of onset of the visual stimulus toward that of the sound (Fendrich & Corballis, 2001; Morein-Zamir, Soto-Faraco, & Kingstone, 2003). A dramatic example of crossmodal identity perception comes from the McGurk effect, whereby the sound /ba/ may be misperceived as /da/ when people watch the lip movements normally associated with /ga/ (McGurk & MacDonald, 1976). Researchers sometimes combine two of these factors in order to test human multisensory interactions, such as 'when' and 'where' (e.g., Slutsky & Recanzone, 2001) or all three – 'what', 'where', and 'when' (e.g., Jones & Jarick, 2006). The above well-known phenomena therefore demonstrate that human

perception in each sense is not completely isolated from perception in the other modalities.

In this thesis, I will focus on the crossmodal interactions taking place between vision and audition because these are the two sensory modalities on which humans primarily rely in order to make sense of the outside world. Specifically, the experiments reported here investigate how the presentation of accessory sounds influences visual perception in terms of the onset time and semantic congruency between the visual and auditory stimuli. The spatial consistency factor will only be tested in one experiment in this thesis, since the modulation of human visual performance by audition in the spatial domain has rarely been observed when the participants' task involves identifying the visual target (see below). A possible framework will be proposed in order to account for these audiovisual interactions in the temporal and semantic domains. This chapter contains a review of *why* and *when* audiovisual interactions occur, *what* the possible perceptual outcomes are, and finally, *how* visual and auditory information interacts in the human brain.

1.1 Why does audiovisual interaction occur?

Sensation and perception help us to acquire the information present in the outside world: '*Sensation…Its function is that of mere acquaintance with a fact. Perception's function, on the other hand, is knowledge about a fact; and this knowledge admits of*

numberless degrees of complication. But in both sensation and perception we perceive the fact as an immediately present outboard reality, and this makes them differ from 'thought' and 'conception,' whose objects do not appear present in this immediate physical way.' (James, 1890; Chapter 17). That is, we humans understand the surrounding environment by receiving information via the five 'basic' senses and initially representing it separately in each perceptual system. The outcome of perceptual processing should provide an accurate representation of the outside world, such that an observer can generate adequate behavioural responses. Hence, the integration of information from different sensory modalities serves a critical ecological function (e.g., de Gelder & Bertelson, 2003; Stein, London, Wilkinson, & Price, 1996).

Consider encountering a barking dog, different types of signals concerning the dog are delivered (e.g., the light reflected from the surface of the dog's coat, and the sound waves produced from the dog's throat). A human observer accordingly receives the information from the receptors in different sensory modalities which are anatomically separated (e.g., the cones and rods of the retina for the visual system, and the hair cells in the cochlea for the auditory system). In order to correctly represent 'a single barking dog' (rather than two dogs, one silent but visible, and the other, barking but invisible), visual and auditory information should be matched and integrated at some level during perceptual processing (e.g., Bedford, 2001; Beierholm, Körding, Shams, & Ma, 2008). On the other hand, in the situation in which we see a dog but hear a cat's meowing, we

may infer that there are two objects (i.e., a silent visible dog and a meowing invisible cat). This is a case in which auditory and visual information were compared but finally represented independently (i.e., no perceptual integration occurred; see Beierholm et al., 2008; Jackson, 1953). The term 'multisensory (or audiovisual) interaction' is therefore used to cover both integrated and non-integrated conditions.

Note that here I am arguing that the outcome of the audiovisual interactions is the result of inference processing in the perceptual system. That is, as each sensory modality independently receives inputs, the perceptual systems most likely need to rely on additional information in order to correctly infer the origin of these sensory inputs. This additional information is the prior knowledge or heuristics that the brain has regarding the stimulation that it encounters. Hence, perception is indirect (see also Ernst & Bülthoff, 2004; Shams & Beierholm, 2010). Researchers who hold this view regarding multisensory perception tend to focus on how crossmodal information interacts during multisensory information processing. The contrasting view is the ecological approach that sensory inputs provide sufficient information for perception (e.g., think of the optic array, Gibson, 1979). For example, Stoffregen and Bardy (2001) have tried to extend the ecological approach to a multisensory setting. Specifically, they have suggested that multisensory perception can be direct since the ambient arrays (e.g., including the optic and acoustic arrays for the case of audiovisual perception) provide specified and sufficient information, and any additional prior knowledge or inference

rules beyond those existing in the sensory stimulation is not required. Note, however, that the ecological approach proposed by Stoffregen and Bardy (2001, Section 2) holds very different prior assumptions. That is, the stimulus energies for each sensory modality are inseparable, and the sensory and/or perceptual systems are anatomically and neurophysiologically indistinguishable as well. The latter two arguments, however, may challenge the present knowledge to the anatomical and functional structure of human brain (see Section 1.4.3).

1.2 When (in which conditions) do audiovisual interactions occur?

Given that our perceptual systems need to organize sensory information in order to correctly represent external stimulation, audiovisual interactions should therefore occur when visual and auditory inputs simultaneously bombard our perceptual systems; that is, almost all of our waking life. The perceptual system has to group crossmodal information which likely originates from the same external object or event while trying to segregate information which likely comes from different sources (see Bedford, 2001; Shams & Beierholm, 2010; Spence & Chen, in press). The possible modulatory factors have been grouped into two categories: *structural* and *cognitive* (Radeau & Bertelson, 1977; Spence, 2007; Welch & Warren, 1980). Structural (or stimulus-driven) factors are postulated to consist of such things as spatial and temporal coincidence (or disparity), whilst cognitive factors are thought to depend on knowledge or top-down control (see

Table 1.1).

Table 1.1. Structural and cognitive factors that influence multisensory interactions, as summarized by Welch and Warren (1980).

Structural factors	Cognitive factors[1]
Magnitude of the discrepancy (in space)	Awareness of the discrepancy
Timing of the stimulus information	The assumption of unity
Active versus passive placement of the limb	The 'compellingness' of the stimulus situation

1.2.1 Structural factors: Spatial and temporal coincidence

Until now, it has been the spatial and temporal factors relating to the stimulus input that have attracted most of researchers' attention (for recent reviews, see Alais, Newell, & Mamassian, 2010; Spence, 2007). With regards to the important role of spatial and temporal factors in multisensory interactions, the most important milestone came with the publication of Stein and Meredith's (1993) seminal volume, *The merging of the senses*. This review of the results of their single-cell neurophysiological research over the preceding couple of decades revealed that about 30% of the neurons in the superior colliculus (SC) of the cat respond to both visual and auditory inputs. Such neurons are called 'multisensory'. Furthermore, they found that the activity of multisensory neurons

[1] It should be noted that these three cognitive factors cannot easily be distinguished from one another. For example, when an observer is aware of the discrepancy between the stimuli presented in the different modalities, the stimuli will be impossible to be taken as originating from the same external source (i.e., the assumption of unity is not supported). Furthermore, Welch and Warren (1980) have suggested that '*a highly compelling stimulus situation is one in which the assumption of unity is strong*' (p. 649). That said, the cognitive factors proposed by Welch and Warren (1980) may be reduced to the '*assumption of unity*' to provide a more parsimonious single cognitive factor.

elicited by the combination of visual and auditory inputs can occasionally be larger than that elicited by either unimodal input alone (see Stein & Stanford, 2008; Stein, Stanford, Ramachandran, Perrault, & Rowland, 2009, for recent reviews). Stein and Meredith proposed three basic rules of multisensory integration: That stimuli presented in the different sensory modalities should be *spatially* and *temporally coincident* in order to stimulate the same multisensory neuron, and that the outcome of multisensory integration follows the principal of *inverse effectiveness* (i.e., the weaker the stimuli, the larger the facilitation. see Section 1.3.3.4 for further discussion).

The spatial and temporal rules of multisensory integration were based on the characteristics of neural activity in the SC of the cat. For example, the sensory inputs from different modalities have to be presented within the receptive field (RF) of a given multisensory neuron in order to be summed together. That is, the specificity of the spatial rule inversely correlates with the size of the RF of multisensory neurons (see Chapters 8 & 10 in Stein & Meredith, 1993; see also Carriere, Royal, & Wallace, 2008). Similarly, the neural activities elicited by individual sensory inputs have to overlap temporally, thus giving rise to the temporal rule. For example, if a sensory input elicits a neural response lasting for a certain amount of time (say 100 ms), multisensory interactions therefore occur in a certain time window (about 200 ms in this case, see Stein & Meredith, 1993, Chapter 10).

The spatial and temporal rules are accordingly considered as basic principles of

multisensory interactions, and researchers have tried to verify these two rules in human behaviour. So, for example, the presentation of white noise has been shown to enhance a participants' sensitivity to a masked LED as compared to a sound absent condition, when the sound was presented at the same location (or with a spatial disparity of less than 16°) and at the same time (or with an asynchrony of less than 100 ms, see Bolognini, Frassinetti, Serino, & Ladavas, 2005; Frassinetti, Bolognini, & Ladavas, 2002). The fact that human performance in response to crossmodal stimuli obeys the spatial and temporal rules has sometimes been used to confirm that crossmodal interactions have indeed occurred (rather than some sort of general alerting effect resulting from the presentation of a stimulus in another sensory modality, see van Vleet & Robertson, 2005).

1.2.2 Cognitive factor: Semantic congruency

In contrast to the idea that the semantic systems for the different senses are completely independent (cf. McCarthy & Warrington, 1988), the semantic systems may be either completely amodal (e.g., Pylyshyn, 1984), or else the semantic systems for each sensory modality may be highly connected (e.g., Plaut, 2002; Shallice, 1988). The latter two notions can be used to explain audiovisual interactions in terms of the semantic congruency of the component unimodal stimuli if any effect is observed (see Figure 1.1). These two possibilities concern the question of whether information

concerning the source modality is still preserved with in human semantic representations (see Barsalou, 1999; Pennartz, 2009).

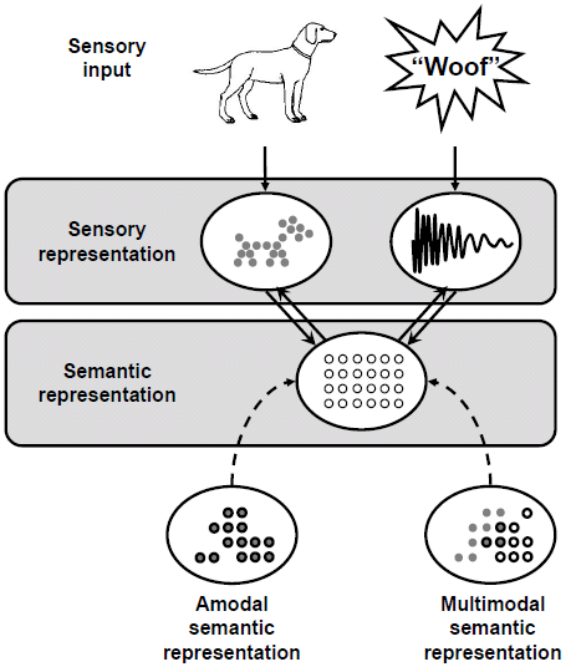


Figure 1.1. A visual input (the picture of a dog) and an auditory input (the sound of a dog's woofing) initially go through sensory processing separately before accessing their respective semantic representations. The gray dots and the open circles represent the semantic nodes for the visual and auditory modality, respectively. According to the amodal semantic representation system, visual and auditory semantics share the same nodes. However, according to the multisensory semantic system view, visual and auditory semantic nodes are non-overlapping but nevertheless still inter-connected.

One of the first examples of crossmodal semantic congruency was reported by Jackson (1953) in a study of the spatial ventriloquism effect. Specifically, he showed that the spatial ventriloquism effect occurred over a larger range of spatial disparities for realistic pairs of visual and auditory stimuli (e.g., the whistling sound and the sight of a steaming kettle) than for arbitrary stimulus pairings (e.g., the sound of a bell and the

illumination of a light). That is, Jackson's results suggested that people can tolerate a larger spatial separation between congruent (as compared to incongruent) pairs of audiovisual stimuli and still bind them together. Welch and Warren (1980) suggested that semantically-congruent stimuli from different sensory modalities should be inferred as coming from the same source, which they called the 'unity assumption'.[2] Much more recently, other researchers have followed-up on the notion of the unity assumption in order to investigate whether people can bind visual and auditory stimuli while tolerating a larger *temporal* separation between pairs of stimuli that happen to be semantically congruent versus semantically incongruent (e.g., Vatakis, Ghazanfar, & Spence, 2008; Vatakis & Spence, 2007; though see Vatakis & Spence, 2008).[3]

1.2.3 The roles of spatial, temporal, and semantic factors in audiovisual interactions

Presumably, spatial and temporal coincidences are concrete properties shared by crossmodal stimuli in natural situations, while the semantic congruency is an abstract association. The latter also involves the observer's familiarity regarding the crossmodal stimulus pair. Hence, for example, spatial and temporal factors may have an advantage

[2] Note that semantic congruency is only one of the possible factors that can induce the unity assumption (see Spence, 2007). It can also result from experiencing a reliable pairing of crossmodal stimuli (Ernst, 2007; Lippert, Logothetis, & Kayser, 2007), or specific experimental instructions (e.g., to imagine that both the visual and auditory stimulus have a common source; see Alais & Burr, 2004; Sugita & Suzuki, 2003).

[3] The pairs of stimuli in the congruent and incongruent conditions used by Jackson (1953) were, however, quite different in terms of their complexity and meaningfulness. That said, Vatakis and her colleagues provide a better control by using the same meaningful visual and auditory stimuli while manipulating their semantic congruency (i.e., matched or not, see Vatakis et al., 2008; Vatakis & Spence, 2007, 2008).

in terms of governing multisensory interactions when the spatiotemporal structure of crossmodal stimuli is obviously coincident or disparate. On the other hand, semantic congruency may sometimes override spatiotemporal factors when the spatiotemporal structure of crossmodal stimuli is highly uncertain. That is, when the stimuli are presented within a threshold range where they are perceived as either spatially (or temporally) coincident or else as disparate (see Jackson, 1953; Vatakis et al., 2008; Vatakis & Spence, 2007).

Welch and Warren (1980) suggest that some of the factors may be dominant over others in certain experimental settings, which they refer to as the 'compellingness' of the stimulus situation. A question is therefore raised: Will every factor be evaluated and weighted during the processing of multisensory interactions, or will only the most dominant factor determine the outcome of multisensory interactions in a compelling situation? This question involves how multisensory information interacts during information processing, which will be discussed in Section 1.4.1.

In summary, multisensory interactions occur when information from different sensory modalities potentially originate from the same object or event. That is, crossmodal sensory inputs which are spatially/temporally coincident, and/or semantically congruent, are likely compared and finally integrated into a unified multisensory representation. During the processing of multisensory information, however, spatial, temporal, and semantic factors may not play an equivalent role.

1.3 What are the possible outcomes of audiovisual interactions?

Partan and Marler (1999) provided a useful classification scheme of the relationships between the individual sensory inputs and the response output (see Figure 1.2). According to their classification scheme, redundancy refers to 'response redundancy' since both signals (A & B) correspond to the same response (A, in the example in the upper part of the panel). However, psychologists interested in human information processing prefer to use *response nonredundant* stimuli in order to investigate multisensory interactions. That is, the participant perhaps only needs to respond to one of the two stimuli. The researchers then manipulate the accessory stimulus in terms of whether it is informative or not (or even conflicting) with respect to

	Unimodal		Multisensory		
	Signal	Response	Signal	Response	Category
Redundancy	A →	A	A + B →	A	**Equivalence** (intensity unchanged)
	B →	A	A + B →	**A**	**Enhancement** (intensity increased)
Nonredundancy			A + B →	A + B	**Independence**
	A →	A	A + B →	A	**Dominance**
	B →	B	A + B →	**A** (or A)	**Modulation**
			A + B →	C	**Emergence**

Figure 1.2. Classification of multisensory signals and the response summarized by Partan and Marler (1999).

the perceptual processing of the target. Here, I will review previous studies that have demonstrated a perceptual modulation resulting from multisensory interactions. Note that in the two conditions, 'equivalence outcome' and 'independence outcome' when multisensory signals are presented, it is hard to infer any internal information processing for multisensory interactions, and hence, they are excluded from further discussion.

1.3.1 Perceptual enhancement by redundant multisensory signals

When two stimuli (either from the same modality or from different sensory modalities) map onto the same response, they are considered as 'redundant' (Miller, 1982). Reaction times (RTs) to redundant signals are normally faster than to the individual unisensory signals. This constitutes the response *enhancement* in the redundancy condition. Raab (1962) suggested that this redundancy facilitation may be attributable to participants' responses being triggered by the first stimulus that they happen to perceive (i.e., a statistical facilitation of responses might have been observed). Miller (1982) subsequently developed a race model which provides a method with which to simulate the RT distribution due to statistical facilitation. He argued that any violation of the predictions of the race model would therefore indicate that an interaction between the two redundant signals had occurred.

For example, Miller (1982, Experiment 3) arbitrarily assigned the visual stimulus 'X' and a 780-Hz tone as targets (and the visual stimulus 'O' and a 360-Hz tone as

distractors). The participants were instructed to press the same key as soon as they detected the presence of either target stimulus, or their combination (i.e., a redundant target signal). Miller found that the RT distribution in the redundant condition (to the combination of the 'X' and the 780-Hz tone) violated the race model predicted from the RTs to the 'X' and 780-Hz tone when presented alone. Note, however, that Miller attributed the crossmodal interaction to coactivation occurring in the process of deciding to respond (Miller, 1982, p. 262).

Three recent studies have used a similar methodology in order to explore whether human visual (or auditory) performance can be enhanced when a semantically congruent stimulus is presented in the other modality. For example, when the semantically-congruent picture (e.g., a cow) and sound (e.g., a cow's mooing) correspond to the same response, detection latencies tend to be shorter for bimodal trials (i.e., the picture of a 'cow' and the 'mooing' sound are presented together) than the RTs predicted by the race model (see Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004; Molholm, Ritter, Javitt, & Foxe, 2004; Suied, Bonneel, & Viaud-Delmon, 2009). However, such crossmodal facilitation can in fact be observed once the visual and auditory target map onto the same response even when they are semantically incongruent (e.g., a picture of 'traffic light' paired with the sound of a cat's 'meowing', see Sinnett, Soto-Faraco, & Spence, 2008). That said, RT facilitation can simply be attributed to the response redundancy while irrelevant to the semantic congruency, as

suggested by Miller (1982). A possible solution to control the response redundancy may be to compare the RT facilitation for the crossmodally-congruent pairs of stimuli versus incongruent pairs under those conditions in which both of the stimuli map onto the same response (i.e., the response redundancy in both conditions were matched, see Experiment 1 in Miller, 1991). Presumably, the RT facilitation should be larger in the semantically congruent condition than in the incongruent condition.

1.3.2 Perceptual dominance by nonredundant multisensory signals

As mentioned earlier, spatial ventriloquism is a phenomenon in which vision dominates over audition in the perception of spatial location. It used to be thought of as resulting from the 'visual capture of sound'. On the other hand, 'auditory capture' has also been demonstrated in the temporal domain. For example, when presenting a pair of visual flickering and auditory fluttering stimuli consisting of different temporal frequencies, the perceived visual flicker rate tends to be biased toward the rate of auditory flutter, but not *vice versa*. This is known as the '*auditory driving effect*' (see Gebhard & Mowbray, 1959; Recanzone, 2003; Shipley, 1964). Such an asymmetrical modulation from audition to vision in the temporal domain can also be observed in the perception of the duration of an audiovisual event. Specifically, the perceived duration of a visual event can be modulated by the duration of an accompanying sound, but not the other way around (Chen K.-M. & Yeh, 2009; Walker & Scott, 1981; though see

Schutz & Kubovy, 2009; van Wassenhove, Buonomano, Shimojo, & Shams, 2008). That said, audition is typically dominant over vision in the temporal domain in multisensory perception.

Shams, Kamitani, and Shimojo (2000, 2002) reported an illusion in which, when a visual flash was accompanied by the presentation of two brief beeps, observers perceived two flashes rather than just one. This sound-induced illusory flashing may be considered as providing evidence that audition is dominant over vision in the temporal domain. However, when two flashes were accompanied by one beep, participants reported that they perceived two flashes accurately (Anderson, Tiippana, & Sams, 2004; Shams et al., 2002). That said, even though audition may have an advantage in terms of better temporal resolution than vision (e.g., Gebhard & Mowbray, 1959), audition does not always dominate vision in the temporal domain (see Anderson et al., 2004; Burr, Banks, & Morrone, 2009; Shimojo & Shams, 2001).

It should be noted that, when carefully measuring human performance in perceptual dominance phenomena, the participants do not make their response entirely according to the information presented in the dominant modality. Instead, they reported the sound location or visual flicker rate that lies somewhere in-between the locations or temporal frequencies provided by each of the component modalities (e.g., Pick, Warren, & Hay, 1969; Welch, DuttonHurt, & Warren, 1986). Hence, modality dominance is typically not a winner-takes-all phenomenon but is instead a weighted sum of the

information provided by each modality. Recent studies have also shown that the nature of the stimuli, such as the variance regarding its spatial location or presentation duration, can modulate the relative weighting of the modalities (e.g., Alais & Burr, 2004; Heron, Whitaker, & McGraw, 2004; Witten & Knudsen, 2005). This issue will be further discussed in the section of the mechanisms underlying multisensory interactions (see Section 1.4.1).

1.3.3 Perceptual modulation by nonredundant multisensory signals

Most researchers have adopted the strategy of measuring human performance in one sensory modality (such as vision) while providing additional information in another sensory modality (such as audition). Note that some of the modulatory effects may be attributable to the presentation of sound simply alerting participants crossmodally (Nickerson, 1973; Robertson, Mattingley, Rorden, & Driver, 1998)[4], or to the sound directing participants' attention toward (or away from) the spatial location of the visual target (e.g., McDonald, Teder-Sälejärvi, & Hillyard, 2000; for a review, see Spence, 2010a). Here I will mainly review the results which can genuinely be attributed to audiovisual interactions between accessory sounds and visual targets in the spatial, temporal, and semantic domains, respectively.

[4] Saying that the presentation of a sound alerts a participant means that the observer may be more vigilant and prepared to encounter or respond to any stimulus (Nickerson, 1973). Alerting is considered as a fast and non-specific modulation of human behaviour (Driver & Noesselt, 2008). Even though alerting is able to induce a certain form of crossmodal facilitation, psychologists are more interested in other specific mechanisms affecting human information processing.

1.3.3.1 Modulating visual perception as a function of the spatial properties of a sound

The presentation of a simultaneous sound can enhance the sensitivity of a participant's visual detection performance, no matter whther it is spatially coincident (e.g., Bolognini et al., 2005; Frassinetti et al., 2002), or not (see Hughes, Reuter-Lorenz, Nozawa, & Fendrich, 1994; Noesselt, Bergmann, Hake, Heinze, & Fendrich, 2008). It is possible that the sound was spatially ventriloquised to the visual target, and thus the spatial disparity between the visual and auditory stimuli simply went unnoticed. Due to the fact that, in most cases, stimuli in the visual modality provide more precise spatial information than stimuli in the auditory modality, the perceived location of a visual stimulus is rarely modulated by the location of a sound. However, in a particular situation in which the location of the visual target is highly uncertain, the spatial location of a sound has been demonstrated to modulate visual perception (e.g., Alais & Burr, 2004; Heron et al., 2004). The stimulus uncertainty factor therefore provides researchers with an important clue with which to investigate the mechanisms underlying audiovisual interactions (see Section 1.4.1).

1.3.3.2 Modulating visual perception as a function of the temporal properties of a sound

The crossmodal facilitation of visual performance by a simultaneously-presented sound, on the other hand, has been demonstrated in various phenomena and experimental paradigms. For example, when trying to localize a moving visual object at

a given point in time, an auditory temporal cue enables participants to respond more accurately as compared to a visual temporal cue (Fendrich & Corballis, 2001; Vroomen & de Gelder, 2004). Besides, people's sensitivity in a visual detection task can be enhanced by the presentation of a synchronous sound as compared to when the target is presented by itself (Bolognini et al., 2005; Frassinetti et al., 2002; Lippert, Logothetis, & Kayser, 2007).[5] Finally, any transient change of a visual target (e.g., luminance change or brief offset) can be detected more accurately when a sound changes synchronously, irrespective of whether the direction of change (e.g., higher or lower sound intensity) is consistent with the visual change or not (Anderson & Mamassian, 2008; Noesselt et al., 2008).

In addition, the presentation of a sound which is only informative with regards to the onset time of the visual targets (while irrelevant with regards to their spatial location and identity) can also enhance the *identification* of that visual target. That said, any response criterion shifting attributable to the presentation of a sound cannot be used to explain this crossmodal facilitatory effect, since the participants' response relies on the identity of the visual target. Such effects have now been demonstrated in the following two situations: In the rapid serial masked visual presentation paradigm, the presentation of a distinct sound can enhance visual perception of the visual target which is embedded

[5] It should be noted that the presentation of the sound also resulted in participants lowering their response criterion in the studies of Bolognini et al. (2005), Frassinetti et al. (2002), and Lippert et al. (2007). That is, the presentation of the sound seemingly encouraged the participants to make a response.

in a series of distractors and masks. This is known as the 'freezing effect' (Ngo & Spence, 2010a, b; Vroomen & de Gelder, 2000; for a recent review, see Spence & Ngo, in press). In dynamic visual search displays, the colour change of the visual target is embedded in a series of colour-changing distractors. The onset of a sound which occurs simultaneously with the target colour change can enhance its perception, as compared to when no sound was presented. This is known as the 'pip-and-pop' effect (see Ngo & Spence, 2010c; van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008a).

The presentation of a sound can also reduce two visual deficits observed in normal participants: In the repetition blindness (RB) paradigm, an observer may fail to detect or report the second occurrence of a repeated item in a rapid serial visual presentation (RSVP) stream (Kanwisher, 1987). In the attentional blink (AB) paradigm, an observer may fail to detect or identify the second target if it happens to be presented 200-500 ms after the presentation of the first target in a RSVP stream (Raymond, Shapiro, & Arnell, 1992). In both phenomena, the presentation of a sound simultaneous with the visual target can improve participants' visual identification performance (see Chen Y.-C. & Yeh, 2008, 2009; Olivers & van der Burg, 2008, for the crossmodal facilitation in RB and AB, respectively).

Note that Ngo and Spence (2010a, c) have recently demonstrated that the presentation of a sound which is both spatially and temporally coincident with respect to the visual target can lead to a greater facilitation of visual performance in the freezing

effect and the pip-and-pop effect than the presentation of a sound that is simply temporally coincident. However, the RTs in Ngo and Spence's (2010a, c) studies generally exceeded 2000 ms. Hence, the auditory cue likely induced a shift of spatial attention that enhanced participants' visual target identification (see Spence & Ngo, in press).

The presentation of a sound can also modulate the nature of a participant's visual perception. For example, when two identical objects move toward each other, coincide, and continue to move away from each other, people may perceive these two objects as either streaming through one another or else as bouncing off each other (Metzger, 1934). When a sound is presented at the time of coincidence in such a bistable streaming/bouncing situation, observers report the percept of bouncing more frequently than when no sound was presented (Sekuler, Sekuler, & Lau, 1997; Watanabe & Shimojo, 2001). Note that in this phenomenon, the multisensory interaction simply modulates the proportion of trials on which the two possibilities are perceived, rather than generating a completely new percept (see Section 1.3.4). That the bouncing percept can be induced by the presentation of a sound may be attributable to the fact that, in daily life, the collision of objects is always accompanied by a collision sound.[6] In addition, the timing of the onset of a sound can also be used to pull the perceived timing

[6] Shimojo, Watanabe, and Scheier (2001), however, suggested that the increased rate of perceiving bouncing may be partly due to the presentation of the sound distracting participants' attention from the visual display. Hence, the observers cannot fully monitor the moving trajectory of the objects which, in turn, increases the possibility of perceiving bouncing.

of visual stimuli by means of temporal ventriloquism. As a result, the presentation of the sound can bias the direction of visual motion perception in apparent motion displays (Bruns & Getzmann, 2008; Freeman & Driver, 2008) or the type of element versus grouped motion perceived in Ternus displays (Shi, Chen, & Muller, 2010).

1.3.3.3 Modulating visual perception as a function of the semantic properties of a sound

The presentation of a naturalistic sound (e.g., a dog's bark) can facilitate people's performance regarding a semantically-congruent picture (e.g., a dog's picture in this example) in a picture categorization task (Schneider, Engel, & Debener, 2008) or a picture localization task (Iordanescu, Grabowecky, Franconeri, Theeuwes, & Suzuki, 2010; Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008). Audiovisual semantic interactions between naturalistic sounds (or spoken words) and pictures (or written words) have also been investigated, typically using Stroop-like interference paradigms (e.g., Roelofs, 2005; Shimada, 1990; Stuart & Carrasco, 1993; Yuval-Greenberg & Deouell, 2009). However, the crossmodal semantic modulation of human performance can only be observed under certain specific conditions. For example, the presentation of spoken colour words has been shown to induce a Stroop effect on the naming of visual colour patches, but the presentation of colour patches does not appear to influence performance when shadowing spoken colour words (Roelofs, 2005). Besides, the presentation of spoken words only induced a Stroop effect

in the picture naming task but not in the picture categorization task (Stuart & Carrasco, 1993).

Another popular approach to investigate audiovisual semantic congruency has been to utilize the temporal order judgment (TOJ) task in which participants have to report which stimulus (in vision or audition, in this case) was presented (or onset) first. The results of numerous TOJ studies have now revealed that a larger temporal asynchrony is needed for semantically-congruent pairs of visual and auditory stimuli, than for semantically-incongruent stimulus pairs, in order to correctly judge the temporal order of occurrence of these two events (van Wassenhove, Grant, & Poeppel, 2007; Vatakis et al., 2008; Vatakis & Spence, 2007). A critical implication of these results is that the temporal window of audiovisual integration can be modulated by the semantic congruency between the stimuli. Note, however, that the studies that demonstrated semantic congruency effects primarily used speech stimuli (such as verbal sounds and videos of lip movements).

1.3.4 Perceptual emergence by nonredundant multisensory signals

The most dramatic perceptual outcome of multisensory interactions is that the integrated visual and auditory information produce a new percept which is different from either unimodal stimulus. The McGurk effect (McGurk & MacDonald, 1976) can therefore be considered as the most obvious example of the emergence of a new percept

following multisensory integration. However, until now, it seems that the emergence of a new percept resulting from multisensory integration is hard to observe in other situations (see Spence & Chen, in press, for a review). For example, Huddleston, Lewis, Phinney, and DeYoe (2008) tested whether visual and auditory stimuli can be grouped and construed as giving rise to smooth apparent motion: They presented two lights from the top and bottom of a display, and two sounds from the left and right of the display, respectively. With an adequate interstimulus interval (ISI), the participants perceived both vertical visual apparent motion, and horizontal auditory apparent motion; however, no matter what the ISI they never perceived a unified multisensory circular apparent motion.

It is, in fact, hard to define what exactly is meant by a 'new perception' emerging as a result of multisensory interactions. In the McGurk effect, for example, the perceptual outcome is categorically (or qualitatively) different from either of the original unimodal stimuli. In the sound-induced illusory flashing phenomenon, when one visual flash and three beeps are presented, participants often report that they only perceived two (rather than three) flashes (Anderson et al., 2004; Shams et al., 2000, 2002). However, this result may be attributed to the third sound having been presented outside of the optimal temporal window with respect to the onset of the real visual flash in order to induce illusory flashing (see Shams et al., 2002). Hence, the sound-induced illusory flashing phenomenon may be properly considered as a 'perceptual dominance'

effect whereby audition is dominant over vision in the temporal domain (see Section 1.3.1; though see Anderson et al., 2004). Given that the McGurk effect is the only robust phenomenon which demonstrates that the interaction of audiovisual information gives rise to a new percept, speech perception is perhaps unique. Indeed, this claim has been discussed for many years now (e.g., Fowler, 2004; Liberman & Mattingly, 1985; Massaro, 2004; Vatakis et al., 2008).

1.3.5 The inverse effectiveness rule in neurophysiological and behavioural studies

Stein and Meredith (1993) summarized their single-cell recording results and proposed the *inverse effectiveness rule*: That is, the combination of weak unimodal stimuli produces the greatest enhancement. For example, Meredith and Stein (1986) demonstrated that the multisensory enhancement revealed by neural activity within the same cell in the SC increased from 110% to 483% when the activity elicited by unimodal stimuli became weaker (see also Stein & Meredith, 1993, p. 145). The outcome of multisensory enhancement can be further classified into subadditive (when the neural activity elicited by combination of visual (V) and auditory (A) stimuli is smaller than the sum of those elicited by V or A stimulus individually, i.e., $VA < V + A$), additive ($VA = V + A$), and superadditive interactions ($VA > V + A$; see Stein et al., 2009). Note that the baseline against which performance is compared for the inverse

effectiveness and superadditive multisensory enhancement are different.[7]

When investigating audiovisual interactions using functional magnetic resonance imaging (fMRI), Werner and Noppeney (2010b) reported that multisensory enhancement can be observed only when the visual and auditory stimuli are degraded but not when they are clearly perceptible. In addition, they also demonstrated that the multisensory enhancement observed in the left superior temporal sulcus (STS, an area involving multisensory object recognition, see Doehrmann & Naumer, 2008) indexed by the blood oxygenation level-dependent (BOLD) response for multisensory stimuli was inversely correlated with the sum of the unimodal BOLD responses. That is, the multisensory enhancement was larger when the individual unimodal stimuli elicited a smaller BOLD response. This result can be considered as showing that the responses of a group of neurons to multisensory stimulation is consistent with the rule of inverse effectiveness (see also Stevenson & James, 2009; though see Meyer, Wuerger, & Greenlee, in press).

In human behavioural studies, researchers may make a participant's task difficult (such as by adding a visual mask following a visual target), so that any crossmodal

[7] Note that the multisensory enhancement compared the neural activity elicited by multisensory inputs to the stronger activity elicited by either of unimodal sensory inputs using the equation: $((CM - SM_{max}) / SM_{max}) * 100$, where CM = the number of impulses evoked by the combined-modality stimulus, and SM_{max} = the response to the most effective single-modality stimulus (Meredith & Stein, 1983; Stein & Meredith, 1993).
On the other hand, the question of whether superadditive multisensory enhancement occurred is calculated using the equation: $((CM - (V + A)/(V + A)) * 100$, where V = the number of impulses evoked by the visual stimulus, and A = the number of impulses evoked by the auditory stimulus (see King & Palmer, 1985; Holmes & Spence, 2005).

facilitation elicited by presenting a spatially and/or temporally coincident sound is easier to observe (e.g., Bolognini et al., 2005; Frassinetti et al., 2002). However, when the difficulty of the task is itself manipulated, the largest multisensory enhancement effect may be observed at intermediate levels of task difficulty (see Lippert et al., 2007; Ma, Zhou, Ross, Foxe, & Parra, 2009; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). That said, although inverse effectiveness has been observed in neurophysiological studies (at the single cell level in anaesthetized animals), it does not mean that inverse effectiveness will necessarily occur in human (or for that matter animal) behaviour (see Holmes, 2009).

In this Section, the outcomes of audiovisual interactions reported in previous studies have been reviewed according to Partan and Marler's (1999) classification scheme (see Table 1.2). In summary, the results that have been obtained from response-redundant multisensory signals should be explained carefully, since they may be attributed to interactions taking place at the decisional rather than at the perceptual level. Most evidence of multisensory interactions comes from results showing that perception in one sensory modality can be enhanced or modulated by a task-irrelevant signal presented in another sensory modality. For the cases of audiovisual interactions that this thesis will focus on, visual perception is often reported as being modulated by auditory information in terms of its temporal properties, while less frequently in terms of its spatial or semantic properties (see Spence, 2007, for a review). Finally, to date, no

Table 1.2 Summary of perceptual outcomes resulting from the influence of audition on vision in terms of Partan and Marler's classification

Category	Example	Studies
Redundant signals, Enhancement	Visual and suditory stimuli correspond to the same response.	Laurienti et al. (2004); Miller (1982); Molholm et al. (2004); Suied et al. (2009); Sinnett et al. (2008)
Nonredundant signals, Dominance	Audition dominate vision in the temporal domain	Gebhard and Mowbray (1959); Recanzone (2003); Shipley (1964); Chen K.-M. and Yeh (2009); Walker and Scott (1981);
	Sound-induced illusory flashing	Shams, Kamitani, and Shimojo (2000, 2002); Anderson, Tiippana, and Sams (2004)
Nonredundant signals, Modulation	Audition can modulate vision in spatial domain when the visual spatial information is highly uncertain	Alais and Burr (2004); Heron et al. (2004)
	The presentation of a simultaneous sound can enhance visual perception in various tasks	Visual localization lask: Fendrich and Corballis (2001); Vroomen and de Gelder (2004) Visual detection task: Anderson and Mamassian (2008); Bolognini et al. (2005); Frassinetti et al. (2002); Lippert et al. (2007); Noesselt et al. (2008) Visual identification task: Chen Y.-C. and Yeh (2008, 2009); Ngo and Spence (2010a, b, c); Olivers and van der Burg (2008); van der Burg et al. (2008a); Vroomen and de Gelder (2000)
	Bouncing-streaning bistable perception	Sekuler et al. (1997); Watanabe and Shimojo (2001)
	Direction of visual motion perception induced by temporal ventriloquism effect	Bruns and Getzmann (2008); Freeman and Driver (2008); Shi, Chen, and Muller (2010)
	The presentation of meaningful sound modulate visual object perception in terms of their semantic congruency in various tasks	Visual categorization task: Schneider et al. (2008) Visual localization task: Iordanescu et al. (2008, 2010) Stroop-like interference paradigm: Roelofs (2005); Shimada (1990); Stuart and Carrasco (1993); Yuval-Greenberg and Deouell (2009) TOJ task: van Wassenhove et al. (2007); Vatakis et al. (2008); Vatakis and Spence (2007, 2008)
Nonredundant signals, Emergence	McGurk effect	McGurk and MacDonald (1976)

obvious evidence has been published to demonstrate that audiovisual integration can give rise to a new perception, except in the case of speech perception (i.e., in the McGurk effect).

1.4 How do visual and auditory information interact?

In answer to the final question addressed in this chapter, Welch and Warren (1980) in their early review provided three hypotheses with which to explain how unimodal sensory input interacts with each other. They also proposed a multisensory interaction framework in which all of the structural and cognitive factors conjointly affected the perception of a crossmodal event. Recently, Driver and Noesselt (2008) reviewed the neurophysiological evidence concerning multisensory interactions and proposed three possible multisensory interaction networks in the human brain.

1.4.1 Modality appropriateness, modality preciseness, and the directed-attention hypothesis

The spatial ventriloquism effect demonstrates that vision is typically dominant in the spatial domain, while the auditory driving effect demonstrates that audition is typically dominant in the temporal domain (see also Freides, 1974; O'Connor & Hermelin, 1972). These modality dominance phenomena can be attributed to the fact that the visual system has a better spatial resolution while the auditory system has better temporal resolution. This is the central idea underlying the *modality appropriateness*

hypothesis. It is true that, up until now, most studies have demonstrated the influence of audition on vision in the temporal domain (see Section 1.3.3.2), and hence the modality appropriateness hypothesis seems to provide a parsimonious explanation. Note, however, that the dominant modality is weighted higher than, rather than completely overriding, the non-dominant modality in multisensory information processing (see Section 1.3.2).

In contrast to the modality appropriateness hypothesis, resent studies suggest that the weighting of each modality is not fixed but can be adjusted according to the reliability of the sensory input. For example, sound location can bias the perceived location of visual stimuli when visual spatial information is highly uncertain (Alais & Burr, 2004; Heron et al., 2004). Besides, it has recently been shown that apparently moving sounds can induce visual apparent motion for visual stimuli presented in the periphery (Hidaka et al., 2009). This result is presumably attributable to the fact that the acuity of human vision is lower in the periphery than at the fovea. It may also be related to the fact that the visual cortex responsible for peripheral vision receives some direct neural projections from auditory cortex (see Falchier, Clavagnier, Barone, & Kennedy, 2002; see Cappe, Rouiller, & Barone, 2009, for a review). Similarly, the influence of sound can also be reduced by decreasing the preciseness of the timing cued by a sound (e.g., the variance of the sound duration is larger; see Heron et al., 2004). That said, the weighting for the sensory input in each modality is correlated with the preciseness of the information provided by that modality during their interaction. This is the *modality*

preciseness hypothesis (Welch & Warren, 1980).

Human perception, along with modality preciseness, has been simulated in a statistically optimal fashion according to Bayesian rules. This is Maximum-Likelihood Estimation (see Ernst & Banks, 2002; see also Alais & Burr, 2004; Burr et al., 2009; for reviews, see Ma & Pouget, 2008; Witten & Knudsen, 2005). According to the modality preciseness hypothesis, the stimulus with a smaller variance provides more precise information. That is, the weighting for each sensory input in the multisensory interaction is proportional to the inverse of its variance (Alais & Burr, 2004; Ernst & Bank, 2002). Note, however, that Battaglia, Jacobs, and Aslin (2003) have suggested a hybrid model that combines the modality appropriateness and modality preciseness hypotheses.

The third account holds that the relatively attended modality is the dominant one (directed-attention hypothesis, Welch & Warren, 1980). For example, Posner, Nissen, and Klein (1976) suggested that people voluntarily (or endogenously) directed their attention toward the visual modality in order to make up for the fact that visual stimuli are generally less alerting than stimuli presented in the other modalities. Hence, according to this claim, the visual modality may be dominant in certain multisensory interaction phenomena due to the modulatory influence of attention. For example, Colavita (1974) reported that when a visual and an auditory target are presented together, people often simply respond to the visual target while failing to respond to the auditory

target (this is known as the Colavita visual dominance effect). The Colavita effect may be accounted for in terms of the consequence of allotting more attention to the visual modality than to the auditory modality (Posner et al., 1976; though see Spence, 2009; Spence, Parise, & Chen, in press, for recent reviews).

The directed-attention hypothesis holds that sensory modality, like spatial location, can serve as a target for attentional selection (e.g., Spence, Nicholls, & Driver, 2001). That is, fully concentrating on one sensory modality can be considered as a form of attentional selection to that modality. Such attentional selection may occur at a perceptual level (this is known as early selection, e.g., Broadbent, 1958; Treisman, 1969; see Figure 1.3A) or at a decisional level (this is known as late selection, e.g., Deutsch & Deutsch, 1963; Norman, 1968; see Figure 1.3B). According to these notions, the need for attentional selection is attributed to a structural 'bottleneck' in human information processing, because only one sensory input can pass through the bottleneck at any given moment. According to the hypothesis underlying structural theories, multisensory interactions can occur only when the bottleneck is late (so that information in both modalities is processed), but not when the bottleneck is early (in this case, information in the unattended modality is completely filtered out without being processed).

According to the capacity theory (see Figure 1.3C), by contrast, there is no actual bottleneck, but only limited resources are available for use in human information processing (e.g., Knowles, 1963; Lavie, 1995; for reviews, see Lavie, 2005). Hence,

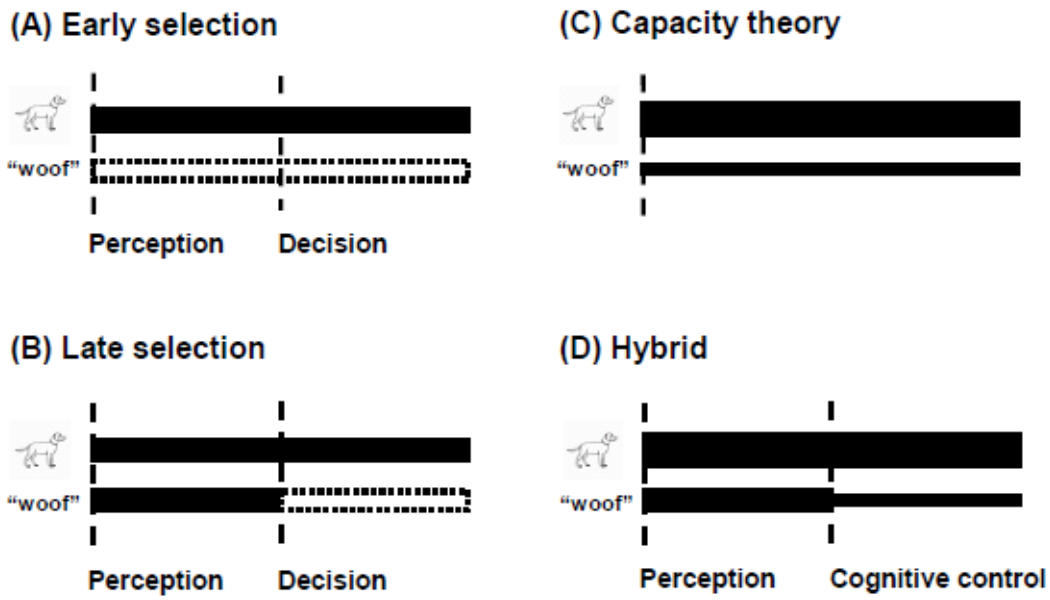


Figure 1.3. Four attentional selection hypotheses regarding the processing of visual and auditory inputs. (A) The bottleneck for attention is located at the perceptual stage of information processing. Hence, only the stimulus in the selected/attended modality (in the figure, vision) can be processed. (B) The bottleneck for attention is located at the decisional stage. Both visual and auditory stimuli go through perceptual processing; however, only the visual stimulus keeps on being processed through to the decisional stage, since it has been selected/attended. (C) There is no absolute bottleneck but only a limited capacity available for visual and auditory processing. Hence, when visual processing occupies most of the resources, few resources remain for auditory processing. (D) The hybrid hypothesis consists of two stages: At the early perceptual stage, the resources available for each sensory modality are independent; however, at the later cognitive control stage, only one resource pool is shared by all sensory modalities. Hence, once the visual processing occupied most of the resources, auditory process can only use the residual resources.

multisensory interactions would be expected to occur only when the loading of the main task is low so that there are still resources available for the stimulus in the non-target modality to be processed. For example, the presentation of a sound (or light) in the peripheral visual field can capture a participant's spatial attention only when no additional stimulus stream other than the visual (or auditory) target is presented

(Santangelo, Belardinelli, & Spence, 2007).[8]

The debate between structural and capacity theories is even more complicated in the multisensory situation. For example, when processing inputs from two rather than one sensory modality, will the capacity for processing multisensory information be greater or else stay the same (e.g., Wickens, 1984, 2002)? Treisman and Davies (1973, p. 115) suggested that '… *perceptual capacity at two different levels at least, one which is relatively peripheral and modality specific, and one which is shared between modalities, but probably located earlier in verbal processing than the semantic level.*' This notion should be considered as a hybrid view of structural and capacity theories (see Figure 1.3D). Brand-D'Abrescia and Lavie (2008) recently have suggested a similar idea that perceptual capacity may be modality-specific, but the capacity of higher-level cognitive control functions may be unitary. These two levels may map to the perceptual load (Lavie & Tsal, 1994) and working memory load (or cognitive control function) proposed by Lavie (2005). Their separation was based on the evidence that a distractor can influence participants' performance when the perceptual load is low while the working memory load is high (see Lavie & de Fockert, 2005; Lavie, Hirst, de Fockert, & Viding, 2004).

Empirical evidence tends to support such a hybrid view of structural and capacity

[8] In another study, however, Santangelo and Spence (2007) demonstrated that when the distractor is highly salient (such as when the cue was presented in multiple sensory modalities at the same time), the distractor can still capture a participant's spatial attention even in a high-load situation (for a review, see Spence, 2010a). This result is consistent with the salience hypothesis (Eltiti, Wallace, & Fox, 2005), and contrasts with Lavie's (1995) perceptual load theory.

theories, in that the difficulty of a low-level visual task (such as contrast discrimination) does not appear to influence the performance of a low-level auditory task (e.g., a pitch discrimination; see Alais, Morrone, & Burr, 2006; see also Duncan, Martens, & Ward, 1997; Tellinghuisen & Nowak, 2003; though see Klemen, Büchel, & Rose, 2009). Similarly, the difficulty of an auditory task (speech detection vs. discrimination) does not influence the activity of visual motion processing in V5 and MT in visual cortex (Rees, Frith, & Lavie, 2001). On the other hand, the deficits in processing multisensory stimuli can be observed in other tasks which have probed central attentional resource (see Arnell & Larson, 2002; Bonnel & Hafter, 1998; Klemen, Büchel, Buhler, Menz, & Rose, 2010; Sinnett, Costa, & Soto-Faraco, 2006). However, up until now, the boundary segregating perceptual from cognitive control functions remains unclear. In summary, when investigating whether attentional resources (or capacity) modulate audiovisual interactions, one should keep in mind that both the loading of the task and the level at which the audiovisual inputs interact are critical factors (Rapp & Hendel, 2003). This issue will be further discussed in terms of the results observed in the experimens reported in this thesis (see Section 9.3.2).

1.4.2 Unitary or multiple mechanisms of multisensory interactions

Welch and Warren (1980; see also Welch, 1999) have proposed a unitary model that contains every structural and cognitive factor. According to this model, all factors

conjointly affect the final perception of a crossmodal event. They also suggested that when a unified perceptual solution emerges, the observer will perceive a coherent event while failing to notice incongruent information. Evidence that is consistent with Welch and Warren's (1980) hypothesis came from studies apparently showing that people can tolerate a larger spatial (or temporal) separation between congruent (as compared to incongruent or irrelevant) pairs of audiovisual stimuli and still bind them together (e.g., Jackson, 1953; van Wassenhove et al., 2007; Vatakis et al., 2008; Vatakis & Spence, 2007).

Note, however, that the temporal window of audiovisual integration may not always be widened by the manipulation of the semantic congruency of the component unimodal stimuli (see Vatakis et al., 2008; Vatakis & Spence, 2008). Besides, recent evidence has demonstrated that the integration of visual and auditory stimuli resulting from their temporal coincidence can be dissociated from the integration resulting from their stimulus attributes. For example, people can correctly integrate rapidly-changing visual and auditory stimuli when matching the synchrony of their onset at a faster rate than when binding specific attributes (such as matching red/green to high/low pitch, respectively; see Fujisaki & Nishida, 2010). In addition, the temporal window for moving lips and speech sounds to be perceived as simultaneous is somewhat narrower than for them to give rise to the McGurk effect (see Soto-Faraco & Alsius, 2009; van Wassenhove et al., 2007). That said, multisensory interactions in the temporal domain

may be dissociated from those in the semantic domain. These results therefore imply that multiple architectures for multisensory interactions in different domains may be needed (see Driver & Noesselt, 2008; Fujisaki & Nishida, 2010; Soto-Faraco & Alsius, 2009).

1.4.3 Direct and indirect audiovisual interactions

Human cerebral cortex is separated into visual cortex (in the occipital lobe), auditory cortex (in the temporal lobe), somatosensory cortex (in the parietal lobe) and so on, according to their function, which indicates that some brain areas are specific to one sensory modality (for a review, see Felleman & van Essen, 1991). On the other hand, some subcortical (e.g., SC; Stein & Meredith, 1993) and cortical areas (such as STS, ventral intraparietal area (VIP), and lateral intraparietal area (LIP)) are considered multisensory because information from different sensory modalities converges here (for a review, see Driver & Noesselt, 2008; Kaas & Collins, 2004). The issue concerning whether audiovisual information interacts in a direct versus indirect way involves how the audiovisual interactions are implemented in the human brain. That is, visual and auditory information may either interact directly during feedforward projections to multisensory brain areas, or these multisensory areas would further send feedback information to modulate the processing in modality-specific areas. Driver and Noesselt (2008) reviewed three possible architectures in the human brain in order to account for

how multisensory interactions may influence unimodal information processing. The first two can be considered as direct ways and the last one, an indirect way.

For the direct multisensory interactions, Driver and Noesselt (2008) summarized two new views of how to define 'multisensory areas'. In Account A for 'All multisensory', for example, the deep layer of the thalamus (receiving information from SC; see Stein & Meredith, 1993, Chapter 3), and perhaps the whole cortex as well, is considered as multisensory. The argument suggesting that the visual and auditory cortices are actually multisensory is based on the observation that there are neural projections connecting these unimodal areas (see also Cappe et al., 2009; Ghazanfar & Schroeder, 2006; Kayser & Logothetis, 2007; Macaluso & Driver, 2005; Schroeder & Foxe, 2005; see Figure 1.4A). Such direct connections may give rise to very fast audiovisual interactions after stimulus onset (e.g., within about 50 ms, see Giard & Peronnet, 1999; Molholm, Ritter, Murray, Javitt, Schroeder, & Foxe, 2002). This notion, as it suggests, is an extreme view that the whole brain is multisensory. In this case, the 'all multisensory' account might provide the neural basis of the ecological approach of multisensory interactions, since this approach holds the assumptions that the sensory and/or perceptual systems are anatomically and neurophysiologically indistinguishable (see Section 1.1). Note, however, that Driver and Noesselt suggest such direct connections are proportionally small as compared to other kinds of connections, and the function of these direct connections may be nonspecific (perhaps having a role in such

function as alerting).

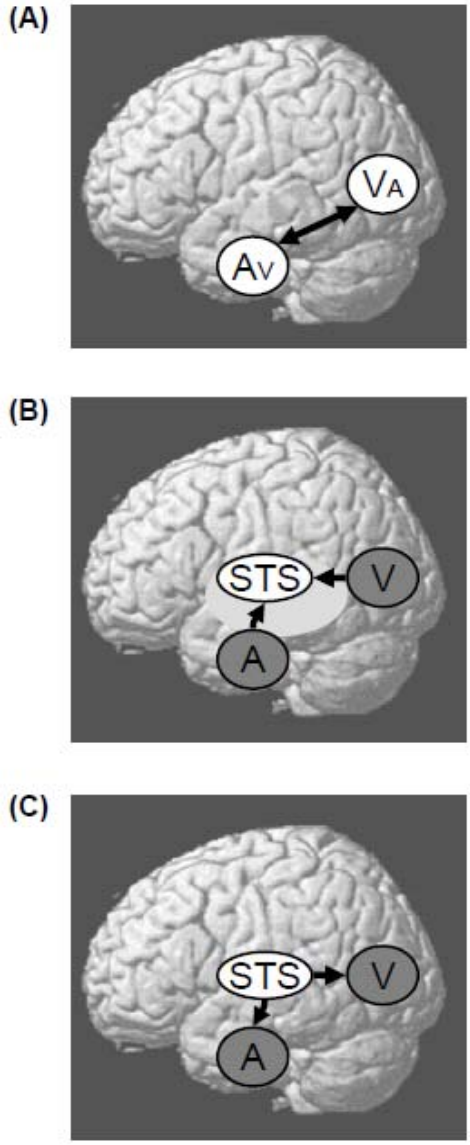


Figure 1.4. The three architectures for multisensory interaction in the human brain proposed by Driver and Noesselt (2008). The areas specified by gray and white represents modality-specific and multisensory areas, respectively. (A) Visual and auditory cortices are connected by direct neural projections, so they may be able to be considered as multisensory (labeled as V_A and A_V). (B) Information processing from modality-specific areas (V and A) to multisensory area (e.g., STS) gradually combines more multisensory information, and hence these intermediary areas are labeled in light gray. (C) After multisensory information interacts in multisensory area (e.g., STS), feedback information is sent from multisensory area to modality-specific areas.

The Account B for 'New Bimodal Brain Areas' holds a less extreme hypothesis that some areas located between modality-specific and already-recognized multisensory areas may be multisensory. That is, multisensory areas can be extended to earlier processing stages than previously thought (see Figure 1.4B). Driver and Noesselt (2008) also suggest that the increased multisensory areas may imply that the multisensory interactions can occur in several parallel and recursive loops. Nevertheless, either Account A or B may imply an important fact – namely, that multisensory processing does not go through a clear-cut unimodal stage and then a multisensory stage, but may gradually combine more and more multisensory information (see Driver & Noesselt, 2008, p. 17; see also Plaut, 2002, for a computational modeling study of audiovisual interactions).

After the sensory inputs from different modalities merge and interact in multisensory areas, these areas may then send feedback information to modality-specific areas. This notion constitutes Driver and Noesselt's (2008) Account C-F-C for 'Critical Role of Feedback Circuitry' (see Figure 1.4C). For example, Noesselt et al. (2007) showed that concurrently-presented visual and auditory stimuli (e.g., a flash and a beep) can induce higher activation in multisensory areas like STS, as well as in modality-specific areas, such as calcarine sulcus (CaS, a visual area) and Heschl's gyrus (HG, an auditory area). They further demonstrated that crossmodal facilitation observed in CaS and HG was modulated by feedback from STS using

directed information transfer measure (DIT, Hinrichs, Heinze, & Schoenfeld, 2006).

Driver and Noesselt (2008) also suggest that audiovisual interaction phenomena may involve both feedforward and feedback connections. For example, using dynamic causal modeling (DCM), Werner and Noppeney (2010a) have recently shown that the best model to explain the stronger activities elicited by simultaneously-presented audiovisual stimulation on the response of HG and CaS includes both direct connections between them and an indirect connection via STS. Driver and Noesselt further suggest that, in order to better understand the feedback modulation on modality-specific areas (and so, in turn, the connection between modality-specific areas), a disruption approach can be adopted, such as conducting transcranial magnetic stimulation (TMS) on human STS.

Nevertheless, according to these new findings, the functional name of each brain area may need to be carefully defined. If neural activity in primary visual and auditory cortex can be modulated by the sensory input from the non-preferred modality (by either a direct or indirect route), are they still sensory-specific (so that whole neocortex is multisensory as proposed by Ghazanfar and Schroeder, 2006, is supported)? Or, how are primary visual and auditory areas to be differentiated from multisensory areas (such as STS)? The response to a unimodal sensory input from the non-preferred modality provides the clue here. Werner and Noppeney (in press) have recently demonstrated that the transient onset of an auditory stimulus would induce a deactivation in the primary

visual cortex (such as CaS), while the transient onset of a visual stimulus would induce a deactivation in the primary auditory cortex (such as HG). In contrast, activity in STS did not show any deactivation to any kind of unimodal or bimodal sensory input. That said, modality-specific and multisensory brain areas may not be differentiated by their response to unimodal input from the preferred sensory versus multisensory input, but by the contrasting activity to unimodal sensory input from preferred versus non-preferred modality.

1.5 Summary

In this chapter, I have reviewed the functional importance of multisensory interactions (why) and the conditions under which multisensory interactions occur (when). Next, I summarized the perceptual outcomes of multisensory interactions (what is the result) according to the classification scheme put forward by Partan and Marler (1999). Finally, I summarized three issues concerning the underlying mechanisms of multisensory interactions (how do multisensory information interact): 1) Whether the sensory inputs are weighted according to their source modality, to their information preciseness, or to the observer's attentional selection in multisensory interactions; 2) Whether there is a unitary mechanism summing up every factor modulating multisensory interactions, or whether instead there are several mechanisms taking charge of interactions between different factors, in order to produce the final perceptual

outcome; 3) Whether multisensory interactions occur in multisensory brain areas, in modality-specific areas modulated by higher-level feedback, or both. In Chapter 2, I will address the specific aims and scope of the experiments to be reported in my thesis.

Chapter 2

Aims and Scope

The focus in this thesis is on situations in which visual perception is *modulated* by the presentation of a nonredundant auditory signal. That is, participants in the studies reported here were instructed to respond to visual targets, while they may be presented with an accessory sound. The influence of two main factors, temporal coincidence and the semantic congruency between the visual target and the accessory sound was investigated. The questions to be addressed regard *when* and *how* the presentation of a sound modulates visual perception. That is, I investigate the conditions under which audiovisual interactions occur, and what possible underlying mechanisms these results may imply.

2.1 Temporal coincidence in audiovisual interactions

In terms of temporal coincidence, I focus on how the presentation of a sound that signals the onset of a visual target can facilitate participants' visual target identification performance in Chapters 3 and 4. Note that the sound in the experiments reported in

these two chapters will only ever be informative with regards to the time of onset of the visual target, but irrelevant with regards to its identity or spatial location (except in Experiment 4.5).

2.1.1 Is temporal coincidence a sufficient condition for crossmodal facilitatory effects to occur?

Several studies have reported crossmodal facilitatory effects elicited by the presentation of a sound along with the temporal coincidence rule in visual detection and identification tasks (see Section 1.3.3.2). It should be noted that the temporal coincidence rule does not mean a very precise concurrence, but allows for the visual and auditory stimuli to be presented within a certain temporal window (Chen & Yeh, 2008; Lewald, Ehrenstein, & Guski, 2001; Lewald & Guski, 2003; Spence & Squire, 2003). Furthermore, this temporal window of simultaneity for pairs of visual and auditory stimuli may be narrowed due to the presentation of other temporally-close visual stimuli (Roseboom, Nishida, & Arnold, 2009) or expertise/knowledge concerning the event (Petrini et al., 2009; Petrini, Russell, & Pollick, 2009). Besides, the temporal window of simultaneity can also be widened or shifted when the visual and auditory stimuli happen to be presented from further away (Engel & Dougherty, 1971; King, 2005; Kopinska & Harris, 2004; Sugita & Suzuki, 2003; though see Arnold, Johnston, & Nishida, 2005; Stone et al., 2001).

With regards to the question of whether temporal coincidence is sufficient for crossmodal facilitation to occur, one may wonder whether this factor modulates visual perception in a purely bottom-up fashion, or it needs to further elicit certain cognitive factors to help modulate visual perception. For example, Olivers and van der Burg (2008) presented a sound randomly with one of 11 items (two targets, called T1 and T2, and 9 distractors) in the RSVP stream in their multisensory study of the AB. Hence, the sound had a validity of 18% that it would accompany either T1 or T2. They found that although the sound was only occasionally presented simultaneously with T2, it could still facilitate the identification of T2 and hence reduce the magnitude of the AB (as compared to the average of six conditions where the sound was presented with visual distractors). Hence, Olivers and van der Burg suggested that the crossmodal facilitation elicited by the presentation of a simultaneous sound is, at least partly, bottom-up.

On the other hand, considering the daily example of thunder and lightning, although they always happen close together in time, people tend to consider them as two separate events since their onset time with respect to each other is changeable and their temporal asynchrony is noticeable. Lippert et al. (2007) presented a sound randomly at one of five different stimulus onset asynchronies (SOAs) relative to the onset of the visual target (0, ±150, and ±400 ms), which gives rise to an audiovisual co-occurrence rate of 20%.[9] However, they did not find any facilitation at the 0 ms SOA as compared

[9] Note that there were sound absent trials with equal probability as each SOA condition in Lippert et al.'s

to the sound absent condition. Lippert et al. therefore suggested that the crossmodal facilitation resulting from the presentation of a simultaneous sound actually reflected a 'cognitive top-down effect' rather than simply a bottom-up effect. That is, the temporal coincidence of visual and auditory stimuli may not be sufficient (though it is necessary). What is also needed is that the crossmodal stimuli have to *reliably co-occur* together. Note that Lippert et al. only reported the null results between the 0 ms SOA and the sound absent condition, while the results at the other SOAs were omitted.[10] Hence, it is unclear that whether the participants processed the sound but the sound did not integrate with the visual target (so that the sound would probably induce an alerting effect when it is presented before the visual target), or else the participants simply ignored the sound since the time of its onset was uninformative with respect to the visual target.

In those situations demonstrating auditory dominance effects (such as temporal ventriloquism or the sound-induced flashing illusion), it has been shown that the sound can modulate visual perception regardless of whether it is presented at either same or different location with respect to the visual stimulus (see Innes-Brown & Crewther, 2009; Vroomen & Keetels, 2006). However, Ngo and Spence (2010b, c) have recently demonstrated that the largest crossmodal facilitation effects are induced by the

(2007) study. Hence, the validity of the sound should be considered as only 16.7%. On the other hand, there was no sound absent condition in Olivers and van der Burg's (2008) study. Besides, a critical difference is perhaps the fact that the sound only accompanied a visual stimulus (i.e., the target) in 20% of trials when it presented in Lippert et al.'s study, while the sound always accompanied a visual stimulus (either target or distractor) in Olivers and van der Burg's study.

[10] Direct requests for this data from the authors have not been successful (personal communication on 08/04/2010).

presentation of a sound which is both spatially and temporally informative with regards to the visual target. The latter results therefore once again imply that the temporal coincidence of visual and auditory stimuli may be a necessary, but not a sufficient, condition for crossmodal facilitatory effects to be observed.

In this thesis, I will address the issue of whether temporal coincidence is a completely bottom-up factor or whether instead it also serves as a top-down modulatory factor (see also Lippert et al., 2007; Ngo & Spence, 2010c). Furthermore, I will compare the crossmodal facilitatory effect between conditions in which the sound is simply temporally coincident with the visual target versus when the sound is temporally as well as spatially coincident with the visual target.

2.1.2 How does the presentation of a simultaneous sound enhance visual perception?

The presentation of a sound can enhance visual perception via several possible mechanisms and at various different levels of visual information processing. These include: alerting, signal enhancement, temporal uncertainty reduction, stronger multisensory object representation, and temporal attention capture (see Table 2.1 for a summary).

The presentation of a simultaneous sound may simply alert a participant in a behavioural experiment and thus speed-up his/her responses (Nickerson, 1973). Note

that the alerting effect has primarily been demonstrated in speeded tasks that researchers believe mainly to facilitate the execution of a participant's response, rather than necessarily improving the sensitivity or accuracy of their visual perception (see Nickerson, 1973, p. 499).

Table 2.1. Five possible mechanisms of crossmodal facilitation elicited by the presentation of a simultaneous sound on visual perception.

Hypothesis	**The effect of a sound on a participant's perception of a visual target**	**The predicted effect on performance**
Alerting	The sound prepares the participant to respond to any stimulus, including to a visual target (Nickerson, 1973)	An overall improvement in the peed of visual detection and identification performance. Perhaps no beneficial effect on accuracy or sensitivity measures
Signal enhancement	Intersensory summation of the visual signal and the auditory signal (Bernstein et al., 1970; Stein et al., 1996)	An improvement in visual detection and identification performance, which can be seen in the threshold of the psychometric function decreasing when the sound is presented
Temporal uncertainty reduction	The onset of the auditory stimulus signifies the timing of the visual stimulus (Lippert et al., 2007)	An improvement in visual detection and identification performance. The threshold as well as the slope of the psychometric function will decrease when the sound is presented (Pelli, 1985)
Stronger object representation	The visual and auditory inputs give rise to a multisensory object representation that is more consolidated than a unimodal one (Chen Y.-C. & Yeh, 2009)	Crossmodal facilitation occurs when the visual and auditory stimuli can be segregated from other stimuli presented within the same modality
Attentional capture	People's attention can be directed to the location of the visual target by the presentation of the sound (Talsma et al., 2010)	Crossmodal facilitation occurs when the auditory stimulus (or cue) is distinct or salient from other auditory stimuli (Spence & Ngo, in press)

The other four possible mechanisms reflect different ways in which audiovisual interactions can occur. For example, the signal of the visual and auditory input may sum

together and therefore both the visual and auditory information are enhanced (Bernstein, Rose, & Ashe, 1970; Hershenson, 1962). This hypothesis is supported by Stein et al.'s (1996) study showing that participants rate a light as looking brighter when it is accompanied by a sound than when it is not.[11] When holding this hypothesis, however, one needs to further consider the potential influence by the background noise or other distracting events in the visual modality (such as any visual masking). That is, the presentation of a sound is likely to enhance the neural signal associated with the visual target as well as the accompanying noise or visual distractors.

On the other hand, since the synchronous auditory stimulus signifies the onset time of the visual target, the temporal uncertainty with regard to when in time the visual target occurred may plausibly be reduced (Lippert et al., 2007).[12] Observers therefore only need to monitor the visual information on which their decision will be based within a specific time window that is cued by the onset of the sound. As a result, the estimated visual sensitivity will be higher when the participants' temporal uncertainty regarding

[11] Note that when the participants' response criterion was well controlled, at least four studies have failed to replicate this result (see Arieh & Marks, 2008; Lippert et al., 2007; Marks, Ben-Artzi, & Lakatos, 2003; Odgaard, Arieh, & Marks, 2003). Therefore, Stein et al.'s (1996) results should be attributed to participants' response bias rather than to a genuine audiovisual integration. On the other hand, the presentation of a light seems to genuinely increase the perceived loudness of a simultaneously-presented sound (Lovelace, Stein, & Wallace, 2003; Odgaard, Arieh, & Marks, 2004; though see Marks et al., 2003).

[12] Anderson and Mamassian (2008) have suggested that the onset of the auditory signal may enhance the transient information of the visual target. That is, the participants can better detect any change of the visual target irrespective of the content of the change (e.g., luminance increase or decrease in their study). They called this the *transient hypothesis*. Accordingly, the transient hypothesis predicts that the crossmodal facilitatory effects can only be observed in a visual detection task but not in a visual discrimination or identification task. Therefore, the transient hypothesis is not listed here since it cannot provide a comprehensive explanation. Even so, this hypothesis is still tested in Chapter 3.

the visual target is reduced (Earle & Rowe, 1971). Note that, in this case, the visual and auditory information may not necessarily integrate. The signal enhancement hypothesis and temporal uncertainty reduction hypothesis can be tested by measuring the psychometric function of visual target sensitivity. That is, the reduction of the visual target threshold by the presentation of a simultaneous sound will indicate whether the visual signal has been enhanced. In the meantime, any reduction of slope together with the reduction of target threshold in the psychometric function will imply that the presentation of a simultaneous sound also reduced the participants' temporal uncertainty regarding the visual target (Pelli, 1985; Yu, Klein, & Levi, 2002).

Considering the situation in which a visual and an auditory stimulus are inferred as originating from the same external event, they may be integrated in order to construct a multisensory object representation (e.g., Ernst, 2007; Shams & Beierholm, 2010). A multisensory object representation is likely to be more consolidated in memory for responding as compared to the purely unimodal one, since the former contains more sensory information (e.g., Chen Y.-C. & Yeh, 2008, 2009). This hypothesis may assume that crossmodal facilitation occurs at a higher level than suggested by the two previous hypotheses, since a complete object representation needs to bind various signals (say, at least its space, time, and identity information, see Bowman & Wyble, 2007; Chun & Potter, 1995; Fujisaki & Nishida, 2010; Kanwisher, 1987).

The final hypothesis holds that crossmodal facilitation is mediated by the boosting

of attention (for reviews, see Spence & Ngo, in press; Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010). That is, the presentation of the sound can increase the saliency of the visual target event (see Noesselt et al., 2008), and it is this salient event that then captures a participant's attention (both temporally and spatially). The beneficial effects of attention, in turn, enhance a participant's representation of the visual target (or multisensory event; e.g., Talsma, Doty, & Woldorff, 2007; van der Burg, Talsma, Olivers, Hickey, & Theeuwes, submitted). Note that a question remains as to whether this attentional effect is in fact elicited by the sound *per se* (so the sound and the visual target may be separately represented) or by the multisensory event (so the sound and the visual target need to be integrated first; Talsma et al., 2010).

It is important to note that these five possible mechanisms are not mutually exclusive. However, under certain conditions, one of them may work alone or two (or more) of them may work together. The experiments reported in this thesis were designed to investigate which mechanism(s) provide(s) the best account for the crossmodal facilitatory effects observed in visual detection and identification tasks, respectively.

2.1.3 Outline of the studies on temporal coincidence

In Chapter 3, a psychophysical study using equivelant noise paradigm is reported. This study was designed to investigate whether the presentation of a simultaneous

sound modulates participants' sensitivity to a visual target in a dual detection and discrimination task. Any response bias elicited by the presentation of a sound was controlled for as far as possible. By comparing the psychometric functions in the sound present and sound absent conditions, it is possible to test whether the crossmodal facilitation by the presentation of a simultaneous sound, when it occurs, involves any reduction of temporal uncertainty. That is, I aim to provide evidence to suggest that the audiovisual facilitation on visual perception by the presentation of a simultaneous sound cannot be attributed to either alerting or temporal uncertainty reduction (i.e., ruling out hypotheses 1 and 3 in Table 2.1).

In Chapter 4, I investigate whether and how the presentation of a sound can facilitate the perception of a visual target when it is embedded amongst visual masks or distractors. The visual backward masking paradigm is utilized in order to investigate the crossmodal facilitation elicited by a simultaneous sound in a visual identification task. Task difficulty is easily controlled in this paradigm and the mechanisms underlying the phenomenon have been well studied (see Breitmeyer & Ogmen, 2006). Here, I investigate whether the crossmodal facilitation effect can be observed when the visual target has to compete with a physically similar mask, and further, whether the facilitation occurs at a 'lower' sensory level or at a 'higher' cognitive level. In summary, the signal enhancement and stronger object representation hypotheses will be tested in Chapter 4.

2.2 Does semantic congruency influence audiovisual interactions?

In Chapters 5 to 8, the constraints on (and the very existence of) audiovisual semantic congruency effects are explored. That is, I investigate whether the processing of a visual stimulus can be influenced by the presentation of a semantically-congruent auditory stimulus (vs. an incongruent stimulus or noise). In the literature, several different kinds of visual and auditory stimulus correspondence have been given the label semantic congruency, such as synaesthetic associations (e.g., high pitch and high spatial elevation, Evans & Treisman, 2010; O'Leary & Rhodes, 1984; Parise & Spence, 2009; see Spence, submitted, for a review), speech sounds and lips movements (Ross et al., 2007; Vatakis et al., 2008; Vatakis & Spence, 2007), naturalistic sounds and pictures or videos (such as the sound of 'woof' and the picture of a dog or video of dog barking, see Iordanescu et al., 2008, 2010; Koppen, Alsius, & Spence, 2008; Noppeney, Ostwald, & Werner, 2010; Schneider, Engel, & Debener, 2008; Vatakis & Spence, 2008), and spoken words and printed words or colour patches (such as '/rɛd/' and 'RED', see Laurienti et al., 2004; Roelofs, 2005; Shimada, 1990; Stuart & Carrasco, 1993). In this thesis, I will primarily focus on those interactions taking place between pictures and naturalistic sounds (and spoken words which provide a point of comparison).

2.2.1 When do audiovisual semantic interactions occur?

According to the unity assumption, semantically-congruent stimuli presented in the

visual and auditory modalities should originate from a single object (or event), Vatakis and her colleagues have shown that the temporal window of audiovisual integration is significantly wider for congruent pairs of heard syllables and seen moving lips than for incongruent stimulus pairs (Vatakis et al., 2008; Vatakis & Spence, 2007; see also van Wassenhove et al., 2007). However, no such congruency effect has been reported for other kinds of naturalistic audiovisual events (e.g., watching a bouncing ball accompanied by either the sound of a bouncing ball or the smashing of ice by a hammer, Vatakis & Spence, 2008).

It should be noted that the integration of semantically-congruent stimuli from different sensory modalities into a unified object representation may perhaps constitute a special case of multisensory interaction. Considering, for example, semantic congruency effects in the visual modality always occur between objects, such as between a picture and a word (e.g., Glaser & Glaser, 1989) or between an object and its background (e.g., Biederman, 1972; Davenport & Potter, 2004; Palmer, 1975). Hence, it seems reasonable to assume that crossmodal semantic congruency effects can occur between *two* objects (or events) presented in different sensory modalities while not being integrated into a unified representation (cf. Welch & Warren, 1980). In fact, researchers have also adopted classic experimental paradigms, such as the semantic priming paradigm (Schneider, Engel, & Debener, 2008) or the Stroop-like interference paradigm (Roelofs, 2005; Shimada, 1990; Stuart & Carrasco, 1993; Yuval-Greenberg &

Deouell, 2009), in order to investigate semantic congruency effects between visual and auditory stimuli.

When investigating audiovisual semantic interactions, a fundamental issue should be borne in mind. That is, visual stimuli access their corresponding semantic representations more rapidly than auditory stimuli. For example, it has been estimated that the meaning of a visually-presented picture can normally be accessed within about 100 ms of its presentation (Greene & Oliva, 2009; Potter, 1975; Thorpe, Fize, & Marlot, 1996). However, it typically takes longer to process sounds since the meaning of a sound evolves over time (e.g., Aramaki, Marie, Kronland-Martinet, Ystad, & Besson, 2010; Ballas, 1993; De Lucia, Clarke, & Murray, 2010; Murray, Camen, Gonzalez Andino, Bovet, & Clarke, 2006). Hence, if a picture and a naturalistic sound are presented simultaneously, the influence from vision to audition should be larger or occur earlier in time than *vice versa* (see Jesse & Massaro, 2010; Yuval-Greenberg & Deouell, 2009). Consequently, when addressing the question of whether audiovisual semantic congruency would occur, one should take the SOA between visual and auditory stimuli into consideration as well.

2.2.2 How do pictures and meaningful sounds interact?

Given that visual processing is dynamic, a sound presented at different SOAs should influence visual picture processing in different ways (see Figure 2.1). For

example, when a sound is presented prior to the onset of a visual target, it seems to be able to induce a semantic priming effect (e.g., Schneider, Engel, & Debener, 2008). When the picture and the sound are presented simultaneously, it is though plausible that the meaning of the sound can enhance visual feature processing or semantic access if it happened to be accessed rapidly enough (Iordanescu et al., 2008, 2010; Stuart & Carrasco, 1993). Alternatively, however, the sound may only influence information processing at a post-perceptual stage (such as at the stage of memory encoding, decision masking or response execution, see Roelofs, 2005; Suied et al., 2009).

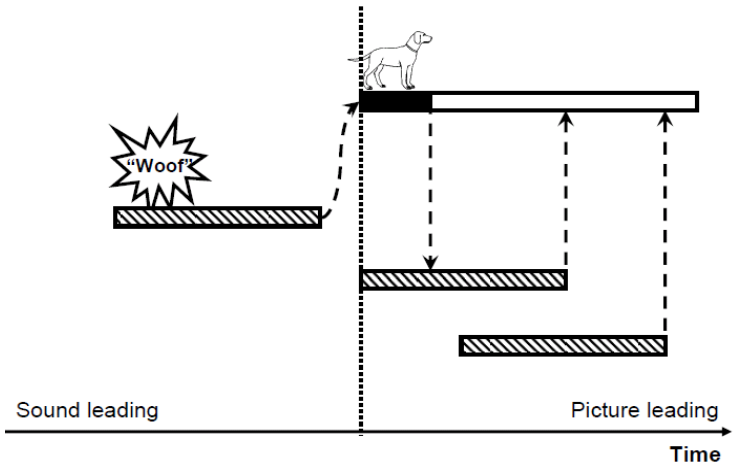


Figure 2.1. A schematic representation of the semantic interactions taking place between a picture (a dog) and a sound (a dog's woofing) at different SOAs. The meaning of a picture can be accessed rapidly (the black band) and then go on to post-perceptual processing (such as memory encoding, decision making, or response execution, the white band). However, the meaning of the sound takes longer to access (the striped band). Hence, when the onset of sound leads the picture by a sufficient time, the meaning of the sound can prime the processing of the picture. When the sound and the picture are presented simultaneously, the meaning of the picture may be able to modulate auditory processing, while the meaning of the sound may only be able to modulate post-perceptual picture processing. However, when the onset of the picture leads the sound, the sound should only modulate post-perceptual picture processing.

Naturalistic sounds and spoken words correspond to non-verbal and verbal auditory stimuli, respectively. The association between naturalistic sounds and visual objects (referring to pictorial rather than linguistic stimuli here) is learned from prior experience in our daily environments, so those universal associations (like a dog and its woofing) can be considered as cross-cultural symbols shared by all humans. Consider the cases that when foreign tourists cannot find a word to express themselves, they have to mimic varieties of naturalistic sounds! On the other hand, the correspondence between spoken words and visual objects is arbitrarily assigned by each language system (though see Berlin, 1994).

Naturalistic sounds and spoken words are commonly used to refer to the same object. The identification of naturalistic sounds and spoken words are both modulated by their familiarity and uncertainty (Ballas, 1993). Furthermore, evidence from neuropsychological studies has demonstrated that the impairment of understanding naturalistic sounds and spoken words are strongly associated in aphasic patients (e.g., Dick et al., 2007; Saygin, Dick, Wilson, Dronkers, & Bates, 2003). One may therefore wonder whether naturalistic sounds and spoken words actually go through a similar process, such as the notion that the identification of naturalistic sounds is mediated by verbalization (Bartlett, 1977) or the processing of naturalistic sounds *per se* should be considered as a form of linguistic processing (Ballas & Howard, 1987).

Dick, Bussiere, and Saygin (2002) explicitly asked the participants in their study to

process naturalistic sounds with or without subvocal naming, and they found that the subvocal naming of naturalistic sounds requires extra time to elicit a response. That is, the naming of a naturalistic sound is *not* necessary to identifying it. I will further tackle this issue using implicit experimental designs. In addition, I will utilize the general reading-naming interference model put forward by Glaser and Glaser (1989) in order to explain the interactions between pictures and naturalistic sounds and spoken words.

2.2.3 Outline of the studies on audiovisual semantic congruency

In Chapter 5, I explore whether the semantic context provided by an auditory soundtrack can modulate the perceptual outcome when the visual system is presented with a bistable figure. Here I use the binocular rivalry paradigm in which two static figures presented to each eye can give rise to a dynamic alternating percept (Alais & Blake, 2005). If the auditory context can be shown to modulate visual perception in terms of its semantic congruency, observers should report the percept that happens to be congruent with the sound that they are hearing as dominant for more of the time.

From Chapter 6 to 8, I will be primarily interested in how humans' identification the visual objects at the basic semantic level (such as animal vs. non-anaimal, or dog, cat, car, and so on) within the first few hundreds of milliseconds after its onset by the presentation of semantically-congruent or incongruent sounds. Hence a static visual picture that was presented briefly and then immediatedly masked was used as the visual

stimulus. Such brief presentation can be compared to situation of the first glance of a visual object.

In Chapter 6 and 7, I examine crossmodal semantic interactions between pictures and both naturalistic sounds and spoken words using a picture categorization task and a picture detection task, respectively. I will also test for the SOAs at which the auditory stimulus can effectively modulate visual picture perception. In these two chapters, I will mainly compare the conditions in which the sounds are presented either earlier or simultaneously with the visual pictures.

The experiments to be reported in Chapter 8 investigate whether the meaning of a sound can still modulate visual perception when a naturalistic sound is presented simultaneously with, or following, the onset of a target picture. The pictures in these experiments were briefly presented and then rapidly masked, and the participants had to name the pictures. On the basis of the results of the experiments reported in Chapters 6 through 8, I will be able to determine whether presenting a sound at different times with respect to the onset of the picture would influence the different stages of visual picture processing.

Chapter 3

The presentation of a synchronous sound enhances visual sensitivity without reducing target uncertainty

3.1 Introduction

Human perceptual systems, especially the early visual system, used to be considered in terms of individual processing modules (e.g., Fodor, 1983; Pylyshyn, 1999). This view was based, at least in part, on the fact the cerebral cortex is anatomically and functionally divided into discrete sensory areas, such as the well-known primary visual cortex, primary auditory cortex, and so on (see Felleman & van Essen, 1991). Recently, however, fMRI studies have demonstrated that activity in primary visual cortex can be modulated by the presentation of a sound at about the same time (Noesselt et al., 2007; Watkins, Shams, Josephs, & Rees, 2007; Werner & Noppeney, 2010a). This neurophysiological evidence raises the possibility that auditory signals can enhance the early stages of visual information processing, either via direct

neural connections between modality-specific brain areas and/or via feedback from multisensory brain areas (see Driver & Noesselt, 2008; Ghazanfar & Schroeder, 2006, for reviews). However, it is still unclear just how such crossmodal interactions contribute to visual perception in humans. In this chapter, the possible mechanisms underlying the crossmodal facilitation of visual target perception elicited by the presentation of a synchronous sound are investigated using psychophysical techniques.

In human behavioural studies, crossmodal facilitatory effects have been demonstrated using various experimental paradigms (see Section 1.3.3.2). The mounting evidence currently suggests the crossmodal facilitation of visual perception elicited by the synchronous presentation of a sound to be a robust empirical phenomenon, one that should extend beyond simplified experimental environments (cf. de Gelder & Bertelson, 2003).

Researchers have also tried to understand the possible mechanisms underlying the crossmodal facilitation elicited by the presentation of a synchronous sound on visual perception. Early studies suggested that the signals of the simultaneously-presented visual and auditory inputs were integrated, such that both visual and auditory signals were mutually enhanced. This is known as the *signal enhancement hypothesis* (Bernstein et al., 1970; Hershenson, 1962). Evidence apparently supporting this hypothesis was provided by Stein et al. (1996). They reported that participants rated a light as appearing brighter when accompanied by a sound than when presented alone.

On the other hand, recent studies have demonstrated that the transient provided by the onset of a concurrent sound can sometimes be critical to eliciting crossmodal facilitation, no matter whether the energy (i.e., the amplitude) of the sound happens to consistently increase or decrease in time with the occurrence of the visual target (e.g., Anderson & Mamassian, 2008; Noesselt et al., 2008). Given that abrupt change is a stimulus feature that extends across modalities, a synchronous abrupt change in the auditory modality is likely to enhance participants' visual change detection performance (see Downar, Crawley, Mikulis, & Davis, 2000). This is known as the *transient boosting hypothesis* (Anderson & Mamassian, 2008). However, the transient information only provides information regarding the onset of the visual stimulus, but no information regarding the identity of the visual target. Accordingly, the transient boosting hypothesis predicts that only human visual detection performance, rather than human visual discrimination (or identification) performance, can be enhanced by the presentation of the sound (see Anderson & Mamassian, 2008). Furthermore, since the synchronous auditory stimulus also signifies the onset time of the visual target, the temporal uncertainty of the participants' perception with regards to when the visual target occurred might plausibly be reduced. As a result, an observer might be better able to sample the information concerning the visual target when the uncertainty is reduced. This should facilitate both visual detection and discrimination (or identification) performance. This is known as the *uncertainty reduction hypothesis* (see Anderson & Mamassian, 2008; Lippert et al.,

2007).

Note, however, that the presentation of a synchronous sound can also induce a more liberal response bias by participants with respect to the visual target than when no sound is presented (e.g., Bolognini et al., 2005; Frassinetti et al., 2002; Lippert et al., 2007). That said, Stein et al.'s (1996) results, though dramatic, have been criticized because of the potential confounding effect of response bias on participants' performance elicited by the presentation of the sound on a subset of the trials. When other researchers controlled the response criterion of participants, they were unable to replicate the finding that the presentation of a synchronous sound modulated perceived brightness of a visual target (see Arieh & Marks, 2008; Marks et al., 2003; Odgaard et al., 2003).

The experiment to be reported in this chapter (Experiment 3.1) was designed to investigate whether crossmodal facilitation of visual target performance elicited by the presentation of a synchronous sound could be observed, and which of the hypotheses (signal enhancement, transient boosting, or uncertainty reduction) can provide the best explanation. A two-alternative forced-choice (2-AFC) procedure was utilized in order to eliminate the influence of any response bias that might have been induced by the presentation of a sound (see Section 3.2.3). Considering that people's performance in visual detection and discrimination tasks can provide a critical clue with which to differentiate the transient boosting hypothesis from the uncertainty reduction hypothesis

(see Anderson & Mamassian, 2008), both tasks were included in the present study.

The visual target, a tilted Gabor patch, was embedded within random dot noise at one of seven different contrast levels (see Dosher & Lu, 1998; Lu & Dosher, 1998). The signal contrast thresholds required to detect the onset of the Gabor patch and to discriminate its orientation at each noise level were estimated. According to the signal enhancement hypothesis, the threshold contrast for the target Gabor should be lower in the sound present condition than in the sound absent condition. According to the uncertainty reduction hypothesis, the threshold should be lowered and be accompanied by a change in the slope of the psychometric function. The above two hypotheses should predict that the crossmodal facilitatory effect can be observed in both detection and discrimination tasks; in contrast, according to the transient boosting hypothesis, crossmodal facilitation should only be observed in the detection task.

3.2 Methods

3.2.1 Participants

The observers consisted of the two experimenters (YCC and PCH) and two naïve participants (CWC and YYH). The naïve participants were paid and informed of their rights according to the ethical standards laid down in the 1990 Declaration of Helsinki. All of the participants had normal or corrected-to-normal vision and hearing by self-report.

3.2.2 Apparatus and stimuli

The visual stimuli were presented on a 17 inch Eizo CRT monitor with a resolution of 1024*768 pixels and a frame rate of 120 Hz. The monitor was controlled by a personal computer with a VSG 2/5 graphics card (Cambridge Research Systems Ltd, UK) with pseudo 15 bit contrast resolution. The monitor was gamma-corrected using luminance measurements with an optical photometer interfaced with the VSG display calibration software (Cambridge Research Systems). The auditory stimuli were presented from a pair of loudspeaker cones, one located on either side of the monitor and aligned with the center of the display in terms of their elevation. The participants were tested in a dimly-lit experimental chamber.

The visual target consisted of an achromatic Gabor patch. The Gabor consisted of a 2D grating multiplied by a Gaussian envelope, tilted either 10-degree clockwise or 10-degree counterclockwise of vertical. The Gabor was defined by the equation:

$$L(x,y,\theta) = L_0 + L_0 \times c \times \cos(2\pi f(x\sin\theta + y\cos\theta) + \rho) \times \exp(-\frac{x^2+y^2}{2\sigma^2})$$

Where L_0 indicates the mean luminance, c the contrast of the Gabor, defined as the Michelson contrast of the grating, f the spatial frequency of the grating, θ the orientation of the grating relative to the vertical, σ the standard deviation of the Gaussian envelope, and ρ the phase of the stimuli with respect to the center of the Gaussian window. The viewing distance was set at 57 cm, and the spatial frequency was 1 cpd with 0 degree

phase angle, the space constant (σ) was 0.8 degrees and consequently the bandwidth of the Gabor was about 0.65 octaves.

The visual noise stimuli consisted of random pixel gray dots sampled from an identical Gaussian distribution with a mean of 128 and variances of 38 ranging from 0 to 255. Seven noise energy levels were chosen by setting different levels of the Michelson contrast (0, 0.02, 0.04, 0.08, 0.12, 0.16, 0.25, equivelent to -∞, -34, -28, -22, -18, -16, -12 dB[13]) of the sampled noise image. The size of noise element was 0.2 x 0.2 degrees of visual angle and the size of the noise frame was 11.4 x 11.4 degrees. The mean luminance of the display background was 50 cd/m^2.

The auditory stimulus consisted of a 17 ms 1000Hz pure tone (including 3 ms fade-in and fade-out). The sound was presented at a sound pressure level (SPL) of 60 dB from free-field loudspeakers. The onset of the Gabor and the sound was measured by an oscilloscope in order to ensure that their onsets were synchronized.

3.2.3 Design and procedure

Three factors, sound (present or absent), noise level (seven levels), and task (detection vs. discrimination) were manipulated. There were two display sequences in each trial, one of which contained the target Gabor (see Figure 3.1). The participants performed a dual 2-AFC task: First, they had to judge which sequence contained the Gabor (first or second, by pressing the left or right arrow key on the keyboard,

[13] dB=20*log(*c*), where *c* is each noise contrast level.

respectively). Second, they had to judge in which direction the Gabor was tilted (clockwise or counterclockwise, by pressing left-arrow or right-arrow key again, respectively). The participants were instructed to respond accurately but not speedily. Two contrast thresholds, the detection of the Gabor and the discrimination of Gabor orientation, were estimated for each noise level in the sound-present and sound-absent conditions. The method of constant stimuli with 7 different signal contrasts was used to generate the psychometric functions. The pure tone was presented in both display sequences in the sound-present condition, but in neither sequence in the sound absent condition. The presentation of the sound therefore provided no information whatsoever with regards to either the detection or discrimination responses (cf. Anderson & Mamassian, 2008).

Each trial consisted of three parts: Interval 1, interval 2, and the response. At the beginning of each interval, a fixation (green dot, 0.25° x 0.25°, 100ms) and a blank field with mean luminance that was the same as the background (200 ms) were presented. In order to manipulate the contrast of the Gabor and the noise independently, an interlaced presentation technique was used in each interval: In a 10 frame sequence (8 ms for each frame at a refresh rate of 120 Hz), a noise was presented in the first, third, seventh, and ninth frames. Either the target Gabor or a blank screen was presented in the sixth frame, leaving blank the remaining frames. Given that the 120 Hz refresh rate of the monitor is beyond the temporal resolution of human visual system (see Kelly, 1974; Robson,

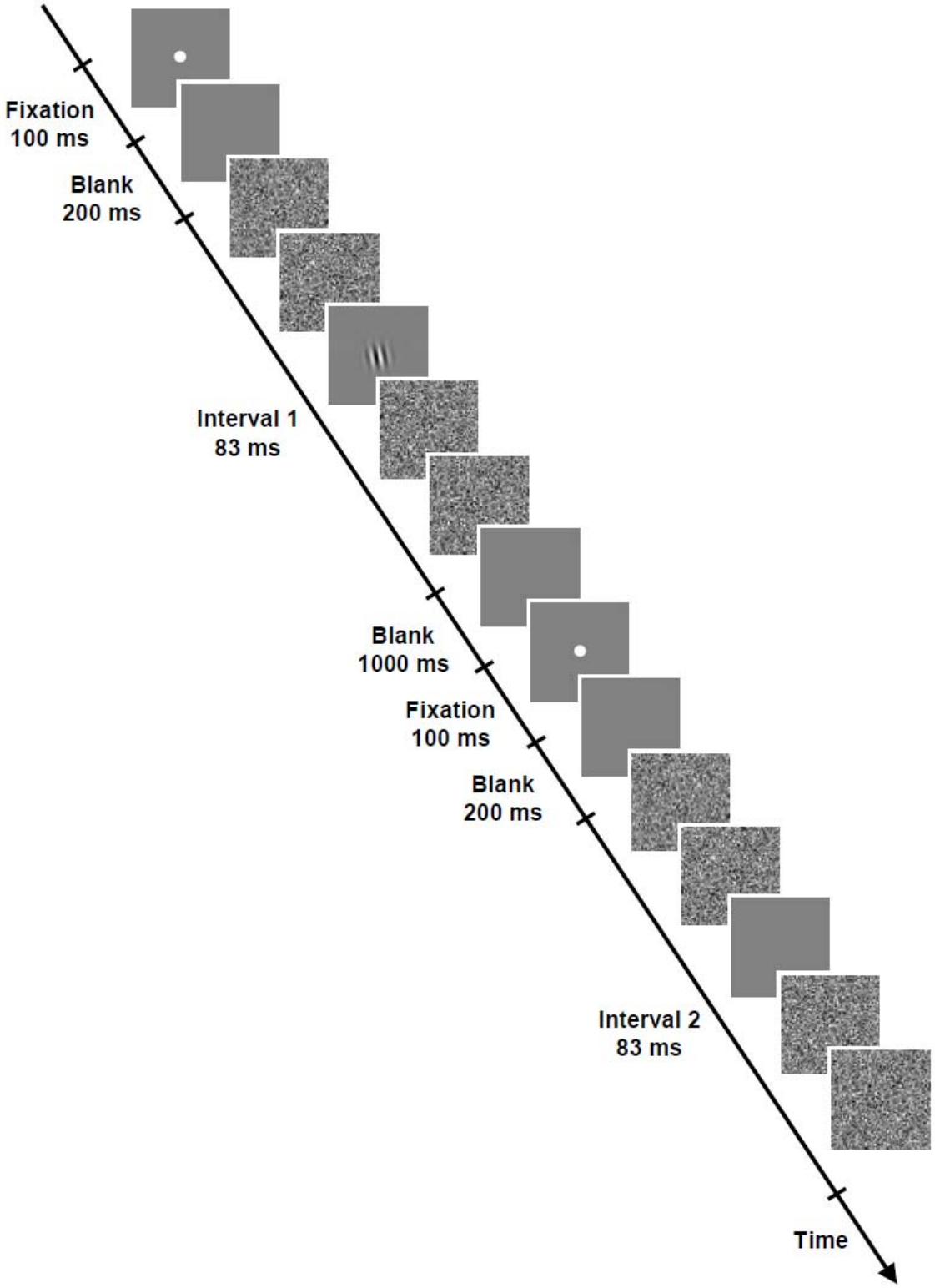


Figure 3.1. The sequence of stimulus displays presented in a trial in Experiment 3.1. Two intervals contained four noise frames and either a visual target (a Gabor patch) or a blank frame. Note that size and time are not represented to scale in the figure. The participants performed a dual 2-AFC task, detecting which interval contained the target Gabor, and discriminating the orientation of the tilted Gabor. In this example, the participant would have had to report 'first interval' and 'counterclockwise'.

1966), people will only have perceived the noise and the target Gabor (providing that they were above threshold) but not the blank frames. In addition, the perceived stimulus contrast of the noise and target Gabor should be halved to reflect the temporal integration of the target (or the noise frame) and adjacent blank frame.

A blank frame with a duration of 1000 ms segregated the two display sequences. After the participants responded using the keyboard placed in front of them, the next trial started 1000 ms later. All of the conditions (7 target contrasts x 7 noise contrasts x 2 sound conditions) were tested once in each block. The order of these trials was randomized. The block was run 40 times. Testing was divided into six approximately one hour testing sessions conducted on different days.

3.3 Results

For each participant, the threshold vs. noise (TvN) functions in the sound present and absent conditions for both detection and discrimination performance were computed. At each noise level, the percentage correct in the detection and orientation discrimination tasks at each signal contrast level were fitted with a logistic function using a bootstrapping procedure with the following equation (see Wichmann & Hill 2001a, b):

$$\Psi(x) = \gamma + (1 - \gamma - \lambda) \times \left(\frac{1}{1 + \exp\left(\frac{\alpha - x}{\beta}\right)}\right)$$

Where γ is the guessing rate (0.5), and λ is the lapsing rate which is set between 0 and

0.05. α is related to the threshold and β is related to the slope of the psychometric function. The threshold signal contrast was set at 75% correct. The signal contrast was then plotted as a function of the noise contrast in terms of dB values (i.e., the TvN function, see Figure 3.2).

3.3.1 Results of the grouped data

The threshold of signal contrast in each condition was submitted to a three-way analysis of variance (ANOVA) with the factors of sound (present vs. absent), noise level (7 levels), and task (detection vs. discrimination). The three main effects were all significant: The signal contrast threshold was lower in the sound present than in the sound absent condition ($F(1,3) = 9.50$, $MSE = 0.74$, $p = .05$). The threshold increased with increasing noise levels ($F(6,18) = 91.80$, $MSE = 4.60$, $p < .0001$). A *post-hoc* Tukey test revealed that the signal contrast thresholds were similar at the first four noise levels (i.e., -∞, -34, -28, -22 dB noise levels). They were all significantly different from those in the remaining three noise levels (i.e., -18, -16, -12 dB noise levels, all ps $< .01$). The signal contrast thresholds at the -18, -16, -12 dB noise levels all differed from each other (all ps $< .05$). The threshold was significantly lower in the detection than in the discrimination task ($F(1,3) = 44.83$, $MSE = 2.36$, $p < .01$).

Two two-way ANOVAs (sound x noise level) were then conducted on the signal contrast threshold in the detection and discrimination tasks, respectively. In the

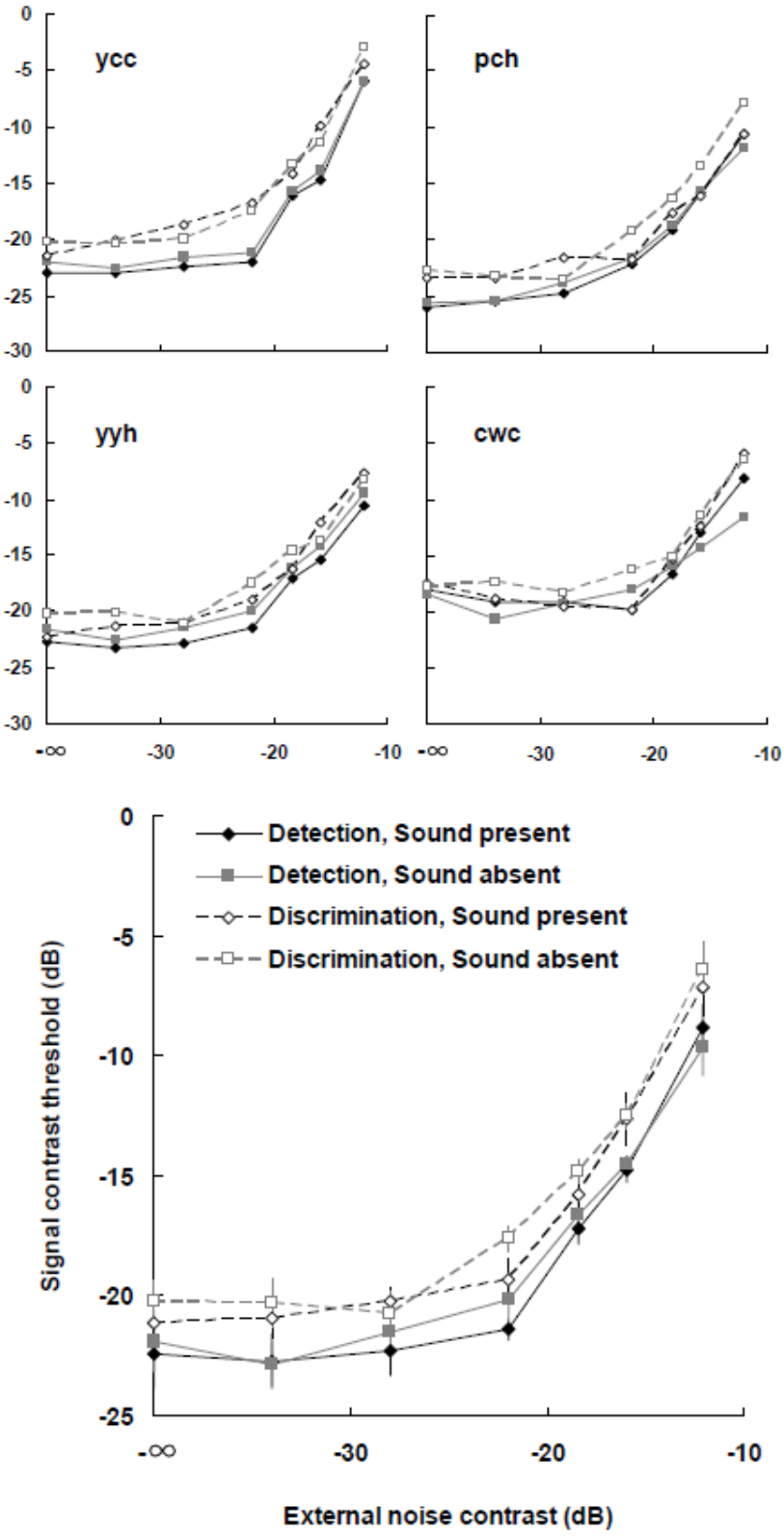


Figure 3.2. The results of Experiment 3.1. The signal contrast threshold was plotted as a function of the noise contrast level using a dB scale (i.e., a TvN function; see Lu & Dosher, 1998). The upper panels represent the results for the four individual participants, while the lower panel represents the average results for the four participants. The error bars indicate ±1 standard error of the means.

detection task, the results highlighted a significant main effect of noise level ($F(6,18) = 57.59$, $MSE = 3.51$, $p < .0001$). The interaction between sound and noise level was also significant ($F(6,18) = 3.13$, $MSE = 0.29$, $p < .05$). The simple main effect showed that the signal contrast threshold was lower in the sound present condition than in the sound absent condition only at the fourth noise level (i.e., -22 dB, $F(1,21) = 5.45$, $MSE = 0.52$, $p < .05$). In the discrimination task, the results showed that the main effect of the presentation of the sound was marginally significant ($F(1,3) = 7.47$, $MSE = 0.83$, $p = .07$), while the main effect of noise level was significant ($F(6,18) = 121.31$, $MSE = 1.82$, $p < .0001$). Although the interaction term was not significant ($F < 1$, $p = .50$), planned simple main effects were conducted in order to test at which noise level the sound factor was significant. Again, the simple main effect showed that the signal contrast threshold was significantly lower in the sound present than in the sound absent condition only at the fourth noise level (i.e., -22 dB, $F(1,21) = 5.80$, $MSE = 1.02$, $p < .05$). That said, the presentation of the sound enhanced the participants' visual sensitivity to the synchronous Gabor patch only at the -22 dB noise level in both the detection and discrimination tasks.

3.3.2 The psychometric function at the intermediate (-22 dB) noise level

In order to understand whether the enhancement in performance (i.e., the decrease of signal contrast threshold) elicited by the presentation of a synchronous sound reflects

reduced temporal uncertainty, the change of threshold and slope between the psychometric functions in the sound present and sound absent conditions at the -22 dB noise level was compared (e.g., Hairol & Waugh, 2010; Yu et al., 2002). A lower uncertainty should result in a lower threshold as well as a flatter slope (i.e., a smaller slope value) of the psychometric function (see Pelli, 1985; Yu et al., 2002). That is, according to the uncertainty reduction hypothesis (Anderson & Mamassian, 2008; Lippert et al., 2007), the presentation of a synchronous sound should systematically reduce both the threshold and the slope, so that their correlation would be positive. On the other hand, according to the signal enhancement hypothesis, the presentation of a synchronous sound should only reduce the threshold without systematically changing the slope (see Hairol & Waugh, 2010; Yu et al., 2002).

The threshold and the slope of the psychometric functions for the sound present and sound absent conditions at the -22 dB level for individual participants are plotted in Figure 3.3. Note that the slope is derived at the 75% performance level (i.e., at the threshold). All four participants consistently showed the tendency for the signal contrast threshold to be lower in the sound present condition than in the sound absent condition in the detection task (see Figure 3.3A). However, only one participant revealed a pattern of results in which the slope was smaller in the sound present than in the sound absent condition. The correlation between the change of threshold and slope due to the presentation of a synchronous sound was not significant (r = -0.21, p = .40, one-tailed).

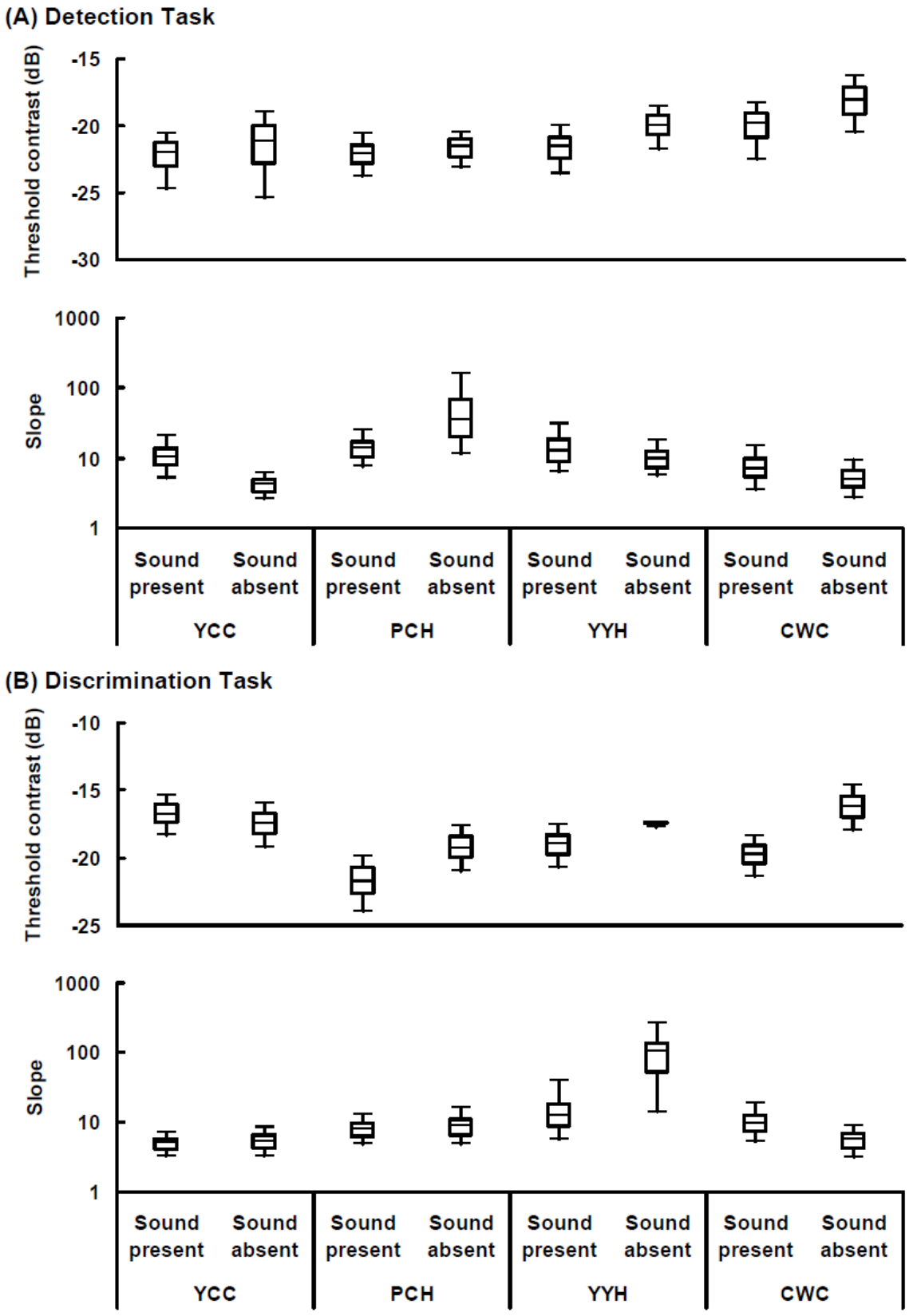


Figure 3.3. The threshold (dB) and slope (in log scale) of the psychometric functions at the -22dB noise level for each participant in (A) the detection task; and (B) the discrimination task in Experiment 3.1. Each box represents the mean, ±1 standard error, and 95% confidence interval estimated by a bootstrapping procedure (Wichmann & Hill 2001a, b).

In the discrimination task (see Figure 3.3B), three of the four participants showed the tendency for the signal contrast threshold to be lower in the sound present condition than in the sound absent condition. However, two participants showed similar, one showed a smaller, and the last one showed a larger, slope in the sound present condition than in the sound absent condition. The correlation between the change of threshold and slope due to the presentation of a synchronous sound was not significant (r = -0.33, p = .33, one-tailed). Therefore, in neither the detection nor the discrimination task, did temporal uncertainty reduction accompany the increased sensitivity resulting from the presentation of a synchronous sound.

3.4 Discussion

In Experiment 3.1, visual detection and discrimination performance was enhanced crossmodally by the presentation of a synchronous sound at an intermediate visual noise level (i.e., the -22 dB noise contrast). In this experiment, a dual 2-AFC task was used, and further, the presentation of the sound was not informative with regards to the response that a participant had to make. Therefore the crossmodal facilitatory effect reported here cannot be accounted for by any form of response bias induced by the onset of the sound (Arieh & Marks, 2008; Lippert et al., 2007; Odgaard et al., 2003). Furthermore, after examining the participants' psychometric functions for the visual target at the -22 dB noise contrast level, no evidence for the slope being systematically

reduced with the threshold by the presentation of a simultaneous sound was observed.

Given that the crossmodal facilitatory effect elicited by the presentation of a synchronous sound was observed in both visual detection and discrimination tasks, the transient boosting hypothesis cannot provide a thorough explanation. Besides, the change of threshold and the slope of the psychometric functions were not positively correlated (see also Lippert et al., 2007). Hence, there was no evidence to suggest that the participants' uncertainty was systematically reduced along with the improved sensitivity regarding the visual target elicited by the presentation of the sound. Therefore, the uncertainty reduction hypothesis, though intriguing, has not received any univocal empirical support to date.

Before concluding that the crossmodal facilitation reported in the present study genuinely attributed to the visual signal enhancement by the presentation of a synchronous sound, two relevant issues need to be considered. First, it is well-known that the presentation of a sound can alert an observer, and hence prime them to respond (Nickerson, 1973). However, the alerting elicited by the sound in the present study should induce a general facilitation effect irrespective of the contrast level of the noise. Similarly, a simple version of the signal enhancement hypothesis should predict that the visual signal can be enhanced by the presentation of a sound, so the signal-to-noise ratio (SNR) can be generally increased. However, the results reported in this chapter are inconsistent with both alerting and simple signal enhancement hypotheses. On the other

hand, the SNR can be enlarged by reducing the noise. Hence, the question of whether or not the noise in the visual system can be modulated by the presentation of the sound needs to be considered (see below).

In Experiment 3.1, crossmodal facilitation was observed at an intermediate noise contrast level. This result is similar to the results reported by Ross et al. (2007; see also Ma et al., 2009). In their study, the intensity of the target signal was fixed while that of background noise was manipulated.[14] The largest crossmodal facilitatory effects were observed at the intermediate noise levels. Note that the -22 dB noise contrast level happened to be close to the turning point of TvN function (i.e., where the signal contrast threshold start to rise up). The turning point indicates that the dominating noise for human visual performance is going to transfer from the internal noise embedded within the sensory system to the noise originating from any external stimulation (see Lu & Dosher, 1998, 1999). In other words, the turning point of a TvN function represents the noise level at which the visual system is unstable in terms of either internal or external noise dominating visual information processing, depending on which one happened to be higher than the other. One possible explanation is therefore that the presentation of a synchronous sound can genuinely enhance visual information processing only when the visual system itself is unstable. This notion, in turn, suggests that a certain range of noise levels needs to be tested when one plans to study crossmodal facilitation in terms

[14] Note that in the Ross et al.'s (2007) study, the participants had to identify the spoken words. The lips motion provided in the visual modality is therefore *informative* to the target.

of signal enhancement in human visual perception. A failure to observe any such crossmodal facilitation in terms of signal enhancement may therefore be attributable to an improper noise level in the experimental setting.

The signal contrast threshold observed in the discrimination task was higher than that observed in the detection task, which suggests that the information required for these two tasks may be different (Watson & Robson, 1981). Presumably, visual detection performance can be improved by increasing the stimulus transient, which can be enhanced by the presentation of a synchronous sound (i.e., the transient boosting hypothesis, Anderson & Mamassian, 2008). However, this notion should be qualified on the basis of the results of the present study, since human performance in an unspeeded detection task (as used in Experiment 3.1) does not necessarily rely on transient information. Instead, an unspeeded detection task can also depend on the sustained information which provides detailed identity information (see Ogmen, Breitmeyer, & Melvin, 2003). Critically, the enhancement of transient information cannot explain the results observed in the discrimination task (see Anderson & Mamassian, 2008). Hence, a more comprehensive account for the results of both the detection and discrimination tasks is therefore needed.

Fujisaki and Nishida (2010) recently proposed a framework that attempts to dissociate 'when' and 'what' information in order to account for multisensory interactions in terms of stimulus temporal synchrony and stimulus identity, respectively

(see Figure 3.4). They suggested that multisensory interactions belonging to the former type may occur in the 'when' processing pathway, whereas the latter type requires the combination of information from both the 'when' and 'what' pathways. With regards to the results reported in this chapter, crossmodal facilitation in a detection task may occur in either the 'when' pathway or after 'when' and 'what' information have been combined. By contrast, crossmodal facilitation in the discrimination task can only occur after the combination of these two types of information (and perhaps more information in the 'what' pathway needs to be accumulated). The hypothesis of dissociating 'when' and 'what' processing pathways will be further addressed in the experiments reported in Chapter 4.

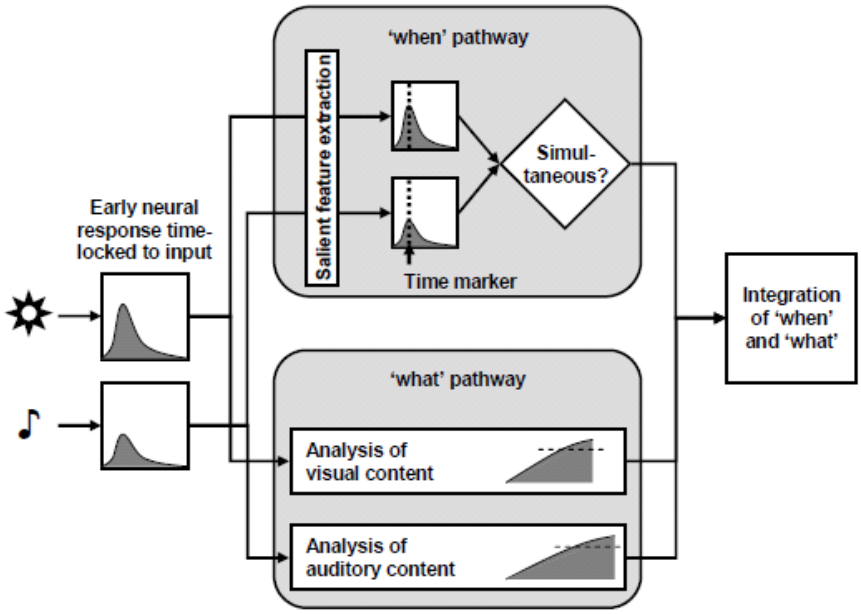


Figure 3.4. A framework proposed by Fujisaki and Nishida (2010) that 'when' and 'what' attributes of a visual or auditory stimulus are processed separately and then integrated at a later stage. The figure is adopted from Fujisaki and Nishida.

In sum, the results reported in the present chapter show that the crossmodal facilitation of visual perception resulting from the presentation of a synchronous sound can still be observed in both visual detection and discrimination tasks when the response criterion of participants is well-controlled. Note, however, that this crossmodal facilitation only occurred when the noise in the outside world just started to interfere with visual processing. With regards to the possible mechanisms underlying this crossmodal facilitation effect, it cannot solely be attributed to the presentation of the synchronous sound enhancing transient information or reducing the participants' uncertainty regarding the visual target. On the other hand, the crossmodal facilitation of visual perception cannot be accounted for by a simple signal enhancement hypothesis, either. A framework dissociating 'when' and 'what' pathways proposed by Fujisaki and Nishida (2010) is therefore considered in order to provide a more comprehensive explanation of this result.

Chapter 4

The crossmodal facilitation of visual object representations by sound: Evidence from the backward masking paradigm

4.1 Introduction

For more than fifty years now, researchers have known that visual performance can be enhanced by presenting a sound at the same time as a visual target (e.g., Bernstein et al., 1970; Hershenson, 1962; Nickerson, 1973). However, the field of crossmodal research really started to expand rapidly following the publication of Stein and Meredith's (1993) seminal volume, *The Merging of the Senses*. Stein and Meredith proposed three basic rules of multisensory integration: The *spatial coincidence,* the *temporal coincidence*, and the *inverse effectiveness* rules (see Section 1.2.1; see also Holmes & Spence, 2005). Stein and his colleagues have also investigated whether the crossmodal facilitation shown at the single cell level could be observed in human behaviour (e.g., Stein et al., 1996). Likewise, in this chapter, the aim was to investigate

the facilitatory effect elicited by the presentation of a sound on visual perception in humans, focusing primarily on the factor of temporal coincidence.

Over the years, several previous studies have demonstrated that the presentation of a sound can genuinely enhance a participant's perception of a simultaneously-presented visual target (see Section 1.3.3.2). Three possible mechanisms have also been put forward in order to account for these crossmodal facilitatory effects. The first is that the presentation of the concomitant sound may simply have alerted the participants and thus increased their preparedness to respond (this is called the preparation-enhancement hypothesis, Nickerson, 1973; Spence, 2010b). However, it should be noted that enhanced preparedness has primarily been demonstrated in speeded response tasks, which researchers believe mainly to facilitate the execution of a participants' response, rather than necessarily improving their visual perception (see Nickerson, 1973, p. 499). The second possibility is that the signal of the accessory sound may have been summed together with the visual signal and, as a result, the energy associated with the visual signal was increased (the signal-enhancement hypothesis, e.g., Bernstein et al., 1970; Hershenson, 1962). Research by Stein et al. (1996) apparently provided support for this hypothesis. Specifically, they showed that participants rated a light as looking brighter when it was accompanied by a sound than when it was not. However, this result, though dramatic, has subsequently been criticized by several researchers on the grounds of the possible influence of response bias on participants' performance. In fact, at least four

studies have subsequently failed to demonstrate that the presentation of a sound genuinely increases the perceived brightness of the visual stimulus under well-controlled laboratory conditions (see Arieh & Marks, 2008; Lippert et al., 2007; Marks et al., 2003; Odgaard et al., 2003). Lippert et al. (2007) reported that the crossmodal facilitation of visual perception by sound only occurred when the sound reliably signaled the onset of the visual target. This should be considered as a cognitive effect since the reliable co-occurrence relationship may induce an assumption that these two stimuli are originated from the same event (Lippert et al., 2007). The concomitant sound hence may have enhanced the participants' perception of 'when and where' the visual stimulus was presented. The precise spatiotemporal information may then have enhanced the object representation, making it stronger, which, as a result, will then be less likely to be influenced by other visual events. This is called the object-enhancement hypothesis (e.g., Chen Y.-C. & Yeh, 2009).

In the present study, the backward masking paradigm was used in order to provide empirical evidence that would be capable to discriminate between the three possible hypotheses regarding the crossmodal facilitation elicited by the presentation of the accessory sound on visual perception outlined above. In the backward masking paradigm, two visual letters are presented, one after the other, from the same location, with the first serving as the target and the second as the mask. This paradigm was chosen for use here not only because the presentation of the visual stimuli and the task

difficulty can be controlled easily. There are also several important reasons for using this paradigm for the first time in a crossmodal setting. Given that participants had to identify the target letter, the presentation of a sound (a pure tone in the present study) cannot provide any information with regard to the possible identity of the visual target letter. Hence, whether or not the sound was presented was independent of how participants responded. Besides, visual masking has actually frequently been used in previous crossmodal studies. For example, the target light was degraded by the subsequent presentation of four surrounding lights in the studies of Bolognini et al. (2005) and Frassinetti et al. (2002; see also McDonald et al., 2000). In the studies where a series of visual stimuli were presented, a pattern was used to mask the potential target stimuli (e.g., Olivers & Van der Burg, 2008; Ngo & Spence, 2010a, b; Vroomen & de Gelder, 2000), or the subsequent item in a RSVP stream can be considered as the mask for the previous item (e.g., Chen Y.-C. & Yeh, 2008, 2009; see also Giesbrecht & Di Lollo, 1998). Hence, the results of the experiments conducted using the backward masking paradigm reported in this chapter can be linked with the above bodies of research which used a visual mask in order to control the visual target duration or visibility (see General Discussion). Finally, and most critically, the mechanisms underlying backward masking in vision have been thoroughly explored by researchers (see Breitmeyer & Ogmen, 2006; Enns & Di Lollo, 2000).

As two letters were used as the target and mask in the present study, this kind of

backward masking is called *masking by structure*, because the spatially overlapping mask shares similar figural features with the target (Breitmeyer & Ganz, 1976). The typical result to have emerged from previous studies of backward masking by structure is that the masking effect decreases (i.e., the accuracy of reporting the target increases) monotonically as the SOA or ISI (the difference of these two values reflects the duration of the target) between the target and mask increases (see Breitmeyer & Ogmen, 2006, p. 73). Researchers have proposed that the mask may influence the target in two ways as a function of the SOA (or ISI) between the target and mask (e.g., Breitmeyer & Ganz, 1976; Kahneman, 1968; Navon & Purcell, 1981; Reeves, 1982; see Breitmeyer & Ogmen, 2006, for a review): At short SOAs (or ISIs), the target and mask may be temporally inseparable; in this case, the mask may serve as the noise to the target (this is known as *integration masking*, e.g., Eriksen, 1966, see Figure 4.1A). However, when the SOA (or ISI) is longer, the processing (or representation) of the target may be interrupted, or substituted, by that of the mask (this is known as *interruption masking*, Spencer & Shuntich, 1970; or *object substitution masking*, Di Lollo, Enns, & Rensink, 2000, see Figure 4.1B).

The mechanisms of integration and interruption (or object substitution) masking, though empirically difficult to separate, may link closely to a participant's perception of whether the target and mask were presented simultaneously or sequentially (Navon & Purcell, 1981; Reeves, 1982). That is, the two masking mechanisms can be further

differentiated in terms of their object representation as follows: In the integration masking case, the target and mask are encoded as a single visual event and thus the competition occurs between the signal and its associated noise within a unitary object representation. By contrast, in the interruption (or object substitution) masking case, the competition occurs between two distinct visual events (i.e., the target and mask). In the experiments reported in this chapter, the ISI refers to the temporal asynchrony between the visual target and the subsequently-presented mask.

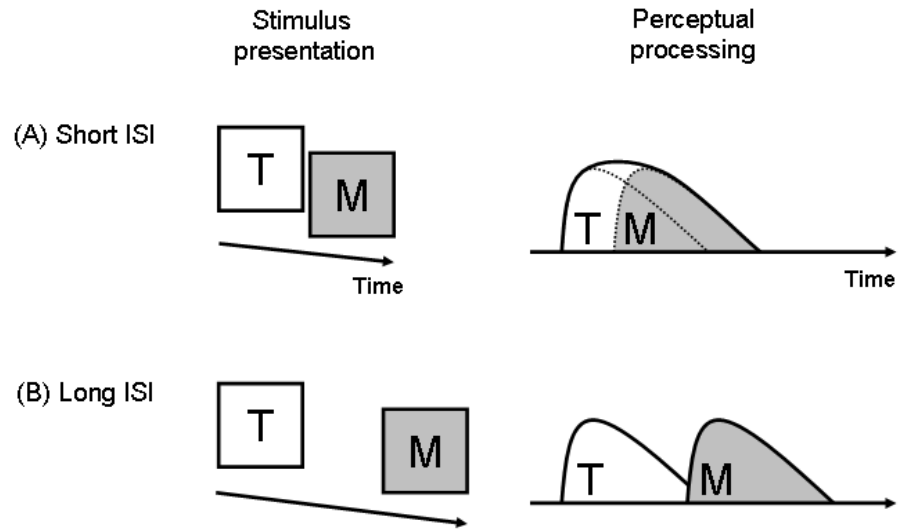


Figure 4.1. Schematic representations of two kinds of backward masking. The white square represents the target (T), and the grey square, the mask (M). (A) When the target and mask are presented at a short interstimulus interval (ISI), participants may be unable to temporally segregate them. In this case, the target and the mask may be encoded within the same perceptual event (solid black border). (B) When the ISI separating the target from the mask is increased, the two stimuli can more easily be separated perceptually. In this case, the target and mask are encoded as distinct perceptual events.

Considering the three possible mechanisms outlined above concerning how a

concomitant sound facilitates participants' visual target identification (preparedness-enhancement, signal-enhancement, and object-enhancement), different predictions can be made. If the presentation of a sound simply alerts participants to prepare to make a response, a general crossmodal facilitation elicited by the sound should be observed across a relatively broad range of ISIs before the accuracy of participants' responses reaches ceiling. That is, crossmodal facilitation should always be observed if the masking effect is still present (as determined by comparing performance to the longest ISI condition, see Figure 4.2A). If the sound enhances the signal of the accompanying visual stimulus, the signal-to-noise ratio should be increased when the sound is presented together with the target. Such enhancement of the perception of the visual target should reduce the influence of integration masking which is due to the signal associated with the visual target being degraded by the mask. Hence, the crossmodal facilitation resulting from the presentation of the sound should start at an ISI of 0 ms and last until the target and mask can effectively be segregated (see Figure 4.2B). If the sound enhances the object representation associated with the visual target, it will be less likely that the target will be interrupted or substituted by the mask. Hence, this mechanism should be able to reduce the influence of interruption/substitution masking; that is, the crossmodal facilitation resulting from the presentation of the sound should start at an intermediate ISI at which target and mask can just be segregated (see Figure 4.2C).

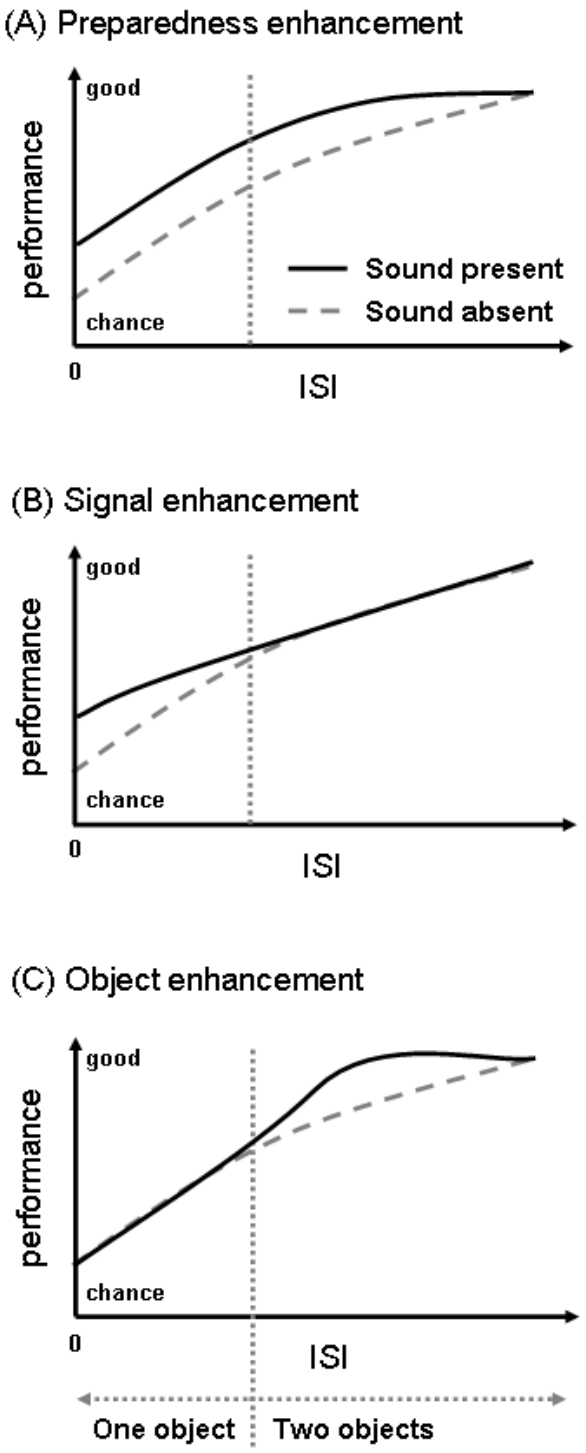


Figure 4.2. Predictions regarding the crossmodal facilitation elicited by the presentation of a sound on participants' visual target identification performance in the backward masking paradigm in terms of (A) the preparedness-enhancement hypothesis; (B) the signal-enhancement hypothesis; and (C) the object-enhancement hypothesis. The critical ISI in terms of differentiating the signal-enhancement hypothesis from the object-enhancement hypothesis occurs when participants can just segregate the target and mask as two distinct events.

In this chapter, a series of experiments designed to demonstrate the facilitation of visual perception by the presentation of a sound are reported (see Table 4.1 for a summary). The results robustly highlight a crossmodal facilitation over a relatively narrow range of ISIs. In Experiment 4.1, the timecourse over which the presentation of a sound facilitated participants' masked letter identification performance, and the timecourse over which the two letters (target and mask) can be segregated as distinct events, were investigated. In Experiment 4.2, the timecourse over which crossmodal facilitation elicited by the presentation of the sound and that induced by increasing the relative strength of the visual signal (i.e., making the visual target brighter) was tested. In Experiment 4.3, the sound was presented at the same time as the visual mask in order to test whether or not the visual masking effect would be increased. In Experiment 4.4, the importance of a reliable temporally coincident relationship between the sound and the visual target in order for obtaining the crossmodal facilitation effect was tested. Experiment 4.5 was designed to investigate whether the facilitation by sound would be larger when the sound and visual target were both spatially and temporally coincident than when they were simply temporally coincident. In all of the experiments reported here, the participants had to report the identity of the target letter, while they might also hear an accessory sound (always a pure tone). That is, the sound only provided information regarding the onset time of the visual target (except in Experiment 4.3 where the sound was presented with the visual mask), rather than its identity (or its

location, excepting Experiment 4.5).

Table 4.1. The design of each experiment in Chapter 4.

Experiment	Onset timing of the sound	ISI between visual target and mask	The third factor
4.1	Target	0, 13, 27, 40, 80, & 133 ms	
4.2	Target	0, 40, & 133 ms	Relative luminance between the target and the mask
4.3	Mask	0, 13, 27, 40, 80, & 133 ms	
4.4A	Target	0, 40, & 133 ms	The onset time of the visual target and the sound (if prsent) were either fixed or jittered
4.4B	6 SOAs with respect to the target: -106, -53, -40, 0, 40, & 80 ms	40 ms	
4.5	The sound was presented either at the consistent or inconsistent location with the target	0, 13, 27, 40, 80, & 133 ms	

4.2 Experiment 4.1

In the Experiment 4.1, participants performed a letter identification task and a letter detection task separately. The letter identification task was designed to investigate the timecourse of the crossmodal facilitation elicited by the presentation of a sound in the backward masking paradigm. The inclusion of the letter detection task was designed to investigate the minimum ISI needed in order for participants to be able to successfully segregate the visual target from the mask (i.e., the minimum interval needed in order to

perceive them as distinct perceptual events). The onset of the sound was always simultaneous with that of the visual target.

4.2.1 Methods

4.2.1.1 Participants

Eighteen volunteers (9 males, mean age of 23 years) took part in this study in exchange for course credit or 10 pounds (UK Sterling) in gift vouchers. They are students at the University of Oxford. All participants had normal or corrected-to-normal vision and hearing by self-report, and were naïve as to the purpose of the experiment. All of the participants were informed of their rights in accordance with the ethical standards laid down in the 1990 Declaration of Helsinki and signed a consent form.

4.2.1.2 Stimuli and apparatus

The stimuli were presented on a 15-inch CRT monitor (75 Hz refresh rate) controlled by a personal computer. The participants sat 58 cm from the monitor in a dimly-lit experimental chamber. Each trial consisted of the presentation of a fixation cross (0.69°×0.69°), a target letter, and a mask letter (see Figure 4.3). All of the stimuli were presented in white (65.0 cd/m^2) against a black background (0.1 cd/m^2). The targets and masks consisted of eighteen uppercase letters, excluding the letters F, G, I, L, O, P, Q, and V. The mask for a given target could be any one of the letters used except the target itself or the next letter in the alphabet. The letters were presented in Arial font,

which ranged in size from 0.79° × 1.28° to 1.19° × 1.28° (width × height). The target and mask letters were both presented for 40 ms.

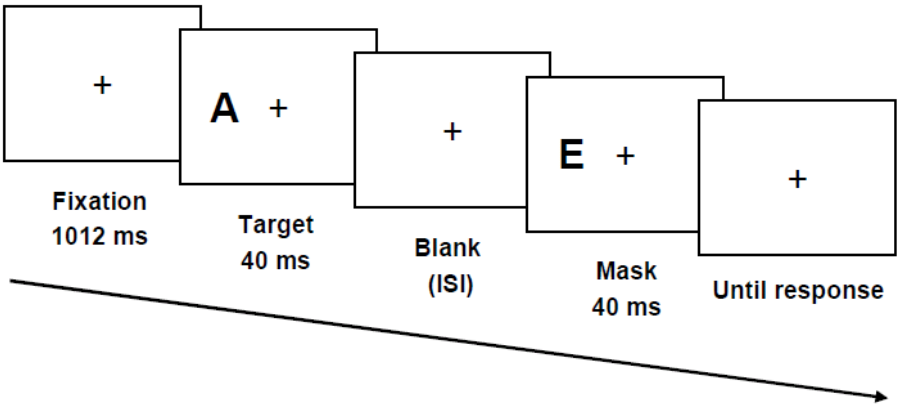


Figure 4.3. The sequence of five frames presented in each trial: A fixation cross, a target letter, a fixation cross with duration of various ISIs, a mask letter, and a fixation cross again until response. The target and mask were presented on the same side (either left or right) of fixation, which varied randomly from trial to trial.

The auditory stimulus consisted of a 1000 Hz pure tone, which was presented simultaneously with the onset of the target and lasted for 27 ms. The free-field sound was presented at 54 dB SPL (the background SPL in the experimental chamber was 48 dB). Four loudspeakers were used, two of which (one above and the other one below the monitor) were aligned with the left target location (see below), and the other two were aligned with the right target location. Given that the sound was presented simultaneously, and with equal-amplitude, from all four loudspeaker cones at the same time, it appeared to originate from the center of the monitor.

4.2.1.3 Design

The sound (present or absent) and ISI (0, 13, 27, 40, 80, and 133 ms) factors were manipulated. In the letter identification task, the participants had to report the target (i.e., the first letter) by entering their answer using the keyboard placed in front of them. A sound was randomly presented in half of the trials. Participants were informed that in some trials they might hear a beep, and, if the sound was present, it would always accompany the first letter (i.e., the target). Six ISIs between target and mask were presented in order to demonstrate the masking effect that typically shows up in terms of the accuracy of participants' target identification monotonically increasing with increasing ISI. The target and mask were always presented from the same location, randomly on either the left or right (7.17° from fixation). The location was varied in order to avoid participants fixating the target location. Target location was thus merely a control factor and target side was collapsed in the data analysis. All 18 letters were presented as the target once in each condition (2 sound conditions x 6 ISIs x 2 locations), meaning that each participant completed a total of 432 trails. These trials were divided into three blocks with balanced conditions. The presentation order of these blocks was counterbalanced across participants. The participants were allowed to take a short break between blocks.

In the letter detection task, the visual and auditory stimuli in the two-letter trials were exactly the same as in the letter identification task. Additionally, 108 catch trials that only contained one letter, half with sound and the other half without, were also

included (20% of the trials in total). Participants had to report whether they saw one or two letters in each trial. In other words, the detection task consisted of a two-alternative forced-choice (2-AFC) task. Again, these trials were divided into three blocks with balanced conditions. The order in which these two tasks were completed was balanced across participants. The experiment lasted for approximately an hour.

4.2.1.4 Procedure

In the identification task, participants initiated a block of trials by pressing the 'ENTER' key on the keyboard. The fixation cross was presented by itself for 1012 ms, and a target was then presented. During the ISI, a frame with only the fixation cross was presented, followed by the presentation of the mask. The participants had to enter the first letter (they were forced to guess if they were unsure) and then press the 'ENTER' key again in order to initiate the next trial. The fixation cross was presented during the intertrial interval, and the participants were asked to fixate that location throughout the experiment. In the detection task, the participants had to report whether they had seen one or two letters (by pressing the 'left arrow' and 'right arrow' keys, respectively, corresponding to the locations of number keys on the keyboard). The next trial started after the participant's response was detected. The participants were instructed to perform as accurately as possible, while they did not have to respond speedily.

A practice session consisting of 24 trials of the identification task and 30 trials of

the detection task was conducted before the main experiment, which included all possible combinations of the manipulated conditions. The experimenter made sure that participants were following the instructions before proceeding onto the main experiment.

4.2.2 Results

In order to make sure that the participants were concentrating on the visual task and had a similar ability to perform the backward masking task, a criterion of 80% accuracy at the longest ISI (133 ms) was used throughout the following experiments reported in this chapter. If the participant failed to reach this criterion, his/her data was excluded from further analysis. The accuracy of participants' responses in each condition were first subjected to arc-sine transformation using the equation $X' = 2 * \sin^{-1}\sqrt{X}$ (Kirk, 1995) in order to linearize the percentage accuracy data.

4.2.2.1 Results of the letter identification task

One participant failed to reach the criterion, leaving the data from 17 participants for further analysis (see Figure 4.4A, solid lines). The transformed data were then submitted to a two-way ANOVA on the factors of sound and ISI. Both the main effect of sound (69.9% vs. 66.5% for sound present and sound absent conditions, $F(1,16) = 6.44$, $MSE = 0.04$, $p < .05$, $\eta^2 = 0.29$) and ISI ($F(5,80) = 142.68$, $MSE = 0.04$, $p < .0001$, $\eta^2 = 0.90$) were significant. A *post-hoc* Tukey test revealed that, compared to the 133 ms ISI,

the accuracy of participants' performance at all of the other ISIs were significantly lower (all *p*s < .01). The interaction between sound and ISI was significant ($F(5,80) = 2.29$, $MSE = 0.02$, $p = .05$, $\eta^2 = 0.13$). The simple main effect demonstrated that the participants' accuracy was higher in the sound present condition than in the sound absent condition at the 27 ms (67.0% vs. 59.3%, $F(1,96) = 8.05$, $MSE = 0.03$, $p < .01$, $\eta^2 = 0.07$) and 40 ms ISIs (71.5% vs. 63.1%, $F(1,96) = 11.41$, $p < .005$, $\eta^2 = 0.10$), but not at any of the other ISIs (all Fs < 1, ps > .44, η^2s < 0.01).

The participants in this experiment may sometimes have erroneously reported the mask letter as the target. In order to understand how often such error was made, the percentage of misreporting the mask letter in each condition was calculated (see Figure 4.4A, dashed lines). The transformed data were submitted to a two-way ANOVA on the factors of sound and ISI. The main effect of sound was significant ($F(1,16) = 11.91$, $MSE = 0.02$, $p < .005$, $\eta^2 = 0.43$). Note that the participants misreported the mask as the target less frequently in the sound present condition (18.3%) than in the sound absent condition (20.9%). The main effect of ISI was also significant ($F(5,80) = 63.84$, $MSE = 0.07$, $p < .0001$, $\eta^2 = 0.80$). A post-hoc Tukey test revealed that, as compared to the 133 ms ISI, the rates of misreporting at 0, 13, 27, and 40 ms ISIs were significantly higher (all *p*s < .01). The interaction between sound and ISI was, however, not significant ($F(5,80) = 0.36$, $MSE = 0.03$, $p = .88$, $\eta^2 = 0.02$). In order to know whether the sound

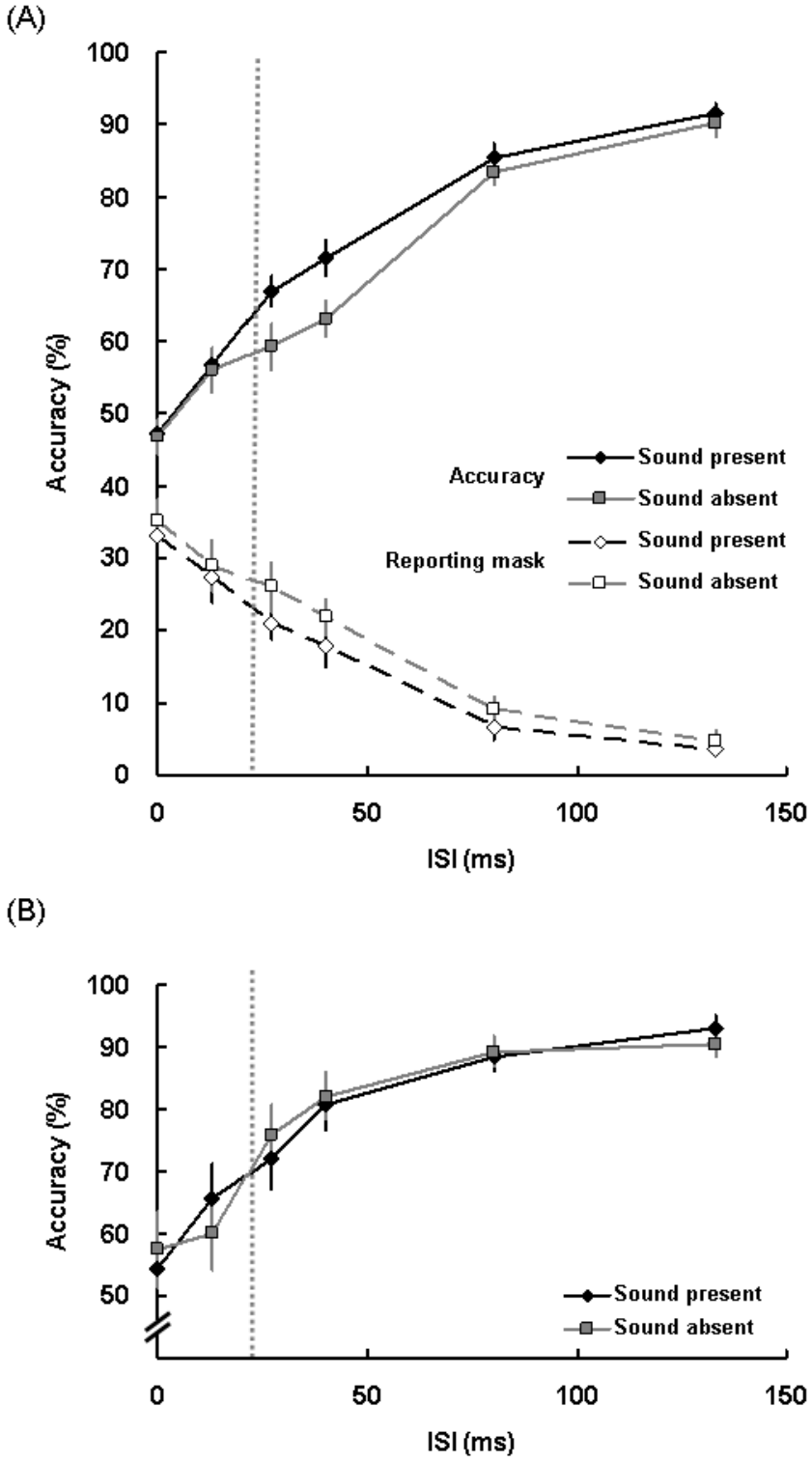


Figure 4.4. (A) The mean accuracy of participants' performance in the letter identification task (solid lines) and mean percentage of misreporting the mask letter as the target (dashed lines); and (B) The mean accuracy of participants' performance in the letter detection task (50% accuracy represents chance level) of Experiment 1. The grey dotted line represents the estimated ISI needed to successfully differentiate the target and mask as distinct letters (i.e., 75% accuracy in detecting the presence of two letters corresponds to an ISI of 21-22 ms). The error bars indicate ±1 standard error of the means.

factor was significant at each ISI, planned simple main effect analyses were conducted. The results showed that the sound factor was only significant at the 40 ms ISI (17.8% and 21.9% for the sound present and absent conditions respectively, $F(1,96) = 3.83$, $p = .05$, $\eta^2 = 0.04$), but not at any of the other ISIs (all Fs < 3.45, ps > .07, η^2s < 0.03). The results therefore show that, generally-speaking, the rate of misreporting the mask as the target was a mirror reflection of the accuracy data, while the magnitude of the effects were smaller and therefore less likely to reach significance. The rate of misreporting the mask as the target will only be reported again in Experiment 4.3, where the sound was presented with the visual mask.

4.2.2.2 Results of the letter detection task

The accuracy data were transformed and analyzed (see Figure 4.4B). In the one-letter trials, the difference in accuracy between the sound present (92.1%) and sound absent conditions (93.0%) was not significant ($F(1,16) = 1.98$, $MSE = 0.02$, $p = .18$, $\eta^2 = 0.11$). The accuracy of participants' performance in the two-letter trials was submitted to a two-way ANOVA on the factors of sound and ISI. Only the main effect of ISI was significant ($F(5,80) = 35.14$, $MSE = 0.14$, $p < .0001$, $\eta^2 = 0.69$). A *post-hoc* Tukey test revealed that, compared to the 133 ms ISI, the accuracy at the 0, 13, and 27 ms ISIs was significantly lower, ps < .01, and the accuracy at the 40 ms ISI was also significantly lower, $p < .05$. The interaction between sound and ISI was also significant

($F(5,80) = 3.53$, $MSE = 0.03$, $p < .01$, $\eta^2 = 0.18$). Analysis of the simple main effect showed that the presentation of the sound led to an increase in the accuracy of participants' performance as compared to the sound absent condition at the 13 ms (65.7% vs. 59.9%, $F(1,96) = 6.19$, $MSE = 0.04$, $p < .05$, $\eta^2 = 0.06$) and 133 ms ISIs (93.2% vs. 90.6%, $F(1,96) = 4.10$, $p < .05$, $\eta^2 = 0.04$).

A log regression was conducted on the mean performance data in order to estimate the ISI at which the participants were able to correctly detect the presence of two letters in the sound present and sound absent conditions, respectively (at 75% accuracy, which corresponds to the threshold in a 2-AFC task; see Kingdom & Prins, 2010, p. 65). The estimated ISI was 21 ms for the sound present condition and 22 ms for the sound absent condition. The ISIs needed to achieve 75% correct for each participant in the sound present and sound absent conditions were also estimated. The values were submitted to a one-way ANOVA; however, the main effect was not significant ($F = 1.00$, $p = .33$, $\eta^2 = 0.06$).

4.2.3 Discussion

The results of the letter identification task revealed that the presentation of a sound only facilitated the participants' identification of the visual target at the intermediate ISIs (27 and 40 ms). In the letter detection task, the presentation of the sound facilitated participants' detection performance at a further shorter ISI (13 ms). Even so, the

estimated threshold ISI that was required to differentiate the target and mask as constituting distinct letters was no different in the sound present versus in the sound absent conditions.

Three critical findings should be highlighted here: First, the crossmodal facilitation by the sound in the identification task occurred within a narrow time-window (namely, at just the ISIs of 27 and 40 ms). That is, there was no crossmodal facilitation at any of the other ISIs (0, 13, 80, and 133 ms). This result is inconsistent with the prediction of the preparedness-enhancement hypothesis (Nickerson, 1973). Critically, no sound-induced facilitation was observed at the 0 ms ISI, which is inconsistent with the prediction of the signal-enhancement hypothesis (Stein et al., 1996). Second, combining the results of the identification and detection tasks, the range of ISIs over which the crossmodal facilitation of visual letter identification performance was observed (i.e., 27 and 40 ms ISI) was just beyond the threshold for participants being able to correctly differentiate the target and mask letters as constituting separate events (at an ISI of around 21~22 ms). These results therefore suggest that the crossmodal facilitation effect in the backward masking paradigm was correlated with our participants' ability to segregate two subsequently-presented visual events (see Navon & Purcell, 1981; Reeves, 1982). That said, the presentation of the sound likely enhanced the participants' processing of the concomitant visual event and hence reduced the likelihood that the object representation would be interrupted/substituted by the mask. This finding is

consistent with the object-enhancement hypothesis.

Third, the crossmodal facilitation elicited by the presentation of the sound occurred at a shorter ISI in the detection task (13 ms) than in the identification task (27 & 40 ms). These results may be attributed to the fact that when humans are trying to bind rapidly-changing visual and auditory stimuli, the threshold temporal frequency is higher when matching the timing of the salient change of crossmodal stimuli than when binding the identity of information across sensory modalities (Fujisaki & Nishida, 2010). Hence, facilitation in the detection task can be observed when the visual stimuli changed faster (i.e., when the ISI was shorter) than the facilitation in the identification task. These results also imply that, even though the task did not explicitly ask the participants to bind the visual and auditory stimuli, crossmodal facilitation should result from integrating them.

In the next experiment, whether the facilitation of participants' letter identification performance induced by the presentation of the sound is different from that induced by increasing the strength of the visual signal was investigated. Spencer and Shuntich (1970) showed that interruption masking is insensitive to the energy-level of the mask (while the energy of the target is fixed). Hence, if the presentation of the sound helped to prevent the accompanying visual target from being interrupted by the mask, it should also be the case that the timecourse of the crossmodal facilitation effect is insensitive to any manipulation of the luminance of the visual stimulus. Besides, it should also be

considered that previous studies have shown that the enhancement of people's ability to identify spoken monosyllabic words that results from viewing congruent lip movements is largest at intermediate levels of task difficulty (Ma et al., 2009; Ross et al., 2007). Hence, one potential confound in Experiment 1 is that the crossmodal facilitation effect reported there may actually have resulted from the intermediate task difficulty seen at the 27 and 40 ms ISIs rather than the temporal aspects of the ISIs themselves.

In order to test these two possibilities in Experiment 2, the relative luminance of the visual target and the mask was manipulated. Comparing the pattern of performance seen in the bright versus dark visual target conditions should therefore reveal the timecourse of letter identification enhancement resulting from the presentation of stronger visual signals. Besides, the bright visual target may increase the accuracy of participants' performance at the 0 ms ISI (e.g., Hellige, Walsh, Lawrence, & Prasse, 1979), which ought to be allowed to evaluate whether or not it is intermediate task difficulty that determines whether or not facilitation by the sound is observed.

4.3 Experiment 4.2

In this experiment, the relative luminance of the target and mask was manipulated. Once again, a sound was presented simultaneously with the visual target on half of the trials. Three critical ISIs were tested in this experiment.

4.3.1 Methods

4.3.1.1 Participants

Twenty-four volunteers (8 males, mean age of 22 years) from the same subject pool as described in Experiment 4.1 took part in this experiment.

4.3.1.2 Stimuli, apparatus, design, and procedure

Three factors, sound (present or absent), ISI (0, 40, and 133 ms), and relative luminance (target brighter or mask brighter), were manipulated. The relative luminance factor was blocked. In the target brighter condition, the target letter was white (65.0 cd/m^2, as in the previous experiment) while the mask letter was 50% gray (12.0 cd/m^2). The luminance of the target and mask letters were swapped in the mask brighter condition. There were three blocks of trials for both the target brighter and mask brighter conditions, and the order of presentation of these blocks was balanced and intermixed between participants. This experiment took about 30 minutes to complete. The other details were as in Experiment 1.

4.3.2 Results and discussion

Three participants failed to reach the criterion of 80% correct at the 133 ms ISI, leaving the data from 21 participants for further analysis (see Figure 4.5). The accuracy of participants' target identification performance was transformed and submitted to a three-way ANOVA with the factors of sound, ISI, and relative luminance. The main

effects of sound (73.6% and 71.9% for the sound present and absent conditions respectively, $F(1,20) = 4.42$, $MSE = 0.03$, $p < .05$, $\eta^2 = 0.18$) and ISI ($F(2,40) = 236.94$, $MSE = 0.09$, $p < .0001$, $\eta^2 = 0.92$) were significant. The two-way interaction between sound and ISI was also significant ($F(2,40) = 3.23$, $MSE = 0.02$, $p = .05$, $\eta^2 = 0.14$). The simple main effect showed that the sound factor only reached significance at the 40 ms ISI (73.6% vs. 69.4% for the sound present and absent conditions respectively, $F(1,60) = 10.12$, $MSE = 0.02$, $p < .005$, $\eta^2 = 0.14$), but not at the 0 and 133 ms ISIs (Fs < 1.19, ps > .28, $\eta^2 < 0.02$). The two-way interaction between ISI and relative luminance was also significant ($F(2,40) = 25.89$, $MSE = 0.10$, $p < .0001$, $\eta^2 = 0.56$). The simple main effect of relative luminance revealed that the participants only responded more accurately in the target brighter condition than in the mask brighter condition at the 0 ms ISI (64.9% vs. 42.7%, $F(1,60) = 33.27$, $MSE = 0.14$, $p < .0001$, $\eta^2 = 0.34$), but not at either the 40 or 133 ms ISIs (Fs < 3.71, ps > .05, η^2s < 0.04). Even though the simple main effect of ISI was significant in both the target brighter and mask brighter conditions (both Fs > 61.78, $MSE = 0.10$, ps < .0001, η^2s > 0.21), the *post-hoc* Tukey tests revealed different results: In the target brighter condition, the accuracy of participants' performance at the 0 and 40 ms ISIs were significantly lower than at 133 ms ISI (ps < .01); however, the accuracy of participants' performance at the two shorter ISIs (64.9% vs. 69.7%) was statistically indistinguishable. In the mask brighter condition, the accuracies at all three ISIs were significantly different from each other

(all *p*s < .01).

In order to confirm whether the crossmodal facilitation effect elicited by the presentation of the sound in Experiment 2 was robust at the 40 ms ISI, planned simple main effects analyses of the sound factor in the target brighter and mask brighter conditions were conducted. The accuracy of participants' performance was higher in the sound present condition than in the sound absent condition for both the target brighter (72.2% vs. 67.2%, $F(1,120) = 5.00$, $MSE = 0.03$, $p < .05$, $\eta^2 = 0.04$) and mask brighter conditions (74.9% vs. 71.5%, $F(1,120) = 3.73$, $p = .05$, $\eta^2 = 0.03$) at the 40 ms ISI. The magnitude of the crossmodal facilitation effect reported at the 40 ms ISI in the target brighter and mask brighter conditions was further examined by subtracting the accuracy

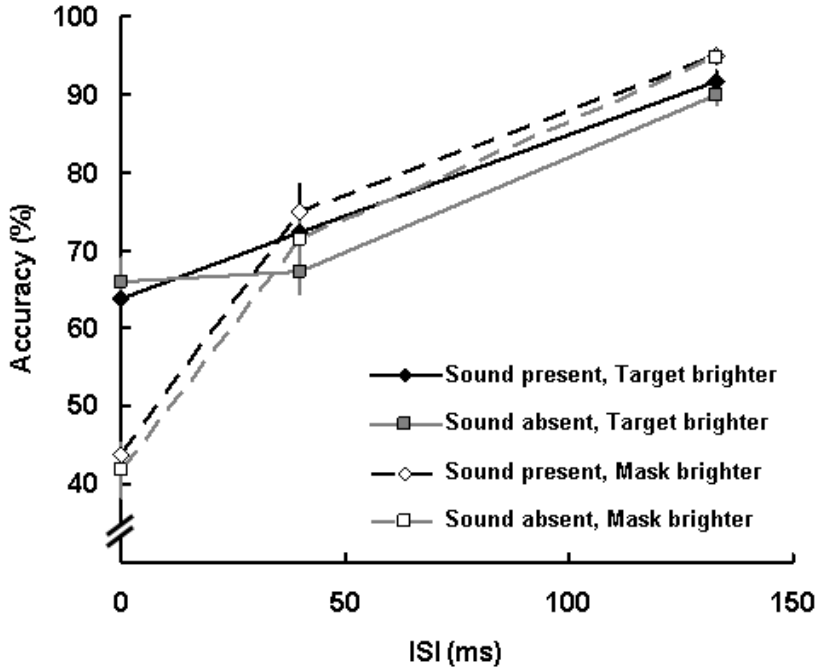


Figure 4.5. The mean accuracy of letter identification performance in Experiment 4.2. The error bars indicate ±1 standard error of the means.

of participants' performance in the sound present condition from that seen in the sound absent condition (5.0% in the target brighter condition and 3.4% in the mask brighter condition, respectively); however, this difference was not significant ($F < 1$, $p = .62$, $\eta^2 = 0.01$). The results suggest that the crossmodal facilitation effect occurred robustly in both the target brighter and mask brighter conditions at the 40 ms ISI, and the magnitude of facilitation was similar.

The results of Experiment 4.2 revealed that the relative luminance of the target and mask influenced participants' letter identification performance at the 0 ms ISI but not at the 40 ms ISI. On the other hand, the presentation of the sound actually modulated participants' performance at the 40 ms ISI rather than at the 0 ms ISI, just as in the previous experiment. The results therefore suggest that the facilitation of participants' masked letter identification performance induced by the presentation of the sound (at the 40 ms ISI) can be dissociated from that induced by the brightening of the visual target (at the 0 ms ISI).

The fact that the task was easier when the target was brighter than the mask occurred at the 0 ms ISI is presumably attributable to the fact that the signal (i.e., the target letter) was stronger than the noise (i.e., the mask letter) in the target brighter condition, and vice versa in the mask brighter condition. Namely, if the presentation of the sound had enhanced the concomitant visual signal (i.e., perhaps by making the target appear brighter, see Stein et al., 1996), crossmodal facilitation by the sound should have

been observed at the 0 ms ISI. Two further results were observed here: First, the manipulation of the input signal only modulated the masking effect at the 0 ms ISI (i.e., at an ISI below 40 ms), which, in turn, suggests that the 40 ms ISI was outside of the time-window for integration masking (e.g., Hellige et al., 1979; Spencer & Shuntich, 1970). Second, in the target brighter condition, the accuracy of participants' performance at the 0 ms ISI was no different from that seen at the 40 ms ISI; however, the crossmodal facilitation effect was only observed at the 40 ms ISI. That said, in order to observe the crossmodal facilitation resulting from the presentation of the sound in the present study, the presence of an empty interval between the target and the mask (i.e., ISI) was critical, rather than necessarily the fact that participants were performing at an intermediate level of task difficulty (cf. Ma et al., 2009; Ross et al., 2007).[15]

A further way in which to test the signal-enhancement hypothesis is by presenting the sound and the mask simultaneously. In this case, if the sound were to enhance the salience of the visual signal, participants' performance should be impaired, since the mask (which actually served as visual noise) is enhanced (especially at the 0 ms ISI, see Experiment 4.2). In addition, a possible bias that is worth considering here is that participants might exhibit a tendency to respond to the letter that happens to accompany the sound, especially when they were able to identify both letters but were uncertain

[15] A critical difference between the studies of Ross et al. (2007) and Ma et al. (2009) and the present study should be noted. In their studies, the movement of the lips on the screen provided useful information regarding the sound of a spoken word; however, in the present study, the sound provided no information concerning the *identity* of the target letter. Whether the accessory stimuli were informative to the task at hand may be the critical factor needed to help explain this discrepancy.

with regards to their correct order of presentation. If this bias influenced the results, the participants should have misreported the mask letter as the target more frequently when the sound was presented together with the mask than when no sound was present. In Experiment 4.3, the sound, if present, always accompanied the visual mask (i.e., the second letter).

4.4 Experiment 4.3

4.4.1 Methods

4.4.1.1 Participants

Eighteen volunteers (8 males, mean age of 26 years) from the same subject pool as described in Experiment 1 took part in this experiment.

4.4.1.2 Stimuli, apparatus, design, and procedure

Two factors, sound (present or absent) and ISI (0, 13, 27, 40, 80, and 133 ms) were manipulated. The participants were not given any information with regards to the onset time of the sound, because the participants may have tried to ignore the sound if they had known that the sound would never coincide with the target. After the experiment, a surprise inquiry was made: '*Which letter do you think the sound was presented simultaneously with?*' This experiment took about 30 minutes to complete. All other aspects of the design were exactly the same as in Experiment 4.1.

4.4.2 Results and discussion

One participant failed to reach the criterion of 80% correct at 133 ms ISI, leaving the data from 17 participants for further analysis (see Figure 4.6, solid lines). The transformed data were submitted to a two-way ANOVA on the factors of sound and ISI. Both the main effect of sound (67.0% and 64.3% for the sound present and absent conditions respectively, $F(1,16) = 12.77$, $MSE = 0.02$, $p < .005$, $\eta^2 = 0.44$) and ISI ($F(5,80) = 217.50$, $MSE = 0.03$, $p < .0001$, $\eta^2 = 0.93$) reached significance. A *post-hoc* Tukey test showed that, compared to the 133 ms ISI, the accuracy at all other ISIs was significantly lower (all ps $< .01$). The interaction between sound and ISI was not significant ($F(5,80) = 1.33$, $MSE = 0.02$, $p = .32$, $\eta^2 = 0.08$). Note that, although the accuracy in the sound present condition (43.1%) was somewhat lower than in the sound absent conditions (43.6%) at the 0 ms ISI, this difference was not significant (planned simple main effect, $F < 1$, $\eta^2 = 0$).

Here, the percentage of misreporting the mask letter as the target in each condition was calculated again (see Figure 4.6, dashed lines). The percentage misreporting data were transformed and submitted to a two-way ANOVA on the factors of sound and ISI. The main effect of ISI was significant ($F(5,80) = 147.51$, $MSE = 0.05$, $p < .0001$, $\eta^2 = 0.90$). A *post-hoc* Tukey test showed that, as compared to the 133 ms ISI, the percentage of trials in which the participants reported the mask letter as the target at all other ISIs was significantly higher (all ps $< .01$). The main effect of sound was, however, only

marginally significant ($F(1,16) = 3.95$, $MSE = 0.03$, $p = .06$, $\eta^2 = 0.20$). It should be noted that participants were less likely to report the mask letter as the target in the sound present condition (21.2%) than in the sound absent condition (23.4%), even though the sound actually accompanied the mask letter. The interaction between sound and ISI was not significant ($F < 1$, $p = .61$, $\eta^2 = 0.04$). These results therefore show that the rate of participants misreporting the mask letter as the target decreased with increasing ISI, and that the presentation of a sound with the visual mask tended to further reduce this rate of misreporting regardless of the ISI.

Concerning the unexpected inquiry, four of the 17 participants reported that the sound was presented with the first letter, seven reported the sound was presented with the second letter, and six gave other answers (such as between two letters, hadn't noticed the sound, or that the sound served as feedback). Their answers have no obvious tendency ($X^2(2) = 0.82$, $p = .66$). When the data were separated into three groups according to the participants' subjective report, this factor was not significant ($F(2,14) = 3.16$, $MSE = 0.05$, $p = .07$, $\eta^2 = 0.31$), neither did it interact with the sound and ISI factors (all Fs < 1.38, ps $> .21$. η^2s < 0.16).

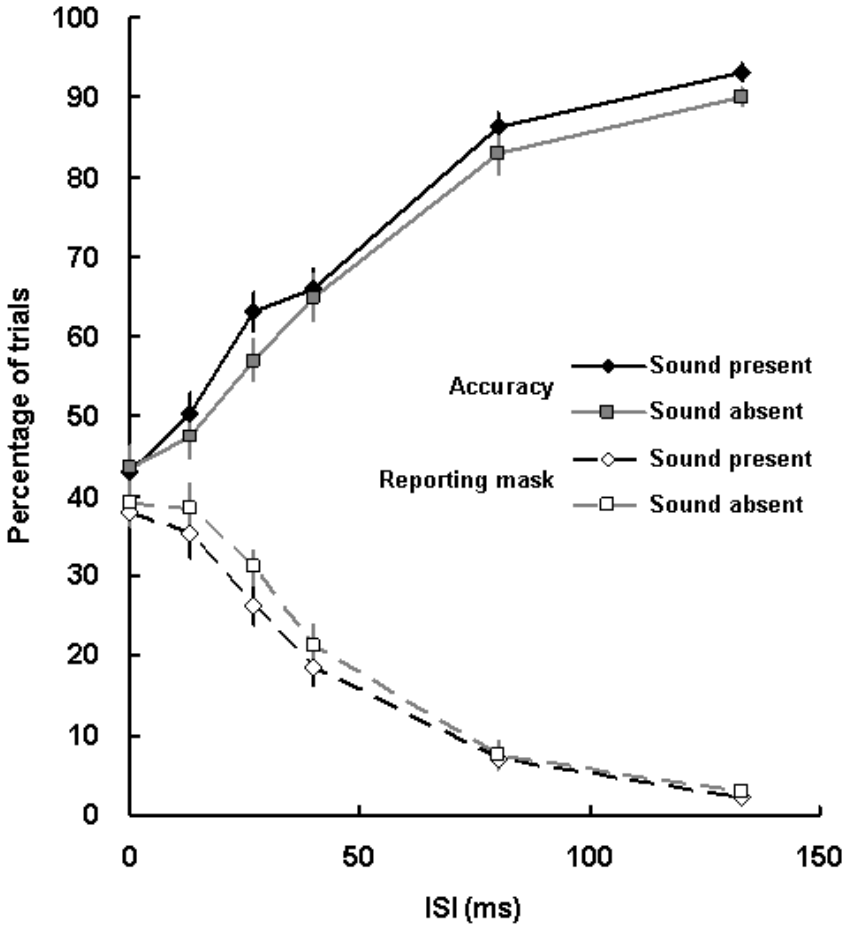


Figure 4.6. The mean accuracy of letter identification performance (solid lines) and mean percentage of misreporting the mask letter as the target (dashed lines) in Experiment 3. The error bars indicate ±1 standard error of the means.

The results of Experiment 4.3 demonstrate that when the sound was added at the onset of the mask, the presentation of the sound nevertheless still facilitated participants' performance in the letter identification task. Therefore, the results of Experiment 4.3 provide no evidence to support the claim that the presentation of the sound enhances the signal of the concomitantly-presented visual stimulus. It should also be noted here that the crossmodal facilitation elicited by the presentation of the sound when it accompanied the visual mask was different in kind from that seen when it

accompanied the visual target, since the observed facilitatory effect was no longer specific to certain ISIs (in Experiment 4.3).

It is surprising that the presentation of the sound accompanying the visual mask still improved participants' visual target identification performance in Experiment 4.3. The results thus suggest that the participants responded to the first letter that they saw rather than to the letter that happened to be accompanied by a sound (also note that no difference was observed between those participants who perceived the onset of the sound at either first or second letter, or others). Besides, this crossmodal facilitation effect cannot be attributed to the fact that the presentation of the sound alerted participants either, since the sound was presented *after* (40-173 ms) the target letter. The final possibility here is that the sound can better signify the onset of the mask letter and thus help to differentiate it from the target, which therefore somehow reduces the inappropriate reporting of the masking letter as the target. A detailed discussion of how this is relevant to the object-enhancement hypothesis can be found in Section 4.8.2.

In Experiments 4.1 and 4.2, the onset of visual target and the sound (if present) always occurred after the 1012 ms fixation frame, and participants were therefore presumably able to predict the onset time of the visual target and sound. The crossmodal facilitation effect reported in Experiments 4.1 and 4.2 therefore likely relied on the occurrence of the visual target and the sound being well-predicted, or that their temporally coincident relationship was critical (e.g., Lippert et al., 2007). In Experiment

4.4A, the sound and visual target were always presented together (i.e., synchronously), while their onset time with respect to that of fixation cross was either fixed or jittered (i.e., varied within a certain temporal range). In Experiment 4.4B, the onset time of the visual target was again fixed, while that of the sound varied.

4.5 Experiment 4.4A

In this experiment, the onset time of the synchronized sound and visual target was either fixed or jittered. The ISI between the visual target and mask was 0, 40, or 133 ms.

4.5.1 Methods

4.5.1.1 Participants

Twenty-two volunteers (three males, mean age of 26 years) from the same subject pool as described in Experiment 4.1 took part in this experiment.

4.5.1.2 Stimuli, apparatus, design, and procedure

Three factors, sound (present or absent), ISI (0, 40, and 133 ms), and target onset time (fixed or jittered), were manipulated in this experiment. The target onset time factor was blocked. The experimental procedure was modified as follows: The fixation cross was again presented at the beginning of each trial. The duration of the fixation frame was 1012 ms in the fixed condition, and 306, 1012, or 2010 ms (one-third of trials for each) randomly interleaved in the jittered condition. There were three blocks of trials

for both the fixed and jittered condition, and the order in which these blocks were conducted was balanced and intermixed between participants. This experiment took about 30 minutes to complete. The other details were as in Experiment 4.1.

4.5.2 Results and discussion

Four participants failed to reach the criterion of 80% correct at 133 ms ISI, leaving the data from 18 participants for further analysis (see Figure 4.7). The accuracy of their letter identification performance was transformed and submitted to a three-way ANOVA with the factors of sound, ISI, and target onset time. The main effect of ISI was significant ($F(2,34) = 306.09$, $MSE = 0.09$, $p < .0001$, $\eta^2 = 0.95$). A *post-hoc* Tukey test revealed that, as compared to 133 ms ISI, the accuracy at the 0 and 40 ms ISIs were significantly lower (ps $< .01$). The main effect of target onset time was also significant ($F(1,17) = 17.35$, $MSE = 0.03$, $p < .001$, $\eta^2 = 0.51$). The higher accuracy reported in the fixed condition (70.0%) than in the jittered condition (66.3%) indicates that the participants found the jittered condition somewhat more difficult. The two-way interaction between sound and ISI was significant ($F(2,34) = 5.22$, $MSE = 0.02$, $p < .05$, $\eta^2 = 0.24$). The simple main effect showed that the sound factor only reached significance at the 40 ms ISI (69.6% and 64.8% for the sound present and absent conditions respectively, $F(1,51) = 9.05$, $MSE = 0.03$, $p < .005$, $\eta^2 = 0.15$), but not at either the 0 or 133 ms ISIs (Fs < 1, ps $> .34$, η^2s $< .02$).

Planned simple main effects were conducted on the sound factor in the fixed and jittered conditions at the 40 ms ISI. Participants responded more accurately in the sound present condition than in the sound absent condition in both the fixed (72.3% and 68.2% for the sound present and absent conditions respectively, $F(1,102) = 4.02$, $MSE = 0.02$, $p < .05$, $\eta^2 = 0.04$) and jittered (66.9% and 61.4% for the sound present and absent conditions respectively, $F(1,102) = 6.19$, $p < .05$, $\eta^2 = 0.05$) conditions. The magnitudes of this crossmodal facilitation effect at the 40 ms ISI in the fixed (4.1%) and jittered (5.5%) conditions were not significantly different ($F < 1$, $p = .52$, $\eta^2 = 0.03$). That said, the crossmodal facilitation effect occurred robustly in both the fixed and jittered conditions at the 40 ms ISI, and the magnitude of facilitation was similar.

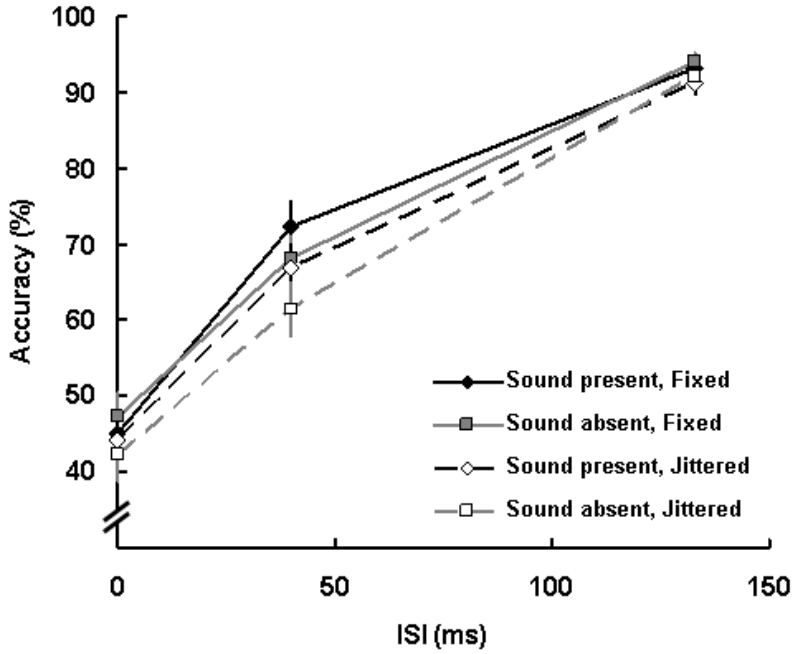


Figure 4.7. The mean accuracy of letter identification performance in Experiment 4.4A, the ISI represents the interval between the visual target and mask. The error bars indicate ±1 standard error of the means.

The results of Experiment 4.4A revealed that when the sound was reliably presented to coincide with the onset of the visual target, crossmodal facilitation was robustly observed at the 40 ms ISI. This was true even though the onset time of the sound and the visual target was less predictable (in the jittered as compared to the fixed fixed condition). Furthermore, the predictability of the onset time of the sound and the visual target did not modulate the magnitude of the crossmodal facilitation effect.

4.6 Experiment 4.4B

In this experiment, the SOA between the sound and the target were manipulated. The ISI between the visual target and mask was always 40 ms.

4.6.1 Methods

4.6.1.1 Participants

Twenty volunteers (9 males, mean age 24 years) from the same subject pool as described in Experiment 4.1 took part in this experiment.

4.6.1.2 Stimuli, apparatus, design, and procedure

There were six possible SOAs between the sound and the target. They were manipulated as follows: The sound could appear before the onset of target (-106, -53, or -40 ms), at the onset of target (0 ms), at the offset of the target (40 ms), or at the onset of mask (80 ms). A sound absent condition was also tested in order to serve as the baseline

to assess the unimodal visual masking effect. A sound was presented in a random half of the trials. This experiment took about 30 minutes to complete. The other details were same as in Experiment 4.1.

4.6.2 Results and discussion

The accuracy of participants' letter identification performance (see Figure 4.8) was transformed and submitted to a one-way ANOVA with sound SOA (6 SOAs and the sound absent condition) as the factor. The main effect of SOA was significant ($F(6,114) = 3.16$, $MSE = 0.02$, $p < .01$, $\eta^2 = 0.14$). A *post-hoc* Tukey test revealed that the accuracy of participants' performance was higher at the -106 ms SOA than at the 80 ms SOA ($p < .05$); critically, the difference between the 0 ms SOA and the no-sound condition was only 0.8%, and these two conditions were not significantly different.[16]

The results of Experiment 4.4B therefore indicate that if the onset time of the sound with respect to the visual target was varied, no crossmodal facilitation was observed when the sound happened to be presented at the same time as the visual target. Lippert et al. (2007) had the participants in their study detect the presence of a Gabor patch, while a sound was presented at one of five SOAs relative to the onset of a Gabor: 0, ±150, and ±400 ms. They too failed to find any significant facilitatory effect when the

[16] A control experiment was conducted in which the sound was always presented in synchrony with the visual target and the ISI between the visual target and the mask was always 40 ms. Another group of 15 participants (four males, mean age of 26 years) from the same subject pool as described in Experiment 4.1 was tested. The data from each participant were transformed and submitted to a one-way ANOVA, which showed that the accuracy was significantly higher in the sound present condition than in the sound absent condition (66.9% vs. 63.9%; $F(1,14) = 4.66$, $MSE = 0.01$, $p < .05$, $\eta^2 = 0.25$).

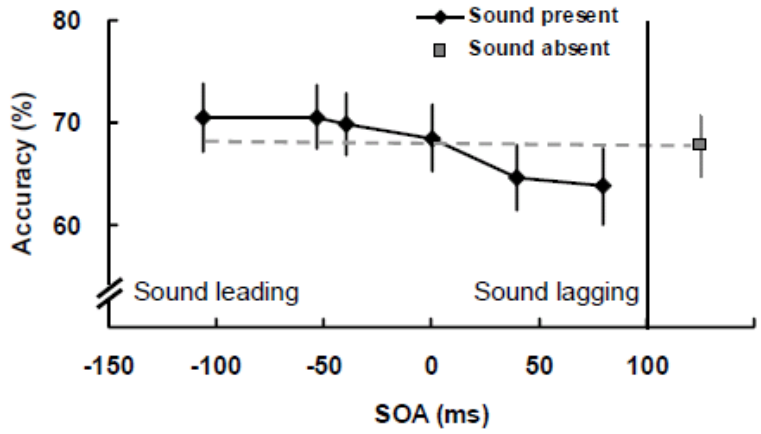


Figure 4.8. The mean accuracy of letter identification performance in Experiment 4.4B, the SOA represents the interval between the sound and the visual target. The error bars indicate ±1 standard error of the means.

sound and the Gabor were presented simultaneously. It should though be noted that in Lippert et al.'s study, only the null results in the simultaneous condition were reported, but not the results when the sound was presented either before or after the onset of the Gabor. Here, a significant difference was observed between those conditions in which the sound was presented 106 ms before the visual target and those conditions when the sound was presented 80 ms after, which may be attributed to an alerting effect induced by the presentation of the sound in the former case. That is, the participants actually processed the sound to some extent (since it influenced their performance); therefore, the null result observed at the 0 ms SOA was not due to participants simply ignoring the sound since its onset time varied with respect to that of the visual target.

Combining the results of Experiments 4.4A and 4.4B, it can be concluded that the reliable temporal coincidence of the sound and the visual target is critical to obtaining

the crossmodal facilitation elicited by that sound, whereas the predictability of the onset time of the sound and visual target was not. Lippert et al. (2007) argued that the effect that they obtained reflected a 'cognitive top-down effect' rather than an automatic sensory effect. That is, the crossmodal facilitation resulting from the presentation of a concomitant sound was modulated by certain cognitive factors, such as whether the visual and auditory stimuli were always presented together, which can be considered as an example of the 'unity assumption' (Spence, 2007; Welch & Warren, 1980, see Section 4.8.1).

Experiment 4.5 was designed to investigate whether spatial coincidence and temporal coincidence would conjointly enhance the crossmodal facilitation effect. which has been demonstrated to provide the largest crossmodal enhancement when compared to those conditions in which either a sound was simply temporally coincident with a visual target or else no sound was presented (e.g., Bolognini et al., 2005; Frassinetti et al., 2002; Ngo & Spence, 2010a, c). The prediction was that the presentation of a sound that was both spatially and temporally coincident should facilitate participants' letter identification performance more than a sound which was simply temporally coincident (but was spatially inconsistent).

4.7 Experiment 4.5

In this experiment, the sound was always presented simultaneously with the target,

while it could appear at either the same or opposite location as the target with an equal probability. There were six possible ISIs between the visual target and the mask.

4.7.1 Methods

4.7.1.1 Participants

Thirty-one volunteers (11 males, mean age 22 years) from the same subject pool as described in Experiment 4.1 took part in this experiment.

4.7.1.2 Stimuli, apparatus, design, and procedure

Two factors, sound location (consistent, inconsistent, and sound absent) and ISI (0, 13, 27, 40, 80, and 133 ms), were manipulated. One-third of the trials were consistent, inconsistent, and sound absent, respectively. As in Experiment 4.1, the target and mask could either be presented from the left or from the right. For the left target, the *consistent* condition occurred when the sound came from the pair of loudspeakers on the left, while the *inconsistent* condition occurred when the sound was presented from the pair of loudspeakers on the right. For the right target, the sound source was right and left for the consistent and inconsistent conditions, respectively. Given that the left and right loudspeakers were aligned horizontally with the position of the visual letters, and they produced the same sound at the same time, the participants should have perceived the sound as originating from the location of the letters in the consistent condition. There were 3 (sound location) x 6 (ISI) x 2 (target location) conditions, and each condition

was tested with 18 letters. Hence, there were 648 trials in total, which were divided into six blocks. It took approximately 45 minutes for participants to complete this experiment. In terms of the data analysis, the results from the left and right targets were collapsed. The other details were exactly the same as in Experiment 4.1.

4.7.2 Results and discussion

Eight participants failed to reach the criterion, leaving the data from 23 participants for further analysis. The accuracy of participants' target identification performance was transformed and submitted to a two-way ANOVA on the factors of sound location and ISI (see Figure 4.9). The main effects of sound location was significant ($F(2,44) = 4.44$, $MSE = 0.03$, $p < .05$, $\eta^2 = 0.17$). A *post-hoc* Tukey test revealed that the accuracy in the incosistent condition (68.7%) was significantly higher than that in the sound absent condition (66.0%; $p < .05$). The main effects of ISI was also significant ($F(5,110) = 255.91$, $MSE = 0.05$, $p < .0001$, $\eta^2 = 0.92$). A *post-hoc* Tukey test showed that, compared to the 133 ms ISI, the accuracy at all other ISIs was significantly lower (all *p*s $< .01$). The interaction between sound location and ISI was also significant ($F(10,220) = 1.89$, $MSE = 0.03$, $p < .05$, $\eta^2 = 0.08$). The simple main effect showed that the sound location factor was significant at the 0 ms ($F(2,264) = 3.44$, $MSE = 0.03$, $p < .05$, $\eta^2 = 0.02$), 27 ms ($F(2,264) = 3.13$, $p < .05$, $\eta^2 = 0.02$), and 40 ms ISIs ($F(2,264) = 7.42$, $p < .001$, $\eta^2 = 0.05$). *Post-hoc* Tukey tests revealed that, at the 0 ms ISI, accuracy was

higher in the inconsistent condition than in the consistent condition (48.5% vs. 42.8% respectively, $p < .05$); at the 27 ms ISI, accuracy was higher in the consistent condition than in the sound absent condition (65.6% vs. 59.9% respectively, $p < .05$); at the 40 ms ISI, accuracy was higher in the inconsistent condition than in both the consistent and sound absent conditions (73.5%, 66.8% , and 65.7% respectively, both *p*s < .01).

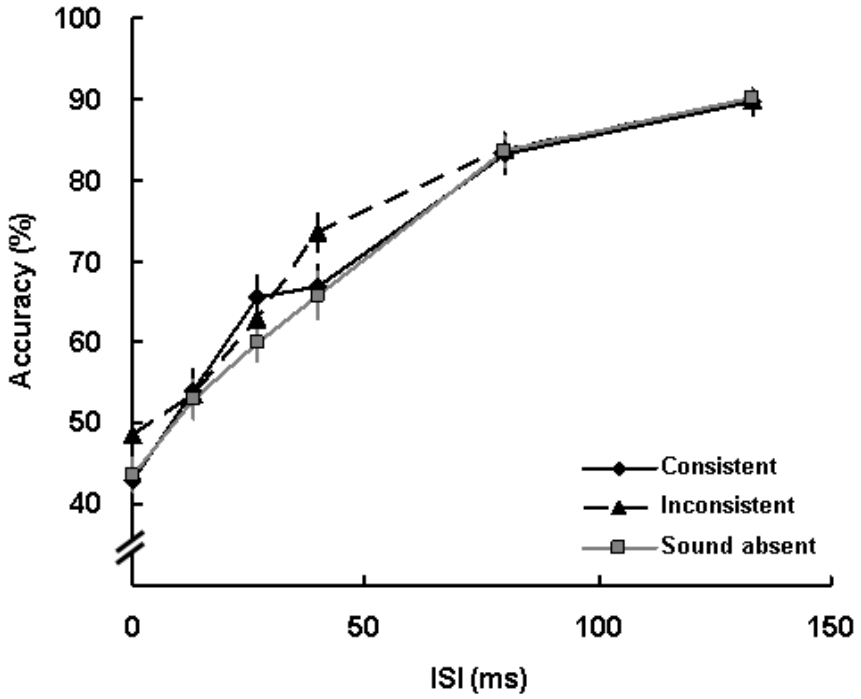


Figure 4.9. The mean accuracy of participants' letter identification performance in Experiment 4.5. The consistency in this experiment was manipulated in terms of the spatial consistency between the sound and the visual target. The error bars indicate ±1 standard error of the means.

The results of Experiment 4.5 showed that both the spatially consistent and inconsistent sound facilitated participants' target letter identification performance as compared to the sound absent condition within the time-window between 27 and 40 ms

ISI reported in Experiment 4.1, with the only difference being that the facilitation induced by the consistent sound (27 ms ISI) occurred at a shorter ISI than that induced by the inconsistent sound (40 ms ISI). That said, even though the spatial consistency between the sound and the visual target modulated the ISI at which the crossmodal facilitation effect was observed, temporal coincidence proved sufficient to elicit the auditory facilitation effect (rather than relying on both spatial and temporal coincidence) in participants' masked letter identification performance (though see Bolognini et al., 2005; Frassinetti et al., 2002; Ngo & Spence, 2010a, c). It has been demonstrated elsewhere that sound can influence visual perception in the temporal domain whether or not the stimuli are presented from the same position (e.g., Innes-Brown & Crewther, 2009; Vroomen & Keetels, 2006; see the results of simple manual response in Experiment 3 in Hughes et al., 1994). Speech perception, as indexed by effects such as McGurk effect, has also been shown to be unaffected by whether the face and voice come from the same location or not (e.g., Jones & Jarick, 2006; see Spence, in press, for a review)[17].

Experiment 4.5 was the first occasion on which any difference in performance elicited by the presentation of the sound at the 0 ms ISI was observed. However, in this

[17] One may wonder whether the magnitude of the spatial disparity between the sound and visual target is critical to demonstrating an effect of spatial coincidence. In those studies that showed a spatial modulation, the spatial disparity between the sound and the visual target was greater than 16° in Bolognini et al. (2005) and Frassinetti et al., 2002, and same/different hemifield in Ngo and Spence (2010a, b). In those studies that failed to show the modulation of spatial coincidence, the spatial disparity was 14° in the present study, 30° in Hughes et al., 1994, 20° in Innes-Brown and Crewther (2009), ranging between 0° to 180° in Jones and Jarick (2006), and ranging between 10° to 90° in Vroomen and Keetels' (2006) study.

case, the difference was induced by the spatial consistency of the sound. It should be noted, given that participants' performance was better in the inconsistent than in the consistent condition, the results therefore still cannot be explained by the signal-enhancement hypothesis (cf. Stein et al., 1996; Stein & Meredith, 1993). It may be that if participants' attention was exogenously directed to the location of the sound (e.g., McDonald et al., 2000; Spence & Driver, 1997), any influences of attention should be more likely on the processing of the visual mask at the 0 ms ISI. That is, in the inconsistent condition, the mask might be attended less and therefore exerted less of an influence over participants' identification of the target. On the other hand, if both the presentation of a spatially-consistent sound or the target letter itself (when no sound was present) can exogenously attract a participant's attention to the location where the letters are presented, the visual mask should thus be enhanced and elicited similar masking effects in both the consistent and sound absent condition. The question of whether this explanation is valid and why the difference only occurred at the 0 ms ISI requires further study.

4.8 General discussion

The experiments reported in this chapter were designed to test whether, and under what conditions, the presentation of a concomitant sound can facilitate the identification of a visual target letter that happens to be masked by a subsequently-presented letter.

The results first demonstrated that the crossmodal facilitation of participants' visual letter identification performance only occurred when the ISI between the visual target and the mask was 27 and 40 ms, a timecourse that happens to lie just beyond the threshold needed to differentiate the target and mask as constituting distinct letters (Experiment 4.1). This crossmodal facilitation effect can be dissociated from the facilitation induced by a brighter visual target that occurred at 0 ms ISI (Experiment 4.2). When the onset of the sound was synchronized with that of the mask, participants' performance was not impaired; in fact, a facilitatory effect was still observed (Experiment 4.3). The following experiments then went on to demonstrate that the crossmodal facilitatory effect only occurred when the sound reliably accompanied the visual target (i.e., when the sound, if it was present, always coincided temporally, with the visual target; Experiment 4.4). However, the results of our final experiment showed that the sound did not need to be spatially coincident with the visual target in order for it to elicit the crossmodal facilitation effect (Experiment 4.5).

One conclusion that can therefore be drawn is that the presentation of a concomitant sound can facilitate the identification of a masked visual target under two conditions: First, the visual target should constitute a distinct event from other visual events (i.e., the masks or distractors). In the present study, when the target and mask were both presented for 40 ms, and were both spatially overlapping and physically similar, the optimal interval between the two events to observe the crossmodal

facilitation effect was 27 to 40 ms. Second, the sound, if present, should always be presented at the same time as (or within a certain interval of, see Roseboom et al., 2009; Spence & Squire, 2003; Zampini, Guest, Shore, & Spence, 2005) the visual target. That is, crossmodal stimulation that simply follows the temporal coincidence rule proved insufficient to guarantee crossmodal facilitation; the visual target and the auditory accessory stimulus also had to co-occur regularly at the same time. It is possible that the reliable temporal coincidence of the visual and auditory stimuli may activate the top-down unity assumption which then modulates audiovisual integration (Experiment 4.4; see also Lippert et al., 2007). Hypothetically, then, these two conditions can be considered as bottom-up and top-down factors that together modulate audiovisual integration (Spence, 2007; Welch & Warren, 1980).

4.8.1 Bottom-up and top-down factors modulating audiovisual integration

In terms of bottom-up factors, for example, the visual and auditory stimuli should be segregated from any other stimuli that happen to be presented in the same modality so that they can be matched and integrated (see Spence & Chen, in press). In previous studies that have demonstrated an audiovisual facilitatory effect, each potential target frame was separated by more than 100 ms (e.g., Chen Y.-C. & Yeh, 2008, 2009; Ngo & Spence, 2010a, b, c; Olivers & van der Burg, 2008; van der Burg et al., 2008a; Vroomen & de Gelder, 2000). This presentation rate may be partly due to limitations of temporal

resolution in terms of detecting audiovisual synchrony (e.g., Fujisaki & Nishida, 2005, 2009; Lewald et al., 2001; van der Burg, Cass, Olivers, Theeuwes, & Alais, 2010; see the General Discussion in van der Burg et al., 2008a, and the Acknowledgments section of Olivers & van der Burg, 2008). Even though the SOA between the visual target and the mask needed in order to observe crossmodal facilitation in the present study was only 67 or 80 ms (i.e., an ISI of 27 and 40 ms, respectively), the interval was sufficiently long for participants to perceive the visual target and mask as distinct perceptual events. This clear temporal structure would have allowed the sound to be matched with the simultaneously-presented visual target and thence likely integrated with it (see also van der Burg et al., 2010, p. 9). That is, the crossmodal facilitation observed in the present study is object- (or event-) based. For those conditions in which the ISI was even shorter (e.g., 0 and 13 ms), the visual target and mask were presumably merged into a single event file and hence, even if the sound integrated with the *event*, both the processing of the target and mask would likely be enhanced. That said, due to the fact that no crossmodal facilitation was elicited by sound at the 0 and 13 ms ISI, the signal-enhancement mechanism therefore cannot work in the situation where a mask was used to control the target duration or task difficulty.[18]

In terms of top-down factors, for example, when a sound and a visual stimulus are

[18] On the other hand, if one plans to test the signal-enhancement hypothesis, the task of detecting or identifying a single target while comparing the sound present vs. sound absent condition may constitute a better choice (see Spence & Ngo, in press). However, at present, there still seems to be lacking any obvious evidence supporting the signal-enhancement hypothesis (see Anderson & Mamassian, 2008; Arieh & Marks, 2008; Lippert et al., 2007; Marks et al., 2003; Odgaard et al., 2003).

always presented together (i.e., synchronized), they can be considered as constituting a single perceptual event. This is called the 'unity assumption' (e.g., Vatakis & Spence, 2007; Welch & Warren, 1980). This top-down modulation facilitates audiovisual integration irrespective of whether the onset time of the sound and visual stimulus was predictable or not (see Experiment 4.4A). However, if the onset time of the auditory and visual stimuli vary with respect to one another, they should be more likely to be perceived as constituting distinct perceptual events, such as in the case of thunder and lighting. When the sound and the visual stimulus do not integrate, no crossmodal facilitation should be observed (as shown in Experiment 4.4B, see also Lippert et al., 2007). It should be noted that even under those conditions in which the participants were not informed that the sound and the visual target would be presented at the same time, as in the present study, participants would nevertheless presumably still have been able to figure out the implicit temporal relationships (see Lippert et al., 2007, p. 103). In other words, the unity assumption can emerge as a result of either knowledge, or by observation of the statistical properties of stimulation across sensory modalities within a particular experimental context (see also Ernst, 2007).

4.8.2 How does the presentation of the sound enhance visual perception?

Considering the three hypotheses outlined in the Introduction: preparedness-enhancement (Nickerson, 1973), signal-enhancement (Stein et al., 1996),

and object-enhancement (Chen Y.-C. & Yeh, 2009), the experiments reported in this chapter support the last hypothesis. The former two hypotheses are contradicted by the fact that the crossmodal facilitatory effect elicited by the presentation of the sound was not observed across all ISIs, and critically, no crossmodal facilitation was observed at the 0 ms ISI. The evidence supporting the object-enhancement hypothesis comes from the fact that the timecourse of crossmodal facilitation was closely linked to the timecourse of the target and mask events being perceptually distinguishable (see Navon & Purcell, 1981; Reeves, 1982). Note also that this crossmodal facilitation was dissociated from the influence of the relative luminance of the visual target and mask (Spencer & Shuntich, 1970). The accessory auditory stimulus was therefore likely integrated with the accompanying visual target to build a more consolidated multisensory object representation, which, as a result, was less likely to be interrupted/substituted by the representation of the subsequently-presented masking event. This multisensory object representation should then be better encoded into memory or used to respond according to the task at hand. It may also be true to say that, when the sound was present at the same time as the visual mask, the multisensory object representation of the mask was clearer and hence was less likely to be confused with the visual target event.

The idea of multisensory object representations can perhaps also be used to help explain other crossmodal facilitation effects that have been reported in the literature in

recent years. So, for example, the fact that the sensitivity of participants' visual target detection is higher for crossmodal events than for unimodal events, since the former contain more information to signify the occurrence of the stimulus (e.g., Bolognini et al., 2005; Frassinetti et al., 2002; Lippert et al., 2007). It should also be likely that the crossmodal event can be better perceived due to its lower likelihood of being interrupted by the subsequently-presented visual event in RSVP or dynamic visual search displays (Ngo & Spence, 2010a, b; van der Burg et al., 2008a; Vroomen & de Gelder, 2000). As a result, crossmodal integration can ameliorate the visual deficit seen in normal observers, such as reducing the occurrence of the RB and AB processing deficits (Chen Y.-C. & Yeh, 2008, 2009; Olivers & van der Burg, 2008), the phenomena which have been argued to result from a failure to construct a complete object representation of the visual target (see Chun, 1997; Chun & Potter, 1995). On perceiving the salient crossmodal event, observers may describe it as brighter (Stein et al., 1996) or as lasting longer than the other events (as in the freezing effect, Vroomen & de Gelder, 2000; see Spence & Ngo, in press, for a review). However, it seems possible that such subjective reports describing the crossmodal event as clearer were based simply on observers' intuitions concerning how a visual stimulus can be better perceived (i.e., either it is brighter or presented for longer, like using their tacit knowledge, cf. Farah, 1988), rather than necessarily reflecting a change in the content of their conscious experience (i.e., genuinely perceived brighter or as perceived as persisting for longer, see also Noesselt

et al., 2008; Odgaard et al., 2003; Spence & Ngo, in press).

In addition, given that some of the phenomena reported above, such as the popping-out of crossmodal events (in the pip-and-pop effect) and the reduction of AB are closely linked to visual attention, crossmodal facilitation is therefore likely mediated by the boosting of attention (e.g., Olivers & van der Burg, 2008; van der Burg et al., 2008a). However, it should be noted that some attention-related factors manipulated in the present study, such as the uncertainty of the onset timing of the audiovisual event in Experiment 4.4A, and the spatial consistency between the sound and the visual target manipulated in Experiment 4.5, did not enlarge or eliminate the crossmodal facilitation effect reported here. Hence, it is likely that the crossmodal event itself elicits and attracts participants' attention in a relatively bottom-up manner (e.g., Santangelo & Spence, 2007; van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008b). This idea is supported by the results of a recent event-related potential (ERP) study showing that simultaneously-presented auditory and visual stimuli may be first integrated, resulting in the construction of a salient crossmodal event; and, as a result, this crossmodal event can attract attention to enhance its later processing (Busse, Roberts, Crist, Weissman, & Woldorff, 2005; Spence, 2010b; Talsma et al., 2007; van der Burg et al., submitted; for a review, see Talsma et al., 2010). Therefore, even though attention may enhance crossmodal facilitation, attention is elicited by the audiovisual event itself. Alternatively, attentional modulation can actually be seen as an emergent property of audiovisual

integration (e.g., Driver & Vuilleumier, 2001; see Spence et al., in press). One open question remains as to whether the audiovisual stimuli would elicit this attentional modulation more rapidly when they were spatially coincident than when they were not (such a finding would be consistent with a channel selection induced by exogenous attentional orienting view, cf. Prinzmetal, McCool, & Park, 2005), such as was observed in Experiment 4.5.

4.8.3 Audiovisual integration at the level of object representation

A critical paradox that needs to be resolved here is: How does the presentation of a sound enhance the object representation of an accompanying visual event without necessarily enhancing its signal? In order to answer this question, one first needs to consider what information is constitutive of an object representation. An object representation presumably contains information regarding not only what it is (i.e., identity information, or 'what', or 'type') but also the time and space that it appeared (i.e., spatiotemporal information, or 'when' and 'where', or 'token'; see Treisman & Kanwisher, 1998). A complete visual object representation therefore relies on the binding together of the activated identity and spatiotemporal information, and attention serves as a critical supplementary component in this binding process (Bowman & Wyble, 2007; Chun, 1997; Treisman & Kanwisher, 1998).

Several cognitive frameworks have proposed very similar dissociations between identity and spatiotemporal information, such as embodied by the notion of *object files*

(Kahneman, Treisman, & Gibbs, 1992), the *visual indexing model* (Pylyshyn, 1989, 2001), and the *type-token framework* (Bowman & Wyble, 2007; Driver & Vuilleumier, 2001; Kanwisher, 1987, 2001; Marcel, 1983). There is also evidence from neuroanatomical studies to support the existence of separate 'where' and 'what' pathways (e.g., Mishkin, Ungerleider & Macko, 1983; Ungerleider & Mishkin, 1982; further dissociated as 'what' and 'how', by Milner & Goodale, 1995). The where and what pathways in vision can further be linked to the magnocellular and parvocelluler pathways, respectively (DeYoe & van Essen, 1988; Grossberg, 1994). More recently, this dissociation between the 'what' and 'where (or how)' pathways has been extended to both the auditory (Alain, Arnott, Hevenor, Graham, & Grady, 2001; Kaas & Hackett, 1999; Kubovy & van Valkenburg, 2001; Rauschecker & Tian, 2000; Romanski et al., 1999) and somatosensory modalities (Chan & Newell, 2008; Dijkerman & de Haan, 2007; Reed, Klatzky, & Halgren, 2005; see also Renier et al., 2009).

One can therefore consider visual and auditory stimulation in terms of a crossmodal type-token binding framework (see Figure 4.10). For the visual target in the present study, the time and location it appeared constituted the token representation (labeled *Vtoken*), while the identity of the letter constituted the type representation (labeled *Vtype*). Similarly, for the sound, the time and location it appeared constituted the token representation (labeled *Atoken*), while the pitch of the sound constituted the type information (labeled *Atype*). According to a framework in which type and token

representation are differentiated, the surface brightness of a visual stimulus should belong to the type representation (see Breitmeyer et al., 2006; Ogmen et al., 2003). Hence, it is possible that unless the visual and auditory stimuli are associated in terms of their semantic meaning (e.g., Chen & Spence, 2010; Schneider, Debener, Oostenveld, & Engel, 2008; Doehrmann & Naumer, 2008) or synaesthestic correspondence (e.g., Evans & Treisman, 2010; Parise & Spence, 2009; Spence, submitted), Vtype should not be enhanced by the presentation of a semantically-irrelevant Atype (e.g., Andersen & Mamassian, 2008; Arieh & Marks, 2008; Odgaard et al., 2003).

On the other hand, the Vtoken pathway should be sensitive to any transient change, even one occurring in a different sensory modality. For example, crossmodal facilitation in terms of detecting an *increment* in visual luminance can still be observed even when the accompanying auditory stimulus actually consists of a *decrease* in the loudness of a sound (and *vice versa*, see Anderson & Mamassian, 2008), or in terms of detecting a blinked-off visual target by the presence of a sound (Noesselt et al., 2008). That is, the crossmodal facilitation of a participant's ability to detect a visual target change can rely solely on the information concerning a transient irrespective of its identity (being either an increment or decrement in loudness, in this case). Therefore, the suggestion is that Vtoken and Atoken should integrate in terms of their temporally-coincident transient, which would provide more precise episodic information than Vtoken when considered by itself (e.g., Fendrich & Corballis, 2001; Vroomen & de Gelder, 2004). The idea of

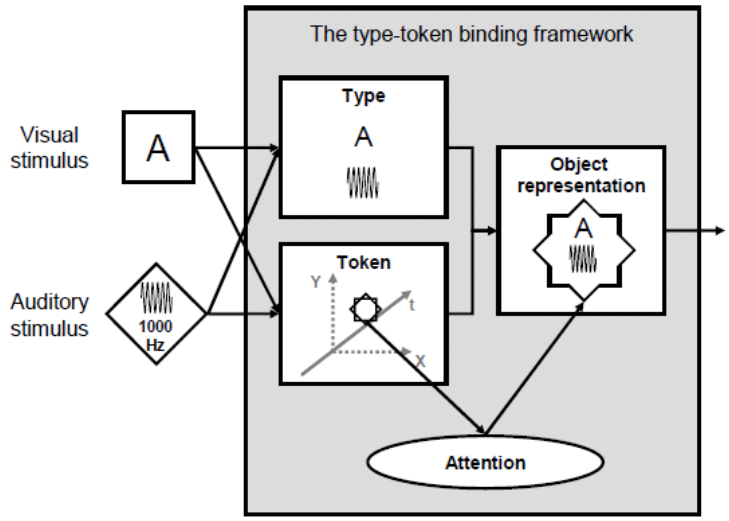


Figure 4.10. Schematic representations of the crossmodal type-token binding framework. Both the visual stimulus (the letter A) and the auditory stimulus (the 1000 Hz pure tone) initially activate their corresponding type and token representations (e.g., Bowman & Wyble, 2007), and eventually bind to form a multisensory object representation. In the type system, Vtype (A) and Atype (1000 Hz tone) were represented; however, since they were neither semantically nor synaethestically associated, they should not enhance (or integrate with) each other. In the token system, given that temporal coincidence was sufficient to obtain the crossmodal facilitation effect in the present study, while spatial coincidence was not, the hypothesis here is that in this system the temporal domain (t) is represented (axis with a solid grey line), whereas the spatial domain (X and Y, respectively) remains unspecified (axes with dotted grey lines). Given that the visual target and the auditory accessory stimulus were presented simultaneously, the Vtoken (□) and Atoken (◇) should integrate; further, this multisensory token may direct attention to enhance the processing of that which they refer to (Busse et al., 2005). The multisensory object representation should then be encoded into memory or else used to select a response.

precise episodic representation may be similar to the notion of *temporal uncertainty reduction* proposed by Lippert et al. (2007), according to which the presentation of an informative sound that signifies the onset of the visual target would enhance the participants' sensitivity to it. This integrated multisensory token then combined the Vtype and Atype information in order to construct a multisensory object representation

(for a similar framework, see Fujisaki, Kitazawa, & Nishida, in press; Fujisaki & Nishida, 2005, 2010). This precise episodic representation may also direct a participant's attention to boost the binding of type and token information and then result in the construction of a complete multisensory object representation (e.g., Busse et al., 2005).

Note that the different results for the detection and identification task reported in Experiment 1 can also be explained in terms of the crossmodal type-token binding framework. Fujisaki and Nishida (2010, p. 2287) suggested that multisensory integration in terms of synchrony occurred in the 'when' pathway, while the multisensory integration in terms of stimulus attributes occurred at a later stage involving the binding of the information from 'when' and 'what' pathways. That said, crossmodal facilitation in the identification task relies on the processing of higher-stage binding, which therefore requires more time to process, as compared to that in the detection task (Fujisaki & Nishida, 2010).

4.8.4 Linking behavioral and physiological evidence

As regards the backward masking paradigm, the results reported in this chapter suggest that the integrated multisensory object representation of the visual target is less likely to be interrupted or substituted by the representation of the mask than a unisensory visual object representation. In terms of interruption masking, the idea is that

the mask prevents the target event from being encoded into memory before its iconic information decays (Spencer & Shuntich, 1970). A stronger episodic representation enhanced by a concomitant sound can therefore help a visual target to become more consolidated (see the previous section). In terms of substitution masking, the idea is that when visual information keeps feeding forward to higher levels, the higher level sends feedback by means of re-entrant connections; when the incoming bottom-up information is mismatched with the re-entrant information, the new information overwrites the old information and therefore substitution masking occurs (see Di Lollo et al., 2000; Enns & Di Lollo, 2000). If audiovisual integration enhances the representation of the target event, the re-entrant information may be weighted more heavily than the new information and therefore reduces the likelihood of masking taking place. These two kinds of masking, though, cannot be differentiated on the basis of the results reported in the present study; however, the distinction needs to be kept in mind since they may have different implications for multisensory integration.

Considering that audiovisual integration may enhance the representation of an object either directly via consolidation or through a re-entrant pathway, these two possibilities can be mapped to the possible accounts of how multisensory integration influences sensory-specific processing proposed by Driver and Noesselt (2008). That is, audiovisual integration either occurs directly in a multisensory brain area (Account B for 'New bimodal brain areas' in their review) or via a top-down feedback loop

(Account C-F-C for 'critical role of feedback circuitry' in their review). The difference between these two possibilities lies not only in whether the pathway is direct or indirect, but also in the fact that the latter account involves attentional modulation (see Di Lollo et al., 2000; Driver & Noesselt, 2008; Enns & Di Lollo, 2000). As the suggestion that audiovisual multisensory integration may increase the token representation of the visual event and attract attention to enhance its representation (see Busse et al., 2005; van der Burg et al., submitted), both mechanisms likely work together to contribute to the audiovisual facilitatory effects reported here.

Finally, it should be noted that, as concerns the three basic rules proposed by Stein and Meredith (1993) on the basis of their neurophysiological data, the results in this chapter actually showed that only temporal coincidence was necessary, while the spatial coincidence was not. Furthermore, Experiments 4.1 through 4.4 did not provide any facilitatory effect induced by sound in the most difficult condition (i.e., 0 ms ISI), which is therefore inconsistent with the inverse effectiveness rule. It is true that when exploring the effect of crossmodal integration on human performance, the task should perhaps be difficult enough so that a facilitatory effect, if any, can be observed (e.g., Bolognini et al., 2005; Frassinetti et al., 2002). However, when task difficulty is itself considered as a function, inverse effectiveness may not necessarily occur (as in the present study, see also Lippert et al., 2007; Ma et al., 2009; Ross et al., 2007). That is, the three basic rules that have been repeatedly demonstrated at the single-cell level in

studies of superior colliculus neurons need to be re-evaluated in terms of human behavioral studies, especially when the participants' task involves target identification (see also Schnupp, Dawe, & Pollack, 2005). It is possible that the enhanced activation of neurons in the superior colliculus serves primarily in terms of the direction of an organism's spatial attention (see Lovejoy & Krauzlis, 2010; Muller, Philiastides, & Newsome, 2005).

4.9 Conclusions

The experiments reported in this chapter have demonstrated that the presentation of an accessory sound can facilitate the identification of a visual target presented in a backward masking paradigm. In terms of the three possible mechanisms (preparedness-enhancement, signal-enhancement, and object-enhancement) outlined in Section 4.1, the presentation of the sound was suggested to enhance the representation of the visual target and hence reduce the likelihood of it being interrupted or substituted by the subsequently-presented mask. By utilizing the type/token framework that is well-established in the study of visual information processing (see Bowman & Wyble, 2007), this crossmodal facilitation can attribute to the presentation of the sound enhancing the token representation of the visual object and, in turn, the whole multisensory object representation containing the visual target end up being better consolidated, and hence easier to report.

Chapter 5

Crossmodal constraints on human visual awareness: Can auditory semantic context modulate binocular rivalry?

5.1 Introduction

When viewing a scene, background context provides useful semantic information that can improve the identification of a visual object embedded within it, such as when the presentation of a kitchen scene facilitates a participant's ability to identify a loaf of bread, say (e.g., Biederman, 1972; Davenport & Potter, 2004; Palmer, 1975; though see Hollingworth & Henderson, 1998). Importantly, however, our environments typically convey contextual information via several sensory modalities rather than just one. So, for example, when we are at the seaside, we perceive not only the blue sea and sky (hopefully), but also the sound of the waves crashing onto the beach, not to mention the smell of the salty sea air. Do such non-visual contextual cues also influence the visual perception of semantically-related objects? In this chapter, I investigate whether the

semantic context provided by stimuli presented in another sensory modality (in this case, audition) modulates the perceptual outcome in vision; namely, visual awareness.

The phenomenon of binocular rivalry provides a fascinating window into human awareness (e.g., Crick, 1996). Binocular rivalry occurs when two dissimilar figures are presented to corresponding regions of the two eyes. Observers perceive one of the figures as dominant (while often being unaware of the other); after a while, the dominance of the figures may reverse and then keep alternating over time. This perceptual alternation has been attributed to the fact that the visual system receives ambiguous information and tries to find a unique perceptual solution, and therefore the two images compete for control of the current conscious percept (see Alais & Blake, 2005, for a review). The fact that a constantly-presented dichoptic figure induces alternating perceptual experiences in the binocular rivalry situation demonstrates the dynamic way in which the brain computes sensory information, a process that gives rise to a specific perception (e.g., Leopold & Logothetis, 1996).

Several studies have tried to understand how visual awareness emerges in the binocular rivalry situation. According to Helmholtz's (1866/1962) early view, the alternation of perceptual dominance is under voluntary attentional control. Subsequently, it has been argued that the phenomenon occurs as a result of competition either between two monocular channels (Levelt, 1965; Tong & Engel, 2001) or else between two pattern representations, one presented to each eye (Leopold & Logothetis, 1996;

Logothetis, Leopold, & Sheinberg, 1996; Tong, Nakayama, Vaughan, & Kanwisher, 1998). More recent models (e.g., Tong, Meng, & Blake, 2006) have suggested that the mechanism underlying binocular rivalry includes not only competition at multiple levels of information processing (for a review, see Blake & Logothetis, 2002; Tong, 2001), but also some form of excitatory connections that facilitate the perceptual grouping of visual stimuli (Alais & Blake, 1999; Kovacs, Papathomas, Yang, & Feher, 1996), as well as top-down feedback, such as attentional control (Chong, Tadin, & Blake, 2005; Meng & Tong, 2004; Mitchell, Stoner, & Reynolds, 2004; van Ee, van Dam, & Brouwer, 2005). That said, the underlying mechanisms giving rise to conscious perception in the binocular rivalry situation likely involve a variety of different neural structures throughout the visual processing hierarchy.

Given that the phenomenon of binocular rivalry is, by definition, visual, one might expect that the perceptual outcome for ambiguous visual inputs should thus be generated entirely within the visual system (cf. Hupé, Joffo, & Pressnitzer, 2008). On the other hand, however, some researchers have started to investigate whether visual awareness can be modulated by the information presented to another sensory modality. So, for example, Sheth and Shimojo (2004) have reported that the presentation of synchronous auditory cues can help to maintain the awareness of a visual stimulus by preventing visual filling-in. Similar evidence has emerged from a binocular rivalry study showing that the dominance duration of a looming (or rotating) visual pattern can

be enhanced when the rate of which happened to be synchronous with a series of pure tones or vibrotactile stimuli (or their combination, see van Ee, van Boxtel, Parker, & Alais, 2009). It is therefore worth testing whether crossmodal interactions in the binocular rivalry situation can be extended to the case of audiovisual semantic congruency, a high-level factor which has recently started to capture the attention of a growing number of researchers (e.g., Chen & Spence, 2010; Iordanescu et al., 2008; Molholm et al., 2004; Noppeney et al., 2008; Schneider, Engel, & Debener, 2008; Taylor, Moss, Stamatakis, & Tyler, 2006).

The experiments reported in this chapter were designed to investigate whether the semantic context embedded in a background auditory soundtrack can modulate the dominant percept under conditions of binocular rivalry. After demonstrating the auditory semantic contextual effect on binocular rivalry, I then go on test whether this crossmodal effect interacts with two visual factors, namely, selective attention and stimulus contrast (Meng & Tong, 2004). An explanation of this crossmodal effect in terms of a response tendency that might have been biased by the meaning of the soundtrack was further ruled out. Finally, these findings are discussed in terms of current models of binocular rivalry.

5.2 Experiment 5.1

The first question to be addressed in the present chapter was whether the semantic

content of a background soundtrack would increase the likelihood of the congruent visual percept being dominant under conditions of binocular rivalry. The participants viewed a dichoptic figure consisting of a bird and a car (see Figure 5.1) and continuously reported which figure was dominant. During the test period, the participants viewed the visual stimuli alone (in the no-sound condition) or else they also heard a bird, car, or restaurant soundtrack at the same time.



Figure 5.1. An example of the dichoptic stimulus pairs used in Experiment 5.1.

5.2.1 Methods

5.2.1.1 Participants

Fourteen volunteers (including the author, five males, mean age 24 years) took part in Experiment 5.1 in exchange for a £10 (UK Sterling) gift voucher. All of the participants had normal or corrected-to-normal vision and hearing by self-report. The participants were tested using depth-defined figures embedded in red-green random-dot stereograms to ensure that they had normal binocular vision.

5.2.1.2 Apparatus and stimuli

The visual stimuli were presented on a 15-inch color CRT monitor (75 Hz refresh rate). The participants sat at a viewing distance of 58 cm from the monitor in a dimly-lit experimental chamber. The visual test stimuli consisted of the outline-drawings of a bird (7.85°×6.39°) and car (8.34°×3.95°) taken from Bates et al. (2003). The two figures were spatially superimposed, with the bird presented in red (RGB (255, 0, 0), CIE (0.621, 0.341)) and the car in cyan (RGB (0, 255, 255), CIE (0.220, 0.347)), or vice versa, against a white background (RGB (255, 255, 255), CIE (0.293, 0.332)). These two color versions of the visual pictures (bird in red and car in cyan, or bird in cyan and car in red) were used to balance the influence of participants' dominant eye when viewing dichoptic figures. The participants wore glasses with a red filter on the left eye and a cyan filter on the right eye during the course of the experiment.

Three sound files, *bird* (consisting of birds singing in a forest), *car* (consisting of car horn and engine-revving sounds in a busy street), and *restaurant* (consisting of the sound of tableware clattering together in a restaurant), which were recorded in realistic environments (downloaded from www.soundsnap.com on 06/11/2008) were used as the auditory soundtracks. The sound files were edited so that the auditory stimulus started from the beginning of the sound file and lasted for 65 seconds. The sounds were presented over closed-ear-cup headphones and ranged in loudness from 55 to 68 dB

SPL[19].

5.2.1.3 Design and procedure

The factor of sound (no-sound, bird, restaurant, or car) was manipulated. On each trial, instructions were presented on the screen informing the participants that they should press the '1' key whenever the image of the bird was dominant, and the '0' key whenever the image of the car was dominant. The participants initiated each trial by pressing the 'SPACE' bar. A blank screen was presented for 5 seconds, followed by the presentation of the dichoptic figures for 60 seconds. The participants were instructed to fixate the area of the bird's wing and car door and to start reporting which figure was dominant as soon as the dichoptic figures were presented; they were also instructed to pay attention to the content of the sound as well (in order to ensure that the soundtracks were processed; see van Ee et al., 2009, Experiment 4). At the end of the trial, the question 'What sound did you just hear?' was presented on the monitor, and the participants had to enter their answer (free report) using the keyboard. The sound was presented from the onset of the blank frame until the offset of the visual stimuli, in order to allow participants sufficient time to perceive the semantic context conveyed by the soundtrack.

Three blocks of experimental trials were presented. In each block, four sound

[19] The sound pressure delivered by the headphones was measured as follows: A CD disc was used to cover the headphone and the sound-level meter was measured 3 cm away from the central hole of the CD.

conditions were tested using the two color versions of the pictures. The order of presentation of these 8 trials was randomized. Prior to the completion of the experimental blocks, a practice block containing 8 no-sound trials was presented in order to familiarize the participants with the task. The participants were instructed to establish their criterion for reporting the exclusive dominance of each of the figures, and to try and hold this criterion constant throughout the experiment. The experiment lasted for approximately one hour.

5.2.2 Results and discussion

Two participants were excluded because they only reported seeing a single figure during the test period in one-third of the trials, leaving the data from 12 participants for further analysis. Three performance indices were used: First, the proportion of time for which the bird percept was dominant (i.e., the dominance duration), was calculated by dividing the sum duration of each bird percept by the sum duration of both the bird and car percept within the test period (thus the dominance duration of the bird and car views were reciprocally-related). The second index was the number of times that *the first percept was bird* out of six trials in each sound condition. The third index was the average *number of switches* between the bird and car percept that took place during test period in each sound condition. This index links closely to voluntary control in the binocular rivalry situation. In particular, when the participants are trying to maintain a

particular percept, they are able to delay the switch to the other percept (see van Ee et al., 2005, p. 51), which should result in fewer switches during the test period. A separate one-way ANOVA was conducted on the sound factor (no-sound, bird, restaurant, and car) for each index.

Analysis of the dominance duration data (see Figure 5.2A) revealed a significant main effect of sound ($F(3,33) = 6.94$, $MSE = 0.01$, $p < .005$, $\eta^2 = 0.39$). Critically, a *post-hoc* Tukey test showed that the dominance duration of the bird percept was significantly higher in the bird-sound condition than in the car-sound condition ($p < .01$). Similarly, an analysis of the first percept (see Figure 5.2B) revealed a significant main effect of the sound factor ($F(3,33) = 3.58$, $MSE = 1.04$, $p < .05$, $\eta^2 = 0.25$). A *post-hoc* Tukey test showed that the participants perceived the bird as the first percept more frequently when listening to the bird sound than when listening to the car sound ($p < .05$). Analysis of the number of switches (see Figure 5.2C) was, however, not significant ($F(3,33) = 1.11$, $MSE = 6.02$, $p = .36$, $\eta^2 = 0.09$).

One may wonder whether the auditory semantic context primarily worked to prolong the dominance duration of the semantically-congruent figure, or to shorten the dominance duration of the semantically-incongruent figure, as compared to those durations in the no-sound condition (or, more fairly, as compared to those durations in the irrelevant-sound condition). This question is raised because the results of previous studies suggest that the modulation of binocular rivalry by voluntary attention resulted

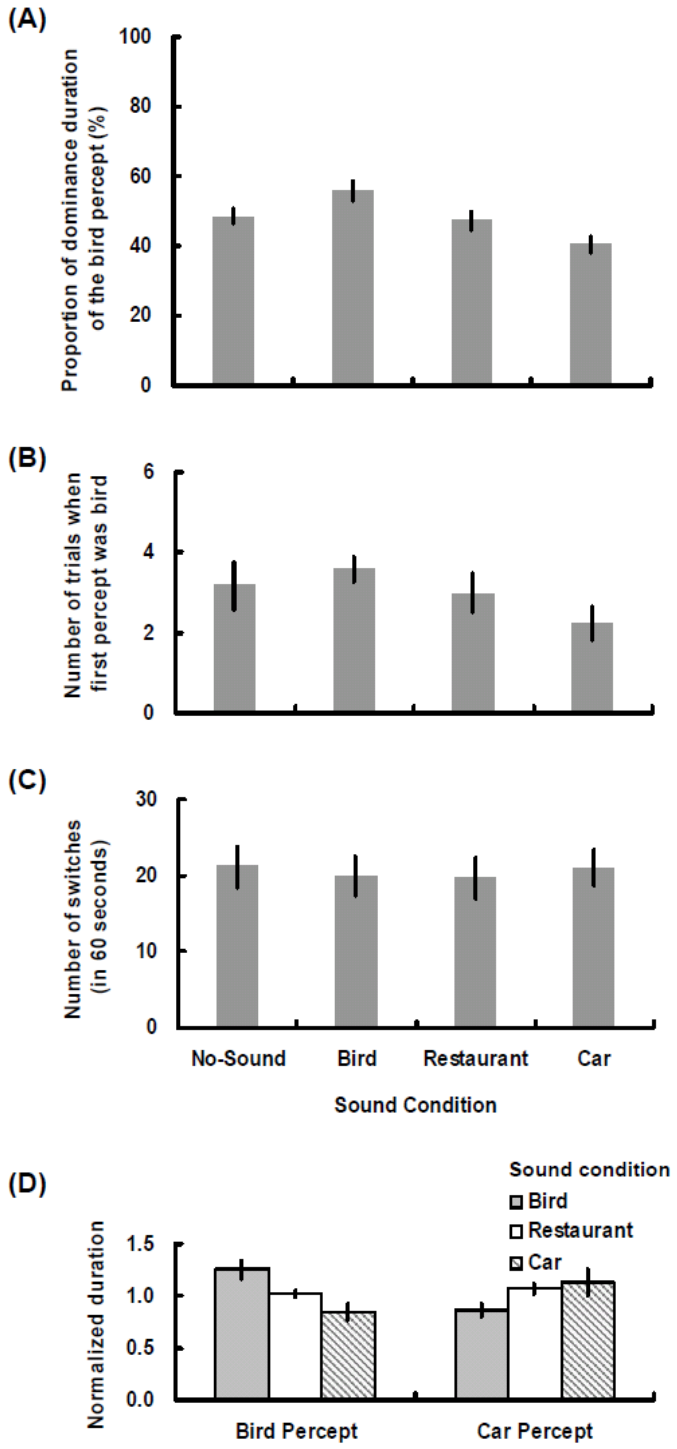


Figure 5.2. Results of the four sound conditions (no-sound, bird, restaurant, and car sound) under conditions of binocular rivalry in Experiment 5.1. (A) The dominance duration of the bird percept (proportion of total dominance duration). The gray lines represent data from each participant, and the black line represents the group mean; (B) The number of times that participants reported the bird view as the first percept (out of 6 trials); and (C) The average number of switches taking place in each trial (during the course of 60 seconds). (D) The normalized duration of the bird and car percept (normalized by the durations in the no-sound condition) in the bird, restaurant, and car sound conditions.

from a prolonging of the dominance duration of the target percept (Chong et al., 2005), whereas increasing the saliency of figure in one eye was proposed to shorten the dominance duration of the figure in the other eye (Levelt, 1965). The mean duration of both the bird and car percepts in the four sound conditions was therefore calculated.

The duration data of each participant in the bird, restaurant, and car sound conditions were normalized by their own duration in the no-sound condition in order to account for individual differences in dominance duration (see Chong et al., 2005). For the bird percept (see Figure 5.2D), the duration was longer in the bird sound condition than in the restaurant sound condition ($t(11) = 2.29$, $p = .021$, one-tailed), and the duration was longer in the restaurant sound condition than in the car sound condition ($t(11) = 2.49$, $p = .015$). The duration of the car percept was shorter in the bird sound condition than in the restaurant sound condition ($t(11) = 2.80$, $p = .009$), while there was no significant difference between the restaurant sound and car sound conditions ($t(11) = 0.51$, $p = .31$). These results show that the presentation of the auditory semantic context was more likely to shorten the duration of the semantically-incongruent percept than to prolong the duration of the semantically-congruent percept, since the former effects were more reliable. However, at this point, this conclusion cannot be strongly supported, because a more dedicated means of estimating the precise dominance duration for each percept is required.

The results of Experiment 5.1 demonstrate that the participants reported perceiving

the bird percept to be dominant more of the time when they were listening to the bird soundtrack than when listening to the car soundtrack. The dominance duration of the bird percept was intermediate between the bird- and car-sound conditions when the participants were tested in silence, or else when they were listening to a semantically-irrelevant soundtrack (restaurant sounds in this study). These results therefore highlight a significant novel crossmodal modulation on the duration of perceptual dominance in the binocular rivalry situation, which resulted from the semantic content of the auditory soundtrack that the observers were listening to. This result can also be observed in the index of participants' first percept of the bird/car dichoptic figures. In the three conditions in which a soundtrack was presented, however, participants' alternation rates (indexed by their switch times) stayed at a level similar to that seen in the typical binocular rivalry situation (i.e., the no-sound condition). Finally, the modulation by auditory semantic context is more likely to shorten the dominance duration of the incongruent percept than to prolong that of the congruent percept. This result resembles to the modulatory effect of manipulating stimulus saliency in the binocular rivalry situation.

Presumably, one may further wonder whether the crossmodal modulation reported in Experiment 5.1 was mediated by either selective attention operating over a specific percept or increasing stimulus saliency, the two visual factors that have previously been shown to modulate perception in the binocular rivalry situation (Meng & Tong, 2004;

van Ee et al., 2005). Hence, in the next two experiments, the crossmodal semantic congruency effect on binocular rivalry was measured while simultaneously manipulating participants' selective attention and stimulus contrast, respectively.

5.3 Experiment 5.2

In Experiment 5.2, the target of participants' selective attention over the dichoptic figure was manipulated independently of the meaning of the sound. That is, the participants were instructed to maintain the bird percept, to maintain the car percept, or to view the figures passively in the control condition (see Meng & Tong, 2004; van Ee et al., 2005). Meanwhile, the participants either heard the birds singing or else the car engine revving soundtracks.

5.3.1 Methods

5.3.3.1 Participants

Seven participants who had shown an auditory modulation of binocular rivalry in Experiment 5.1 (the proportion of dominance durations in the bird- and car-sound conditions had differed by more than 5%) were recruited to take part in Experiment 5.2.

5.3.3.2 Design and procedure

Two factors, sound (bird and car) and selective attention (passive, maintain bird, or maintain car) were manipulated. The 5-second blank frame presented prior to the visual

stimulus display now contained the instruction to 'just look at the figures PASSIVELY', 'try to maintain the percept of the BIRD as long as possible' or 'try to maintain the percept of the CAR as long as possible'. The bird or car soundtrack started at the onset of the attention instruction frame (just as in Experiment 5.1). In order to increase the degree of overlap between the two figures and hence reduce the monocular cues associated with a given figure, the bird and car figures were rotated 15° counterclockwise and clockwise, respectively (see Figure 5.3).

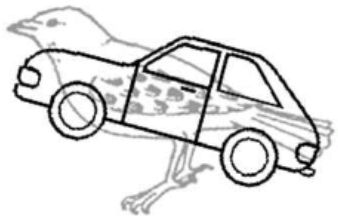


Figure 5,3. An example of the dichoptic stimulus pairs used in Experiments 5.2 to 5.4.

Three blocks of experimental trials were presented. The 12 trials (two sound conditions x three selective attention conditions x two color versions of visual pictures) were presented in a randomized order in each block. A practice block containing 6 no-sound trials, two for each selective attention condition, was conducted prior to the main experiment.

5.3.2 Results and discussion

A two-way ANOVA was conducted on the factors of sound and selective attention.

In the analysis of dominance duration (see Figure 5.4A), there were significant main effects of sound ($F(1,6) = 17.25$, $MSE = 0.004$, $p < .01$, $\eta^2 = 0.75$), and selective attention ($F(2,12) = 12.63$, $MSE = 0.04$, $p < .005$, $\eta^2 = 0.68$). A *post-hoc* Tukey test showed that there was a significant difference between the dominance duration in the *maintain bird* and *maintain car* conditions ($p < .01$). The interaction between sound and selective attention, however, was not significant ($F(2,12) = 1.39$, $MSE = 0.002$, $p = .29$, $\eta^2 = 0.18$). In order to examine whether the auditory modulation of binocular rivalry was significant in all three of the selective attention conditions, planned simple main effects on the sound factor were conducted. The results revealed that the dominance duration of the bird percept was longer when hearing the bird soundtrack than when hearing the car soundtrack in the maintain bird condition ($F(1,18) = 5.57$, $MSE = 0.002$, $p < .05$, $\eta^2 = 0.13$), in the passive condition ($F(1,18) = 16.65$, $p < .001$, $\eta^2 = 0.37$), and in the maintain car condition ($F(1,18) = 4.91$, $p < .05$, $\eta^2 = 0.11$). The magnitude of this auditory semantic modulation effect (indexed by subtracting the dominance duration in the bird-sound condition from that seen in the car-sound condition in each selective attention condition for each participant) was not significant ($F(2,12) = 1.40$, $MSE = 0.004$, $p = .29$, $\eta^2 = 0.19$).

The results of the first percept in each condition (see Figure 5.4B) revealed that the participants reported the bird as the first percept somewhat more frequently when they were either listening to the bird sound or when trying to maintain the bird view, as

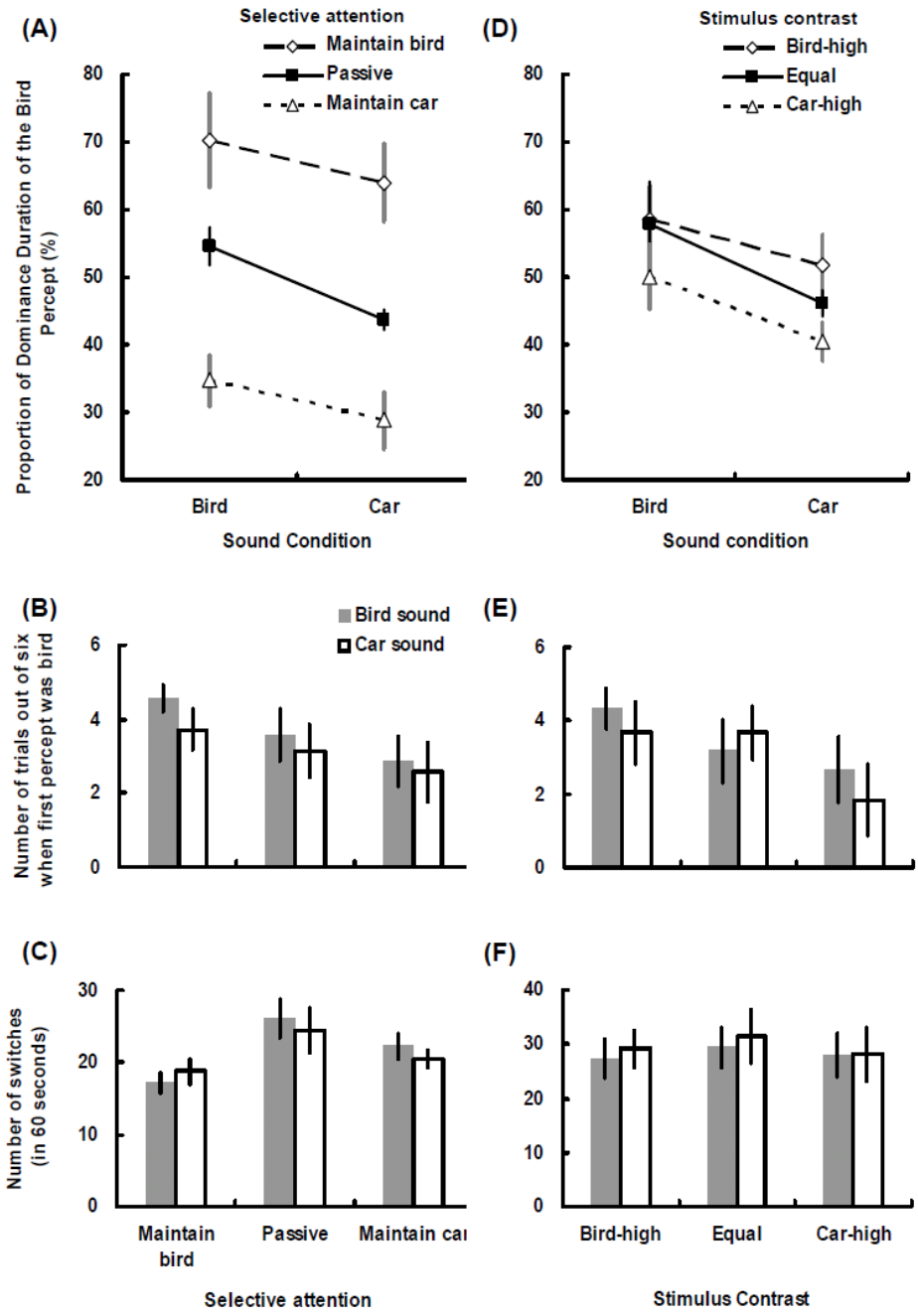


Figure 5.4. Results of the various sound conditions under conditions of binocular rivalry. The upper panels (A and D) depict the dominance duration of the bird percept (proportion to total dominance duration). The middle panels (B and E) show the number of times that participants reported the bird as the first percept (out of 6 trials). The lower panels (C and F) show the average number of switches taking place in each trial (during the course of 60 seconds) in Experiments 5.2 and 5.3, respectively. (A) The normalized dominance duration of the bird percept in the maintain view of bird (dashed line), passive viewing (solid line), and maintain view of car (dotted line) conditions of Experiment 5.2. (D) The normalized dominance duration of the bird percept in the bird-high (dashed line), equal (solid line), and car-high (dotted line) conditions in Experiment 5.3. Error bars represent ±1 standard error of the mean.

compared to when they were either listening to the car sound or else trying to maintain their view of the car. However, neither of the main effects, nor their interaction, reached significance (all Fs < 3.27, ps > 0.12, η^2s < 0.35).

Analysis of the number of switches in each condition (see Figure 5.4C) revealed both significant main effects of sound ($F(1,6) = 8.19$, $MSE = 0.43$, $p < .05$, $\eta^2 = 0.58$) and selective attention ($F(2,12) = 3.92$, $MSE = 48.05$, $p < .05$, $\eta^2 = 0.40$). A *post-hoc* Tukey test showed that the number of switches was higher in the passive condition than in the maintain bird condition ($p < .05$). The interaction between sound condition and selective attention was also significant ($F(2,12) = 5.27$, $MSE = 2.57$, $p < .05$, $\eta^2 = 0.47$). The simple main effect showed that the number of switches was *smaller* in the *bird-sound* condition than in the *car-sound* condition when the participants had to try and maintain their view of the bird ($F(1,18) = 5.39$, $MSE = 1.85$, $p < .05$, $\eta^2 = 0.16$). In contrast, the number of switches was *smaller* in the *car-sound* condition than in the *bird-sound* condition when the participants passively viewed the figures ($F(1,18) = 5.09$, $p < .05$, $\eta^2 = 0.15$), or when they had to try and maintain the view of the car ($F(1,18) = 6.02$, $p < .05$, $\eta^2 = 0.17$).

These results once again demonstrate that the participants reported perceiving the bird percept as being dominant for more of the time when they heard the sound of birds singing than when they heard the sound of cars revving their engines. In addition, selective attention also modulated the dominance duration of the bird percept when the

participants tried to maintain their view of the bird or car, consistent with the results reported by both Meng and Tong (2004) and van Ee et al. (2005). Critically, these two factors modulated dominance duration in a seemingly additive manner. That is, the modulation of auditory semantic context was observed in all of the passive, maintain bird, and maintain car conditions, and the magnitude of crossmodal modulatory effect was similar in these three conditions. The modulation by auditory semantic context therefore can be dissociated from that by visual selective attention in terms of dominance duration in the binocular rivalry situation. In other words, the robust auditory modulation observed in Experiment A.2 cannot be attributed solely to the meaning of the sound guiding participants' attentional selection, either voluntarily or involuntarily; otherwise, the effect of auditory semantic context should have been overridden by the effect of selective attention (see Hsiao, Chen, Spence, & Yeh, 2010).

It should, however, be noted that even though the modulation by auditory semantic context was not equivalent to that elicited by selective attention, the former likely increased the influence of the latter. That is, one possible explanation for the modulation by auditory semantic context is that the presentation of the sound may have enhanced the participants' selective attention when the meaning of the sound happened to be consistent with the target of selective attention, as suggested by van Ee et al. (2009). If so, the results should have shown the modulatory effect of auditory semantic context to be larger in the 'maintain bird' and 'maintain car' conditions than in the passive

condition (see the results reported by van Ee et al., 2009). This notion, however, is inconsistent with the dominance duration results that were actually observed here.

In terms of the number of switches, the results reveal that selective attention effectively reduced the frequency of switches during the test period (see also van Ee et al., 2005). This result is in line with the fact that the occurrence of switches from one percept to the other under conditions of binocular rivalry can be modulated by attention (see Leopold & Logothetis, 1999; Lumer, Friston, & Rees, 1998). This also provides evidence for the claim that the modulation of binocular rivalry by auditory semantic context was dissociated from any modulation of visual attention induced by the soundtrack, since there was no effect of presenting a soundtrack on switch times in Experiment 5.1. It should, however, be noted that the auditory semantic context present in Experiment 5.2 somehow assisted visual attentional control over perceptual switching. This was evidenced by the observation that when the participants were listening to the soundtrack that was congruent with the figure that they were instructed to try and maintain in awareness, the number of switches was smaller than when they heard the sound that was incongruent with it (see the opposite patterns modulated by the sound in the maintain bird and maintain car conditions). It is possible that, since attentional control over a given object representation relies on holding that target in working memory (see Desimone & Duncan, 1995), the presentation of a semantically-congruent auditory soundtrack may have helped the participants in Experiment 5.2 to hold the

target in mind during the test period.

In summary, the modulation by auditory semantic context and selective attention in the binocular rivalry paradigm can be dissociated in terms of determining the dominance duration; that is, these two factors appear to operate additively. Hence, attentional control over a specific percept in the binocular rivalry situation is not a necessary condition for the crossmodal modulation of binocular rivalry by auditory semantic context (cf. van Ee et al., 2009). Even so, auditory semantic context can undoubtedly modulate visual attentional control in terms of switch times. Further discussion of the interplay between auditory semantic context and attention is saved for the General Discussion of this chapter (see Section 5.6.1).

5.4 Experiment 5.3

The next question addressed in this chapter was whether the modulatory effect of auditory semantic context on binocular rivalry would interact with low-level visual factors. The luminance contrast of a figure provides a bottom-up (i.e., stimulus-driven) factor that has been shown to modulate perceptual dominance; that is, a higher luminance contrast figure will likely be perceived for more of the time than a figure with a lower luminance contrast (Meng & Tong, 2004; Mueller & Blake, 1989). The participants in Experiment 5.3 heard either the bird or car sound while presented with one of three levels of luminance contrast.

5.4.1 Methods

5.4.1.1 Participants

Six of the participants who took part in Experiment 5.2 (one dropped out) were tested.

5.4.1.2 Design and procedure

Two factors, sound condition (bird, car) and stimulus contrast (equal, bird-high, car-high) were manipulated. In the equal condition, the dichoptic figures used in the previous experiments were presented. In the bird-high condition, the luminance contrast of the bird figure was constant (the Michelson Contrast value was 67.1% for red and 7.4% for cyan), whereas the luminance contrast of the car figure was reduced (the Michelson Contrast value presented on monitor was 62.5% for red and 6.6% for cyan; that is, reduced by 6.9% for red and 10.8% for cyan, respectively), and vice versa in the car-high condition. In each trial, the attention instruction frame was now blank (just as in Experiment 5.1). The participants were instructed to view the figures passively. The other details were the same as Experiment 5.2.

Three blocks of experimental trials were presented. The 12 trials (two sound conditions x three stimulus contrast conditions x two color versions of visual pictures) in each block were presented in a randomized order. A practice block containing 6 no-sound trials, two for each stimulus contrast condition, was conducted before the

main experiment.

5.4.2 Results and discussion

A two-way ANOVA was conducted with the factors of sound and stimulus contrast. Analysis of the dominance duration data (see Figure 5.4D) revealed significant main effects of sound ($F(1,5) = 7.31$, $MSE = 0.01$, $p < .05$, $\eta^2 = 0.59$) and stimulus contrast ($F(2,10) = 11.76$, $MSE = 0.003$, $p < .005$, $\eta^2 = 0.71$). A *post-hoc* Tukey test showed that there was a significant difference between the bird-high and car-high conditions ($p < .01$), and between the equal and car-high conditions ($p < .05$). The interaction between sound and stimulus contrast was, however, not significant ($F(2,10) = 1.14$, $MSE = 0.002$, $p = .36$, $\eta^2 = 0.20$). The planned simple main effect of the sound factor revealed that the dominance duration of the bird percept was higher when hearing the bird soundtrack than when hearing the car soundtrack in the equal luminance condition ($F(1,15) = 8.82$, $MSE = 0.005$, $p < .01$, $\eta^2 = 0.27$), in the car-high condition ($F(1,15) = 5.97$, $p < .05$, $\eta^2 = 0.18$), and only marginally significant in the bird-high condition ($F(1,15) = 3.03$, $p = .10$, $\eta^2 = 0.09$)[20]. Note that the magnitude of this auditory modulation was not significantly different across the three levels of stimulus contrast ($F(2,10) = 1.12$, $MSE = 0.003$, $p = .37$, $\eta^2 = 0.18$).

Analysis of the first percept data (see Figure 5.4E) revealed a significant main

[20] It should be noted that, in the bird-high condition, all six participants consistently reported the bird percept for more of the time when listening to the bird-sound than when listening to the car-sound. A repeated measure *t-test* revealed this difference to be significant ($t(5) = 2.22$, $p < .05$, one-tailed).

effect of stimulus contrast ($F(2,10) = 7.18$, $MSE = 1.33$, $p < .05$, $\eta^2 = 0.59$). A *post-hoc* Tukey test showed that there was a significant difference between the bird-high and car-high conditions ($p < .05$). The other main effect, that of sound ($F(1,5) = 0.83$, $MSE = 1.20$, $p = .40$, $\eta^2 = 0.14$)[21], and the interaction between these factors ($F(2,10) = 2.02$, $MSE = 0.78$, $p = .18$, $\eta^2 = 0.29$) were not significant. No significant differences were observed in the analysis of the number of switches (see Figure 5.4F; all Fs < 4.01, ps > 0.05, $\eta^2 < 0.44$).

The results of Experiment 5.3 once again demonstrate that auditory semantic context can increase the dominance duration of a semantically-congruent percept under conditions of binocular rivalry. Besides, the finding of the dominance duration being influenced by the relative luminance contrast of the two visual inputs was replicated (Meng & Tong, 2004). Critically, the modulations of dominance duration by auditory semantic context and stimulus contrast were additive. This result also indicates that, even when the participants were listening to the bird (or car) soundtrack, they were still sensitive to the low-level visual properties (i.e., stimulus contrast in this experiment) of the dichoptic figures presented during the test period.

It should be noted that the auditory semantic context modulated the first percept in Experiment 5.1 but not in Experiment 5.3. On the other hand, the stimulus contrast

[21] The first percept data for the same six participants from Experiment 5.1 were examined. The results revealed that they reported the bird as the first percept more frequently when listening to the bird sound than when listening to the car sound ($F(1,5) = 8.45$, $MSE = 0.48$, $p < .05$, $\eta^2 = 0.63$). Hence, the finding that the modulatory effect of auditory semantic context on the first percept failed to reach significance in Experiment 5.3 cannot simply be attributed to a lack of statistical power.

determined the first percept in Experiment 5.3; that is, the figure that had the higher contrast was perceived first. The low-level visual factor of stimulus contrast was thus more salient in terms of determining the first percept than the semantic context of the auditory soundtrack.

Having shown that the modulation of dominance duration by auditory semantic context in the binocular rivalry situation was independent of the modulation by selective attention and stimulus contrast, a critical possible confounding factor involving response bias had to be excluded next. That is, it could be argued that the participants might have merely been adopting a strategy of reporting the percept that was semantically-congruent with the sound (though note that if they had done this, they should, in an extreme case, have reported seeing the semantically-congruent figure throughout the test period – which none of them did). In Experiment 5.4, the possibility that response bias might have affected participants' performance was reduced by presenting the soundtrack which was never directly congruent with participants' visual target.

5.5 Experiment 5.4

In Experiment 5.4, the participants only had to monitor the dominance duration of one of the two figures (such as the 'bird'), while they listened to either the soundtrack that was *incongruent* with the visual target (i.e., car soundtrack, in this case) or the

sound that was *irrelevant* to both figures (i.e., the restaurant soundtrack). A parallel task of monitoring the dominance duration of the car figure was also conducted, and the participants listened to the bird soundtrack in the incongruent condition while listening to the restaurant soundtrack in the irrelevant condition. This aspect of the experimental design effectively reduced the likelihood that participants would simply bias their responding in accordance with whichever soundtrack they happened to be hearing. On the other hand, if the auditory semantic context genuinely prolonged the dominance duration of the semantically-congruent picture (or shortened the dominance duration of the semantically-incongruent picture, see Experiment 5.1), the dominance duration of the visual target should be shorter in the *incongruent* condition than in the *irrelevant* condition.

5.5.1 Methods

5.5.1.1 Participants

Twelve volunteers (three males, mean age 26 years) took part in Experiment 5.4 in exchange for a £10 (UK Sterling) gift voucher. Three of the participants had taken part in the previous experiments, and the other 9 new participants were naïve as to the goal of the present study. They all had normal or corrected-to-normal vision and hearing.

5.5.1.2 Apparatus and stimuli

The visual test stimuli were the same as in Experiment 5.2 but somewhat smaller in

size (bird: 4.44°×2.76°; car: 4.41°×2.27°). The other details were the same as in Experiment 5.1.

5.5.1.3 Design and procedure

Two factors, the sound (incongruent or irrelevant) and visual target (bird or car), were manipulated. Each participant reported the dominance of the bird and car percepts in separate sessions. Under those conditions in which the visual target was the bird, the participants were instructed to press the '1' key as soon as the image of the bird became dominant. The participants were informed that the criterion of the 'bird being dominant' was that they were able to see every detail, such as the texture of the wings, of the figure of the bird. Once any part of the bird figure became vague or else started to be occupied by the features of the car figure, they were instructed to press the '0' key as soon as possible, which indicated the image of the bird was no longer completely dominant. Adopting this criterion was aimed to estimate the dominance duration of the bird percept more conservatively, since it excludes the duration when the car percept being dominant as well as when participants experienced a mixed percept. Similarly, under those conditions in which the visual target was the car, the participants had to press '1' and '0' to indicate the start and end time that they perceived the car percept as being dominant.

As in Experiment 5.1, a blank screen was presented for 5 seconds, followed by the

presentation of the visual stimuli for a further 60 seconds. The soundtrack was played from the onset of the blank screen until the offset of the test stimuli. The participants had to monitor the dominance of their target picture continuously during the test period, and report what sound they just heard after each trial.

In both visual tasks (i.e., when the visual target was a bird and when it was a car), a block of 12 trials was presented (consisting of two sound conditions x two color versions of visual pictures, each conditions were repeatedly tested 3 times). The order of presentation of these 12 trials was randomized. Prior to the completion of the experimental block of trials, a practice block containing 6 no-sound trials was presented. The order of conducting the two visual tasks was counterbalanced between participants.

5.5.2 Results and discussion

A two-way ANOVA was conducted with the factors of sound and visual target. Only two indices, the dominance duration of the target picture and the switch times (here the times that the participants reported the target figure as dominant was calculated), were used in the data analysis. In calculating the dominance duration, the participants may occasionally have pressed the ‘1’ or ‘0’ key twice. In such cases, the shorter duration (i.e., the duration from the second ‘1’ keypress to the first ‘0’ keypress) was used. The sum of the dominance duration for the target picture was divided by 60

seconds.[22] Analysis of the dominance duration data (see Figure 5.5A) revealed significant main effects of sound ($F(1,11) = 25.68$, $MSE = 0.0005$, $p < .0005$, $\eta^2 = 0.71$) and visual target ($F(1,11) = 11.50$, $MSE = 0.01$, $p < .01$, $\eta^2 = 0.51$). There was, however, no interaction between the sound and visual target factors ($F(1,11) = 0.60$, $MSE = 0.001$, $p = .46$, $\eta^2 = 0.07$). The planned simple main effect of the sound factor revealed that the dominance duration of the target picture was shorter when listening to the *incongruent* soundtrack than when listening to the *irrelevant* soundtrack both when the visual target was the bird ($F(1,22) = 6.34$, $MSE = 0.001$, $p < .05$, $\eta^2 = 0.15$), and when it was the car ($F(1,22) = 13.96$, $p < .005$, $\eta^2 = 0.33$). The magnitude of the auditory modulation effect was not significantly different in the bird and car target conditions ($F(1,11) = 0.60$, $MSE = 0.002$, $p = .45$, $\eta^2 = 0.06$). In the analysis of switch times (see Figure 5.5B), no main effects, nor their interaction, were significant (all Fs$(1,11) < 4.10$, ps $> .06$, η^2s < 0.27).

The results of Experiment 5.4 once again demonstrated the modulation of dominance duration by auditory semantic context in the binocular rivalry situation. Critically, the soundtrack that the participants listened to during the test period was never congruent with the visual target.[23] In other words, the crossmodal modulation of

[22] One participant stopped reporting the dominance of the target figure in the first 25 seconds in a trial in the incongruent condition, while keep reporting the dominance percept until the end in the other trials. This trial, as well as a matched condition in the irrelevant condition, was excluded from the data analysis.

[23] A control experiment in which, instead of hearing a soundtrack during the test period, the participants were given the name of the soundtrack in each trial, was conducted. Twelve volunteers (six males, mean age 24 years) took part in this experiment. One of them was the author, and the other 11 new participants were naïve as to the goal of the study. The experimental design was exactly the same as in Experiment 5.4, with the only exception as follows: A word was presented at the start of each trial, and the participants had to report the word after the test period. In the bird (or car) target condition, the word 'car (or bird)' was presented in the incongruent trials, and the word 'restaurant' was presented in the irrelevant trials. If

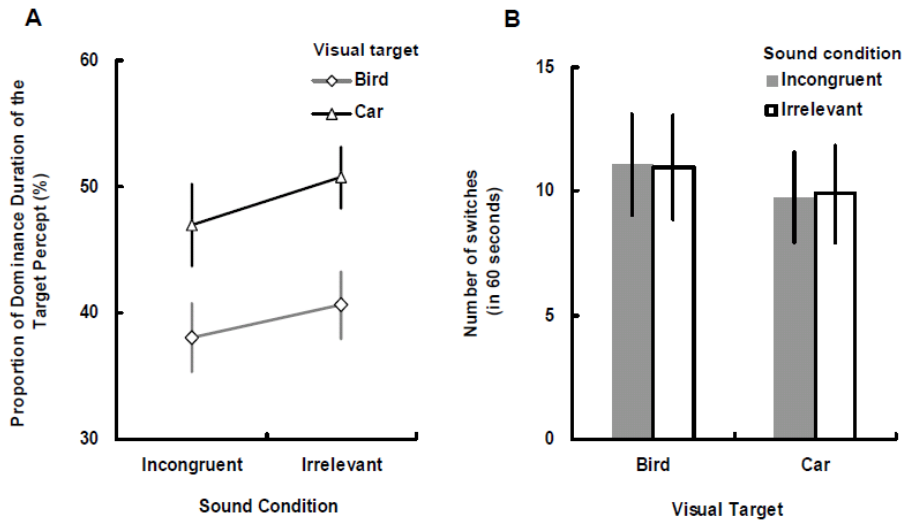


Figure 5.5. Results of the sound conditions (incongruent or irrelevant) under conditions of binocular rivalry in Experiment 5.4. (A) the dominance duration of the target percept (either bird or car, proportion of 60 seconds); (B) the number of times that the participants reported the target picture being dominant in each trial (during the 60-second test period).

binocular rivalry by auditory semantic context was indirect in terms of the incongruent sound either increasing the time to perceive non-target visual picture and/or decreasing the time to perceive the target picture (though the latter explanation is more likely, see Experiment 5.1). Both possibilities lead to a shortening of the dominance duration of the target picture. Therefore, the modulation of auditory semantic context was robustly observed under those conditions in which the meaning of the soundtrack could not

---

the participant reported the word incorrectly, the trial, as well as a trial with the same color picture in the other word condition, were excluded (4% of all trials, in order to control the eye dominance of each participant). A two-way ANOVA was conducted with the factors of word and visual target. In the analysis of dominance duration, the main effect of visual target was significant (36.6% vs. 46.4% for bird and car target conditions, respectively; $F(1,11) = 7.99$, $MSE = 0.01$, $p < .05$, $\eta^2 = 0.42$). However, neither the main effect of word (41.2% vs. 41.8% for incongruent and irrelevant conditions, respectively; $F(1,11) = 0.96$, $MSE = 0.0005$, $p = .35$, $\eta^2 = 0$), nor their interaction ($F(1,11) = 1.49$, $MSE = 0.001$, $p = .25$, $\eta^2 = 0.11$) were significant. In the analysis of switch times, neither main effects, nor their interaction, were significant (all Fs$(1,11) < 1$, ps $> .34$, η^2s < 0.08). Hence, simply maintaining a word in memory during test period did not bias a participant's dominance perception in the binocular rivalry situation. That said, the semantic modulation in the binocular rivalry situation reported in the experiments described in this chapter cannot simply be attributed to a semantic priming effect, but a continuous presentation of a meaningful soundtrack during the test period is critical.

directly bias the participants' response to report the semantically-congruent view as being dominant.

The participants reported the dominance duration when the car was the target was longer than when the bird was the target. This result may be attributable to the fact that the wing of the bird contains small individual elements which constitute the texture of the bird, and the individual element may disappear occasionally (Kovacs et al., 1996). On the other hand, most of the lines of the car figures are connected, so they should construct a unitary element (such as a car door). Since the participants were instructed to report the target figure as being dominant only when they could see all of its features, the better-grouped figure (i.e., the car) should be easier to reach this criterion than the worse-grouped figure (i.e., the bird).

One may wonder whether selective attention might have modulated the dominance duration in the present study, given that the participants only had to monitor one of the figures. Since the visual target was the same in both the incongruent and irrelevant sound conditions, participants' voluntary selective attention should always have been focused on the target stimulus. Furthermore, as shown in Experiment 5.2, the modulation of selective attention should effectively increase the dominance duration as well as reduce the switch times. Here, however, no significant difference in switch times was observed; that is, there was no evidence of the modulation of selective attention involved in the results of the present experiment.

5.6 General discussion

The results of the four experiments reported in the present chapter demonstrate that a participant's visual awareness in the binocular rivalry situation can be modulated crossmodally by the semantic context provided by a concurrently-presented auditory soundtrack. That is, the dominance duration of a given percept was longer when it happened to be semantically-congruent with the sound than when it was not (which was semantically-incongruent with the sound; see Experiments 5.1-5.3). In Experiment 5.2, the modulation by auditory semantic context was additive with the modulation resulting from a top-down factor, visual selective attention in terms of dominance duration. Similarly, in Experiment 5.3, the modulation by auditory semantic context was additive with the modulation resulting from a bottom-up factor, the visual luminance contrast in terms of the dominance duration. In Experiment 5.4, the modulation by auditory semantic context was robust even when the presentation of the sound could not directly bias participants' reporting of the dominance of the visual target.

One of the novel observations provided by the results reported here is that the semantic context of a task-irrelevant auditory soundtrack can effectively modulate participants' visual awareness. In contrast to the previous studies reported by van Ee et al. (2009) and Sheth and Shimojo (2004), in which the visibility of stimuli undergoing visual competition was maintained by a series of synchronized pure tones, the

experiments reported in this chapter demonstrated that the crossmodal modulation by sound was attributed to its semantic content (i.e., rather than to its physical synchrony). The results of the present study therefore provide important evidence that the factors modulating binocular rivalry can reach a semantic level (see Engel, 1956; though see Blake, 1988; Zimba & Blake, 1983), and critically, can occur crossmodally.

It should be noted that even though some studies have shown that a semantic prime can increase the likelihood that a semantically-congruent figure would break through to awareness prior to the other semantically-incongruent figure under conditions of dichoptic viewing (e.g., Costello, Jiang, Baartman, McGlennen, & He, 2009; Rommetveit, Toch, & Svendsen, 1968), the measurement of which picture was perceived first (i.e., the first percept) may merely reflect the result of dichoptic masking rather than true binocular rivalry (see Blake, 1988, p. 140; on this issue, see also Baker & Graf, 2009; van Boxtel, van Ee, & Erkelens, 2007). Furthermore, the modulation of auditory semantic context on the first percept reported by participants was only observed in Experiment 5.1, which may indicate that this effect was not particularly reliable (see below).

5.6.1 Possible mechanisms underlying the modulation of visual awareness by auditory semantic context

In Experiments 5.1 to 5.4, three possible factors (auditory semantic context, visual

selective attention, and visual stimulus contrast) that modulated participants' visual perception under conditions of binocular rivalry was examined (see Table 5.1 for a summary). All three factors effectively modulated the typical index of dominance duration of a given percept. Furthermore, in Experiments 5.2 and 5.3, the factor of auditory semantic context and either visual selective attention or visual stimulus contrast, modulated dominance duration additively. Since Meng and Tong (2004) have shown that visual selective attention and visual stimulus contrast work additively to modulate the dominant percept under conditions of binocular rivalry, it is therefore likely that all these three factors modulated the dominance duration in an additive manner.

Table 5.1. Summary of results in Experiments 5.1 to 5.4. A tick indicates that the modulating factor had a significant effect on the corresponding measurement.

		Measurements		
Experiment	Modulating factors	Dominance duration	First percept	Switch times
5.1	Auditory semantic context	√	√	
5.2	Auditory semantic context	√		√
	Visual selective attention	√		√
5.3	Auditory semantic context	√		
	Visual stimulus contrast	√	√	
5.4	Auditory semantic context	√	N/A	

Each of the three factors, however, may influence the performance in other aspects in the binocular rivalry situation. For example, only the factor of visual selective attention effectively reduced the switch times during the test period (see Experiment 5.2). To the extent that the reduction of switch times can be considered as the signature of selective attention in the binocular rivalry situation (see van Ee et al., 2005), any modulation of binocular rivalry by auditory semantic context should therefore be distinguished from that elicited by selective attention. On the other hand, the stimulus contrast effectively modulated the participants' first percept (Experiment 5.3). However, the auditory semantic context modulated the first percept when it was the only factor to be manipulated (Experiment 5.1), but not when it was manipulated with either one of the visual factors simultaneously (see the results of Experiments 5.2 and 5.3). That said, the modulation by auditory semantic context on the first percept was easily overridden by visual factors. Considering the results of all these three indices, the suggestion would be that the modulation by auditory semantic context can be dissociated from that by visual selective attention and visual stimulus contrast.

It should be noted that even though the modulation of binocular rivalry by auditory semantic context and visual selective attention can be dissociated at some level, they clearly interact, as revealed by the index of switch times (see Experiment 5.2). It should also be noted that, since the participants were asked to attend to the sound during the test period, it cannot be argued that the soundtrack was processed automatically here (cf.

Schneider, Dumais, & Shiffrin, 1984; see also van Ee et al., 2009). It could be suggested that the soundtrack should be selected in, rather than simply filtered out, in the early stages of auditory information processing (e.g., Treisman & Riley, 1969). It is possible that, once the sound had been processed, the auditory semantic context then unavoidably interact with any relevant visual information (see Lavie, 2005; Treisman & Davies, 1973; Yuval-Greenberg & Deouell, 2009).

The audiovisual semantic congruency effect reported here could be implemented in the model of binocular rivalry based on the three mechanisms (inhibitory, lateral excitatory, and feedback connections) proposed by Tong et al. (2006). A prerequisite of audiovisual interactions in terms of their semantic content is that both the meaning of visual and auditory stimuli have been accessed. That said, the possible stage of visual processing that was modulated by auditory semantic context should occur at, or later than, the representation of visual objects. Given that binocular rivalry can occur between competing visual pattern representations (Logothetis et al., 1996; Tong et al., 1998), the auditory semantic context may therefore either enhance the congruent, or inhibit the incongruent, visual object representation (e.g., Chen & Spence, 2010; Iordanescu et al., 2008). Either way, the visual stimulus that happened to be congruent with the sound would, as a result, win the visual competition more frequently. This crossmodal semantic congruency effect can be compared to the perceptual grouping effect (Alais & Blake, 1999; Alais, Lorenceau, Arrighi, & Cass, 2006; Kovacs et al.,

1996) or the contextual constraints (Shimojo & Nakayama, 1990; Treisman, 1962; Watson, Pearson, & Clifford, 2004) within the visual modality that occur under conditions of binocular rivalry.

According to this hypothesis, the activities of brain areas that are modulated by auditory semantic context should be at the locus of visual object recognition, such as inferior temporal cortex (IT) and superior temporal sulcus (STS, see Sheinberg & Logothetis, 1997; Tamura & Tanaka, 2001). On the other hand, previous studies suggest that audiovisual semantic interactions likely occur in brain areas such as STS, middle temporal gyrus, and perirhinal cortex (see Belardinelli et al., 2004; Taylor et al., 2006; see Doehrmann & Naumer, 2008, for a review). The area of STS is, though, reported to perform a dual function in terms of visual object recognition and audiovisual semantic interaction; however, auditory semantic context is likely to enhance the congruent visual object representation via either excitatory connections within multisensory areas and/or via feedback circuits from multisensory areas to visual areas, to modulate the visual pattern competition (see also *Account B for 'new bimodal brain areas'* and *Account C-F-C for 'critical role of feedback circuitry'*, respectively, in Driver & Noesselt, 2008). The critical difference between these two accounts lies in whether the competition between the neural activity corresponding to the two object representations occurs in multisensory brain areas (i.e., where the competing representations are multisensory, so the STS can be a candidate area), or whether instead it occurs in modality-specific

visual areas (say, for example, IT) that receive feedback from higher multisensory areas.

Recently, Pearson, Clifford, and Tong (2008) showed that imaging one of the percepts in a binocular rivalry situation can increase the possibility of that percept winning the competition to awareness. Mental imagery can be considered as a top-down modulation of a particular object representation, though the time requires executing mental imagery is much slower than executing selective attention (see Pearson et al., 2008, Experiment 3). Note that, however, Pearson et al. also showed that, when the background was 100% illuminated (i.e., a white background, as in the present study), their participants performed similarly under the conditions of passive viewing and executing mental imagery. Besides, Segal and Fusella's (1969, 1970) early studies showed that visual (or auditory) sensitivity to detect a visual (or auditory) stimulus was lowered when the observer executing mental imaging in either the visual or auditory modality, and the deterioration was larger when the stimulus and imaging were in the same modality than in different modalities. Recently, Daselaar, Porat, Huijbers, and Pennartz (2010) demonstrated that when imaging a visual or auditory object, activities in the primary visual and auditory cortices were weaker when perceiving a veridical visual or auditory object. They suggest that such suppression in primary sensory areas help to reduce the interruption from external sensory input when generating a mental image or sound. It is true that in most of experiments in Pearson et al.'s study, there was no visual stimulus presented during the imaging period. Therefore, it is unclear whether

mental imagery ability can be apply to enhance a particular percept in the binocular rivalry situation where the visual background was white and visual and auditory stimuli were continuously presented, as in the present study.

5.6.2 Human perceptual awareness: Unisensory or multisensory?

Hupé et al. (2008) recently demonstrated that the perceptual outcome of visual and auditory bistable stimuli which were presented together were generated separately within each sensory modality. This result may imply that the sites where conscious perception emerges are separate for different sensory modalities. This would be consistent with the traditional view that given that the processing of each sensory input is modular, the module itself should be the site responsible for the emergence of conscious perception in that sensory modality (e.g., Pylyshyn, 1999; Zeki, 2003).

The phenomenon of binocular rivalry provides an example of how a conscious experience emerges (Blake & Logothetis, 2002). Here, auditory semantic context modulated visual awareness under conditions of binocular rivalry, which is in line with the following views of the brain as a closely-connected network: In terms of neurophysiology, information is communicated between areas which were traditionally considered to sensory-specific, so brain regions are not structured as discrete unimodal modules (see Driver & Noesselt, 2008; Ghazanfar & Schroeder, 2006). In terms of psychological functioning, more generally, it is worth considering the powerful

constraints that semantics places on the perceptual system as it tries to infer the nature of the environmental stimulation (see Hohwy, Roepstorff, & Friston, 2008). That is, audiovisual semantic congruency can provide heuristics on multisensory integration that modulate what we experience on an everyday basis in the real world (i.e., see the literature on the *unity assumption*, Spence, 2007; Welch & Warren, 1980). The accumulating evidence regarding crossmodal interactions involving semantic congruency implies that the human semantic system does not consist of entirely independent modules for each modality (see McCarthy & Warrington, 1988). Instead, there may be an interconnected semantic network which still retains modality-specific information across the senses (e.g., Barsalou, 1999; Plaut, 2002; Shallice, 1988). Furthermore, the interplay between perceptual and higher-level cognitive systems may also imply that perception and cognition share common representation systems, as has been proposed by the view of grounded cognition (e.g., Barsalou, 1999; 2008).

5.7 Conclusions

The study reported in this chapter provides empirical evidence that auditory semantic context modulates visual perception in the binocular rivalry situation. The results demonstrate that the effect of auditory semantic context is dissociable from the previously reported effects of visual selective attention and visual stimulus contrast (Meng & Tong, 2004). When considering how the dominant percept in binocular rivalry

(and so, human visual awareness) emerges, information from other sensory modalities also needs to be considered; and, in turn, that multisensory stimulation provides a novel means other than unimodal stimulation to probe the contextual constrains on human visual awareness.

Chapter 6

Naturalistic sounds and spoken words modulate the categorization of visually-presented pictures at multiple levels of human information processing

6.1 Introduction

Spoken words (e.g., /dɔg/) and naturalistic sounds (e.g., a dog's 'woofing') constitute verbal and non-verbal stimuli that convey their meaning auditorily. They can be compared to printed words and pictures in the visual modality. Accordingly, either type of sound might be expected to connect to the concept of a common object, such as the above examples being associated with 'dog'. However, although both types of sound provide semantic context, they may interact with visual pictures in qualitatively different ways, due to the idiosyncratic nature of human information processing (see Chapter 7; Saygin, Dick, & Bates, 2005). In this chapter, the influence of naturalistic sounds and spoken words on participants' visual picture processing was investigated

using a picture categorization task. The SOA between the onset of the sounds and pictures was manipulated in order to better understand the timecourse of audiovisual semantic interactions (for a similar approach used to study visuotactile crossmodal interactions, see Shore, Barnes, & Spence, 2006).

Previously, crossmodal interactions between pictures and either naturalistic sounds or spoken words have been tested separately. For example, the presentation of naturalistic sounds has been shown to influence a participant's responses to visually-presented pictures in identification, categorization, and visual search tasks (see Chapter 8; Iordanescu et al., 2008; Schneider, Engel, & Debener, 2008; Yuval-Greenberg & Deouell, 2009). Crossmodal interactions between spoken words and pictures have often been tested using Stroop-like interference paradigms in which the observer has to name or categorize visually-presented pictures or color patches (Roelofs, 2005; Shimada, 1990; Stuart & Carrasco, 1993). Saygin et al.'s (2005) study was perhaps the first to compare these two types of sound at the same time. They demonstrated that the time required to select which of two pictures matched a naturalistic sound was shorter than to pick the picture that matched a spoken word. The aim of their study was, however, to investigate the processing of naturalistic sounds and spoken words, so the candidate pictures were presented 1000 ms before the onset of the sound in order to avoid any crossmodal interference resulting from the processing of the visually-presented pictures. Hence, when investigating the crossmodal modulation of

visual picture processing by these two types of sound, a narrower range of SOAs between the onsets of the pictures and sounds are needed.

Considering the timecourse of audiovisual semantic interactions, a fundamental issue concerns whether visual stimuli access their respective semantic representations more rapidly than do auditory stimuli. The meaning of a visually-presented picture can normally be accessed within about 100 ms of its presentation (Greene & Oliva, 2009; Potter, 1975). However, it typically takes longer to process sounds (e.g., Aramaki et al., 2010; Ballas, 1993). For example, the results to be reported in Chapter 7 demonstrate that a semantically-congruent sound needs to be presented about 350 ms before the onset of the picture in order to induce semantic priming effects that modulate a participant's sensitivity to detect or identify a visually-presented picture. That said, presumably when a naturalistic sound or a spoken word is presented with a sufficient lead (of perhaps 200-300 ms) such that it can be processed by the time that the target picture is presented, both of them can induce semantic congruency effects due to crossmodal priming (i.e., a facilitatory effect in the congruent condition and an inhibitory effect in the incongruent condition, see Roelofs, 2005; Schneider et al., 2008; Shimada, 1990). On the other hand, when the sound and the picture are presented close together in time, two possibilities emerge: If semantically-congruent visual and auditory stimuli are integrated due to their temporal coincidence, a semantic congruency effect should once again be observed (e.g., Iordanescu et al., 2008; Yuval-Greenberg &

Deouell, 2009). However, if semantic access to the meaning of the sound occurs later than to that of the picture, the meaning of the sound may only be expected to influence post-perceptual picture processing, such as that occurring at the decisional stage. In this case, any non-target information may interfere with the decisional processing for the target information (cf. Stuart & Carrasco, 1993).

The second factor that may modulate the time course of audiovisual semantic interactions is how the visual pictures, naturalistic sounds, and spoken words are processed. Researchers have argued that when processing a visual picture, its associated semantic representations are accessed first, followed by the associated lexical representations (i.e., the name of the picture; see Glaser & Glaser, 1989; Potter & Foulconer, 1975; Smith & Magee, 1980). The processing of naturalistic sounds may be comparable to that of visual pictures in that they are likely to access the associated semantic representations directly (e.g., Dick, Bussiere, & Saygin, 2002). By contrast, the processing of spoken words may be comparable to that of visually-presented words in that access to the associated meaning is mediated by their lexical representations (e.g., Coltheart, 2000; Roelofs, 2005). According to this notion, audiovisual semantic interactions may occur more automatically and more rapidly between naturalistic sounds and pictures than between spoken words and pictures (see also Chapter 7; Saygin et al., 2005).

The experiment reported in this chapter (Experiment 6.1) was designed to

investigate the influence of naturalistic sounds and spoken words on the processing of visual pictures by measuring participants' RTs in a picture categorization task. Seven SOAs between the sound and the picture were presented in order to probe the timecourse of audiovisual semantic interaction effects.

6.2 Methods

6.2.1 Participants

Fourteen volunteers (seven males, mean age of 25 years) took part in this study in exchange for course credit or a 10 pound (UK Sterling) gift voucher. All of the participants were native or bilingual English speakers. All of the participants had normal or corrected-to-normal vision and hearing by self-report, and all were naïve as to the purpose of the study.

6.2.2 Apparatus and stimuli

The visual stimuli were presented on a 15-inch CRT monitor (75 Hz refresh rate) controlled by a personal computer. The participants sat at a viewing distance of 58 cm from the monitor in a dimly-lit experimental chamber. Each trial consisted of the presentation of a target picture (see Figure 6.1), which was presented in black in the center of a white background. Thirty outline-drawings taken from the Snodgrass and Vanderwart (1980) and Bates et al. (2003) databases were used as visual targets (half were of animals and the rest of non-animals). All of the target pictures covered an area

of roughly 5°×5°.

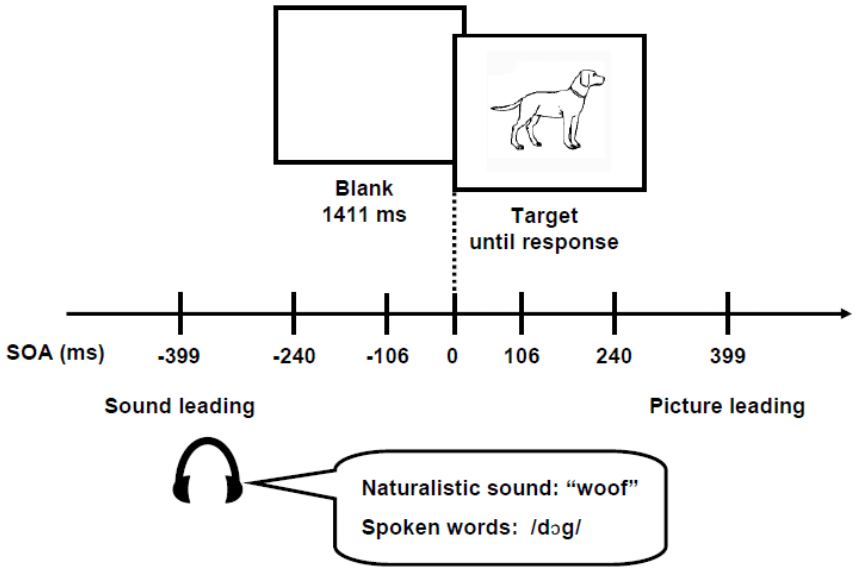


Figure 6.1. The sequence of two frames presented in each trial of Experiment 6.1 are shown: A blank screen followed by a target picture (e.g., dog). The timeline represents the seven possible SOAs between the onset of the visual target picture and the task-irrelevant accessory sound (either a naturalistic sound or a spoken word).

The congruent naturalistic sounds were chosen to match the pictures. For example, the sound of a dog's bark was matched with the picture of a dog; the sound of a creaking door was paired with the picture of a door (see Table 6.1 for further details). All of the sound files were downloaded from www.findsounds.com (on 08/08/2008). Only one sound file was used for each object. The sounds were edited to start from the beginning of the sound file and to last for 350 ms. The congruent words were chosen according to the most commonly agreed name used to refer to a given picture in the studies of Snodgrass and Vanderwart (1980) and Bates et al. (2003). All of the spoken

Table 6.1. Stimulus materials used in the experiments reported in Chapters 6 and 7

Picture	**Sound**	
	Congruent	**Incongruent**
Bird	Bird chirping	Motorcycle
Cat	'Meow'	Telephone
Cow	'Moo'	Door
Dog	'Woof'	Guitar
Duck	'Quack'	Cannon
Eagle	Eagle call	Lock
Elephant	Elephant trumpet	Piano
Fly	Buzzing fly	Hammer
Frog	'Ribbit'	Trumpet
Goat	'Baa-baa'	Whistle
Horse	Naying sound	Car
Mouse	Squeak	Gun
Pig	'Oink'	Drum
Rooster	Crowing sound	Scissors
Tiger	Roaring sound	Switch
Cannon	Cannon fire	Goat
Car	Car ignition sound	Cat
Door	Creak	Eagle
Drum	Drumstick hitting drum	Fly
Guitar	Sound of guitar	Mouse
Gun	Gun shot sound	Dog
Hammer	Sound of hammer hitting	Frog
Lock	Snap of lock	Tiger
Motorcycle	'Vroom-vroom'	Rooster
Piano	Sound of piano	Cow
Scissors	Clipping scissors	Duck
Switch	Clicking of switch	Elephant
Telephone	Telephone ringing	Pig
Trumpet	Sound of trumpet	Bird
Whistle	Whistling sound	Horse

words were recorded by a female native English speaker. Seventeen one-syllable words were compressed down to a total duration of 350 ms, 9 two-syllable words had a duration of 450 ms, while four words with three or more syllables had a duration of 500 ms. White noise lasting for 350 ms served as the meaningless auditory control stimulus.

The sounds were presented over closed-ear headphones, and were edited to have equalized peak amplitudes and ranged in loudness from 62 to 74 dB SPL.

6.2.3 Design

The three factors of semantic congruency (congruent, incongruent, or noise), SOA (-399, -240, -106, 0, 106, 240, and 399 ms), and sound type (naturalistic sound vs. spoken word) were manipulated. The semantic congruency factor was mixed within blocks. In the congruent trials, the sound matched the picture. In the incongruent trials, for fear of eliciting any unwanted semantically-related priming effects, the paired sound and picture belonged to different semantic categories; that is, a picture of an animal was always presented together with the sound of a non-animal item, such as a tool or instrument. The incongruent sound-picture pair was the same across different SOAs and sound types, so that any different effects could be simply attributed to the variation of these two factors rather than to any difference between the stimuli.

The SOA factor was blocked (see Glaser & Glaser, 1989; Roelofs, 2005). The order of presentation of the seven SOA blocks was determined randomly. In each SOA block, all 30 pictures were presented three times, accompanied by either a congruent or incongruent sound, or else by white noise. The presentation order of these 90 trials was randomized. Sound type was blocked, and the blocks containing naturalistic sounds or spoken words were grouped into the first or second experimental sessions. The order in

which the naturalistic sound and spoken word sessions (7 SOA blocks for each) were conducted was counterbalanced across participants.

6.2.4 Procedure

In the main experimental session, the participants initiated a block of trials by pressing the 'ENTER' key on the keyboard in front of them. In each trial, a blank frame was presented for 1411 ms, followed by a target picture which was presented until the participant responded. The participants had to decide whether they saw a picture of an animal by pressing the 'RIGHT ARROW' key or a picture of a non-animal by pressing the 'LEFT ARROW' key. Note that in each semantic congruency condition, participants' RTs and error rates were obtained from equal numbers of animal and non-animal trials. The participants were informed that they should respond to the picture rather than to the sound, and that they should respond as accurately, and rapidly, as possible.

Each participant first completed a no-sound block (i.e., the target picture was presented alone) containing all thirty target pictures used in the main experiment in order to familiarize him/herself with the task. Next, a practice session consisting of 12 trials (including the three semantic congruency conditions) was conducted prior to the main experiment. The SOA in this practice block was 0 ms and the sound type was the same as the participants encountered first during the experiment proper. The stimuli in

the practice session were not used in the main experiment. The experimenter made sure that the participants were following the instructions before proceeding to the main experiment. After a participant had completed the first session, the no-sound block was tested again in order to estimate the baseline RTs for the picture categorization task. The experiment lasted for approximately one hour.

6.3 Results

The RT and accuracy of participants' responses in each condition were calculated. Response latencies exceeding 1500 ms were excluded first (accounting for approximately 0.3% of all trials). The median RT (for correct response trials only) in each condition was used in order to reduce the influence of positive skewed distributions in the RT data when estimating participants' performance (Hays, 1973). See Table 6.2 for the average median RT in each condition.

The RT data were submitted to a three-way ANOVA with the factors of semantic congruency, SOA, and sound type. This revealed a significant main effect of semantic congruency ($F(2,26) = 16.47$, $MSE = 566.31$, $p < .0001$, $\eta^2 = 0.56$). A *post-hoc* Tukey test revealed that RTs were shorter in the congruent (458 ms) and noise (459 ms) conditions than in the incongruent condition (470 ms; ps $< .01$), while the former two conditions did not differ significantly. The main effect of SOA was also significant ($F(6,78) = 3.00$, $MSE = 1746.12$, $p < .05$, $\eta^2 = 0.19$). A *post-hoc* Tukey test revealed that

RT only reached significance between the -240 ms SOA (454 ms) and the 240 ms SOA (474 ms; $p < .05$). The only two-way interaction to reach significance was that between semantic congruency and SOA ($F(12,156) = 9.41$, $MSE = 403.15$, $p < .0001$, $\eta^2 = 0.42$). Critically, the three-way interaction was also significant ($F(12,156) = 2.55$, $MSE = 345.13$, $p < .005$, $\eta^2 = 0.16$). See Table 6.3 for the results of three-way simple main effects and *post-hoc* Tukey tests carried out on the semantic congruency factor at each SOA. The crossmodal priming effects calculated by subtracting the RTs in the congruent (or incongruent) condition from those in the noise condition are highlighted in Figure 6.2.

Table 6.2. Average of median RTs and standard errors (in parentheses) for each condition in Experiment 6.1.

		SOA (ms)						
		Sound leading						Picture leading
Condition		-399	-240	-106	0	106	240	399
Naturalistic sound	Congruent	435 (18)	441 (13)	453 (17)	452 (16)	477 (15)	469 (14)	463 (12)
	Noise	458 (22)	447 (12)	455 (16)	455 (17)	457 (16)	463 (12)	467 (20)
	Incongruent	467 (20)	465 (13)	466 (14)	457 (18)	475 (13)	468 (15)	457 (16)
Spoken word	Congruent	422 (10)	437 (9)	453 (11)	466 (13)	477 (9)	482 (17)	485 (17)
	Noise	464 (16)	458 (13)	451 (15)	448 (11)	463 (7)	485 (17)	458 (15)
	Incongruent	486 (16)	476 (18)	475 (14)	477 (15)	479 (10)	477 (18)	465 (16)

Note: The average median RT in the no-sound condition was 432 (11) ms.

Table 6.3. The results of the three-way simple main effects analysis on the semantic congruency factor and the post-hoc Tukey tests for each SOA (*: p < .05; **: p < .01) in Experiment 6.1.

Condition	SOA (ms)	Simple Main Effect			Tukey Test		
		$F(2,364)$	p	η^2	Congruent vs. Incongruent	Congruent vs. Noise	Incongruent vs. Noise
Naturalistic sound	-399	9.35	< .0005	0.03	**	**	
	-240	5.43	< .005	0.02	**		*
	106	4.22	< .05	0.02		*	*
Spoken word	-399	36.99	< .0001	0.13	**	**	**
	-240	13.09	< .0001	0.05	**	*	
	-106	6.31	< .005	0.02	**		**
	0	7.53	< .001	0.03		*	**
	399	7.22	< .001	0.03	*	**	

In summary, three components were observed in the naturalistic sound condition: A facilitatory crossmodal effect was elicited by the presentation of a semantically-congruent sound at the -399 ms SOA, an inhibitory effect was elicited by the presentation of a semantically-incongruent sound at the -240 ms SOA, and inhibition was elicited by both congruent and incongruent sounds at the 106 ms SOA. Three similar components, though at somewhat different SOAs, as well as a unique component were observed in the spoken word condition. A facilitatory crossmodal effect was elicited by the semantically-congruent words at the -399 and -240 ms SOAs, and an inhibitory effect was elicited by the semantically-incongruent words at the -399 ms SOA. An inhibition was also elicited by both congruent and incongruent spoken

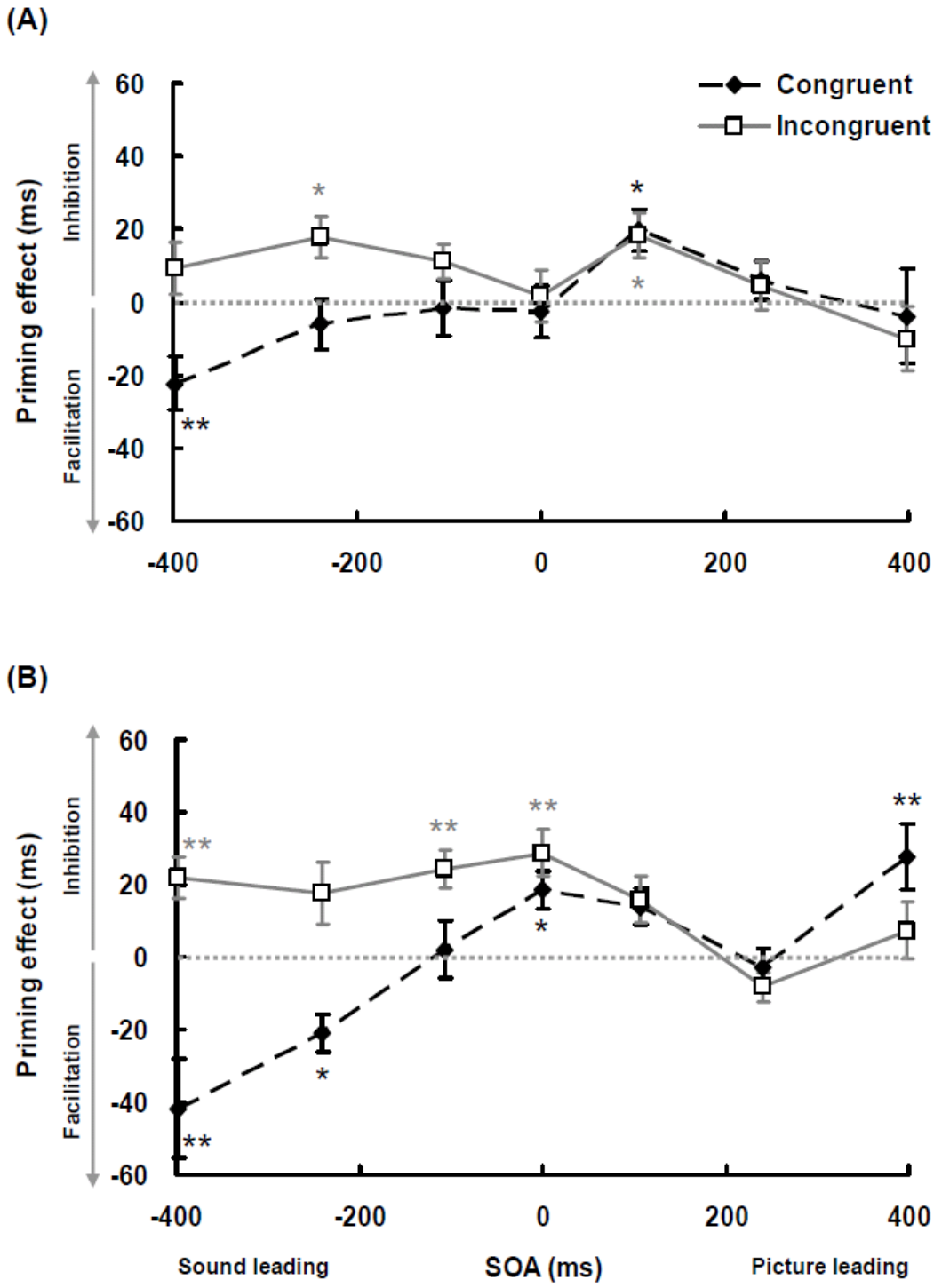


Figure 6.2. The priming effect on RTs in the congruent and incongruent sound conditions when plotted relative to performance in the noise condition for (A) naturalistic sounds and (B) spoken words for Experiment 6.1. 'Negative' priming effects indicate crossmodal facilitation whereas positive values indicate inhibition. Negative SOAs indicate that the auditory stimulus was presented first, while positive SOAs indicate that the target picture was presented first. Error bars indicate ±1 standard error of the means. * denotes a significant difference, black for the congruent condition and gray for the incongruent condition, respectively (*: $p < .05$; **: $p < .01$).

words at SOA of -106 and 0 ms (with a peak at 0 ms). The presentation of semantically-congruent words elicited an additional inhibitory effect at the 399 ms SOA.

The accuracy data (see Table 6.4) were submitted to a three-way ANOVA. The main effect of semantic congruency was significant ($F(2,26) = 6.06$, $MSE = 0.001$, $p < .01$, $\eta^2 = 0.32$). A *post-hoc* Tukey test showed that the accuracy of participants' responses was higher in the congruent (96.7%) and incongruent (96.8%) conditions than in the noise condition (95.9%; *p*s < .05). None of the other main effects or interactions was significant. Given that the accuracy of participants' responses was generally high and no trend toward a speed-accuracy trade-off was observed, the discussion will focus

Table 6.4. Mean accuracy (%), and the standard errors (in parentheses), of participants' responses for Experiment 6.1.

		SOA (ms)						
		Sound leading						Picture leading
Condition		-399	-240	-106	0	106	240	399
Naturalistic sound	Congruent	96.4 (1.3)	98.1 (0.8)	96.8 (1.0)	97.9 (0.9)	97.4 (0.9)	96.5 (1.3)	95.5 (1.6)
	Noise	96.8 (0.9)	95.2 (1.3)	94.5 (1.1)	96.5 (1.3)	95.2 (1.7)	94.7 (1.4)	96.4 (1.0)
	Incongruent	96.0 (1.0)	96.0 (1.5)	97.4 (1.2)	97.5 (0.9)	97.4 (0.8)	96.1 (0.9)	96.6 (0.8)
Spoken word	Congruent	97.9 (0.8)	96.1 (1.1)	95.4 (1.7)	96.7 (1.2)	95.7 (1.1)	96.6 (1.1)	96.2 (1.2)
	Noise	96.6 (1.0)	96.6 (0.8)	95.8 (0.8)	95.7 (1.0)	96.9 (0.8)	95.8 (1.4)	95.6 (1.0)
	Incongruent	96.3 (1.2)	97.3 (1.0)	97.1 (0.9)	98.1 (0.9)	98.1 (1.0)	95.7 (1.1)	95.7 (1.3)

Note: The mean accuracy in the no-sound condition was 97.0% (1.3%).

on the RT data.

6.3.1 Analysis of the one-syllable spoken words

Considering that the duration of the spoken word stimuli having two or more syllables was greater than 350 ms, RTs were calculated for items that contained sounds with a duration of 350 ms in order to provide a comparison to the naturalistic sound condition. There were 17 trials with one-syllable spoken words in the congruent and incongruent conditions, respectively, and all 30 trials in the noise condition. The data were submitted to a two-way ANOVA with the factors of semantic congruency and SOA. The main effect of semantic congruency ($F(2,26) = 13.76$, $MSE = 710.44$, $p < .0005$, $\eta^2 = 0.51$) was significant. A *post-hoc* Tukey test showed that the RTs were shorter in the congruent (457 ms) and noise (461 ms) conditions than in the incongruent condition (476 ms; both *p*s < .01), while performance in the former two conditions was not significantly different. The main effect of SOA was also significant ($F(6,78) = 3.47$, $MSE = 1565.75$, $p < .005$, $\eta^2 = 0.21$). A *post-hoc* Tukey test revealed that RT only reached significance for the comparison between the -399 ms SOA (451 ms) and the 240 ms SOA (484 ms; $p < .01$) and between the -106 ms SOA (456 ms) and the 240 ms SOA ($p < .05$). There was also a significant interaction between semantic congruency and SOA ($F(12,156) = 7.62$, $MSE = 613.89$, $p < .0001$, $\eta^2 = 0.37$). Table 6.5 highlights the results of simple main effects and *post-hoc* Tukey tests conducted on the semantic

Table 6.5. The results of two-way simple main effects analysis on semantic congruency and the post-hoc Tukey tests for each SOA in the analysis of one-syllable spoken words in Experiment 6.1 (*: p < .05; **: p < .01). Negative SOAs indicate that the auditory stimulus was presented first, while positive SOAs indicate that the target picture was presented first.

	Simple Main Effect			Tukey Test		
SOA (ms)	$F(2,182)$	p	η^2	Congruent vs. Incongruent	Congruent vs. Noise	Incongruent vs. Noise
-399	27.73	< .0001	0.18	**	**	
-240	17.30	< .0001	0.11	**	*	**
0	6.73	< .005	0.04	*		**
399	3.81	< .05	0.03		*	

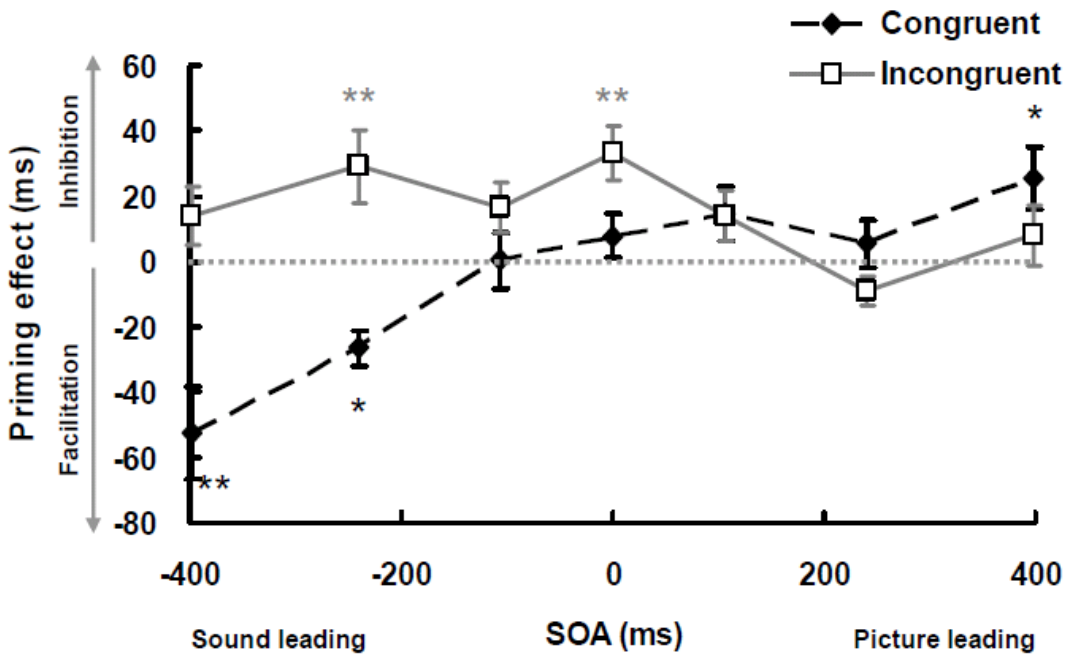


Figure 6.3. The priming effect on RTs in the congruent and incongruent conditions as compared to the noise condition in the analysis of one-syllable spoken words in Experiment 6.1. Negative priming effects indicate facilitation whereas positive values indicate inhibition. Negative SOAs indicate that the auditory stimulus was presented first, while positive SOAs indicate that the target picture was presented first instead. Error bars indicate ±1 standard error of the means. * denotes a significant difference, black for the congruent condition and gray for the incongruent condition, respectively (*: $p < .05$; **: $p < .01$).

congruency factor at each SOA. Figure 6.3 shows the priming effect calculated by subtracting the RTs in the congruent (or incongruent) condition from those in the noise condition.

6.4 Discussion

In Experiment 6.1, crossmodal semantic interactions between visually-presented pictures and both naturalistic sounds and spoken words were assessed. Participants generally took longer to respond when they heard a semantically-incongruent sound than when they heard either a semantically-congruent sound or white noise instead. There was also a trend for participants to respond more rapidly when the sound led the picture than when it followed the picture, which may be attributable to the alerting properties of sound (e.g., Robertson et al., 1998). Critically, the interactions between sounds and pictures differed as a function of the SOA. The crossmodal semantic congruency effects (i.e., the facilitation resulting from the presentation of semantically-congruent sounds and the inhibition resulting from the presentation of semantically-incongruent sounds, as compared to the noise condition) were observed when the onset of the sound preceded that of the picture by 240 ms or more. When the sound and picture were presented within about 106 ms of one another, all meaningful sounds (i.e., no matter whether they were congruent or incongruent) gave rise to inhibition. An unexpected effect was observed in the spoken word condition in that

when a semantically-congruent word was presented 399 ms after the onset of the picture, participants' RTs were slower as compared to both the incongruent and noise conditions.

6.4.1 Audiovisual semantic interactions at both perceptual and decisional stages

The semantic congruency effects observed at the -399 and -240 ms SOAs should presumably be attributed to the crossmodal semantic priming of vision by audition. That is, the sound had to be presented earlier than the picture in order to be processed, so that its meaning could facilitate or interfere with the processing of the visually-presented picture (see also Chapter 7). The crossmodal congruency effects observed when the onset of the sound preceded that of the picture therefore likely occurred at the level where the picture accesses its semantic representation.

On the other hand, both meaningful naturalistic sounds and spoken words elicited inhibitory effects when the sound and picture were presented within 106 ms of each other, irrespective of their semantic congruency. Similar effects between spoken words and pictures have been reported by Stuart and Carrasco (1993); specifically, they demonstrated that participants took longer to categorize a picture when the simultaneously-presented sound was a spoken word (either semantically-related or unrelated to the picture) as compared to noise instead. Note that the results reported in this chapter suggest that it should take around 200-300 ms to access the meaning of the sounds (see also Chapter 7), while at this point the meaning of the picture should

already have been accessed. This crossmodal inhibitory effect therefore plausibly occurs when the participant is about to make a decision (i.e., as to whether the picture is of an animal or not), and the presentation of any meaningful sound may thus interfere with this decisional process. Besides, even though the participants knew that they had to focus on the visual stimuli, they were unable to ignore the sounds completely. The meaningful sounds must therefore have been automatically processed up to a semantic level in order for them to interact with the processing of the visually-presented pictures. Hence, in the present experimental setting, the participants were capable of simultaneously-processing the visual and auditory inputs until a late stage (i.e., when the perceptual decision was about to be made). This result is consistent with the late selection view of attention, according to which, the bottleneck for attentional selection occurs at the decisional rather than at perceptual stage of information processing (see Deutsch & Deutsch, 1963; for a review see Wickens, 1984).

The results of Experiment 6.1 demonstrate that genuine semantic congruency effects occurred at SOAs of -399 and -240 ms. The intervals were presumably sufficient for participants to correctly perceive the sound and picture as being asynchronous (e.g., Stone et al., 2001; Zampini, Guest, Shore, & Spence, 2005). On the other hand, when the sound and the picture were presented within the temporal window of simultaneity (±106 ms here, see Vatakis & Spence, 2010), no advantage for the presentation of semantically-congruent sounds was observed when compared to control (i.e., noise) or

to semantically-incongruent sounds at the peak of inhibition. These results demonstrating audiovisual semantic congruency should therefore be distinguished from the 'unity assumption', which holds that when congruent stimuli from different sensory modalities are presented temporally close to each other, they are assumed to originate from the same source and consequently are more likely to be integrated (see Spence, 2007; Vatakis & Spence, 2007; Welch & Warren, 1980). Instead, the results reported here demonstrate that the audiovisual semantic congruency effect is determined by the time required for both the visual and auditory stimuli to access their semantic representations, rather than by their time of onset. That said, audiovisual integration for simultaneously-presented visual and auditory stimuli in terms of their semantic congruency may probably occur if the visual stimulus also takes time to process, such as when the visual stimuli are degraded (see Chapter 8; Noppeney et al., 2010; Yuval-Greenberg & Deouell, 2009) or time-evolving themselves (as in the case of videos, see Vatakis & Spence, 2007).

6.4.2 Differences between naturalistic sounds and spoken words

The crossmodal facilitation elicited by semantically-congruent naturalistic sounds (at the -399 ms SOA) had a smaller effect size and was shorter-lived than that resulting from the presentation of spoken words (at the -399 and -240 ms SOAs; see Table 6.3). This may be partly attributable to the lower familiarity or higher uncertainty that

participants had with regards to the naturalistic sounds than with the spoken words. Considering both the early and late inhibitory effects, however, a similar trend was observed in that the peaks of inhibition elicited by the spoken words (at the -399 and 0 ms SOAs, respectively) occurred earlier in time than those elicited by the naturalistic sounds (-240 and 106 ms SOAs, respectively). Note that when those trials containing one-syllable spoken words lasting only 350 ms were analyzed, the peak of the early inhibition effect shifted to an SOA of -240 ms while the late inhibitory effect remained centered on the 0 ms SOA. That said, the time required for a sound to be presented and identified should always be considered when investigating any effects induced by the meaning of an auditory stimulus. Nevertheless, the peak of late inhibition still occurred at an SOA that was earlier for spoken words than for naturalistic sounds. Note that the SOAs were probed more densely closer to synchrony than at other SOAs, and so the results with respect to the SOA manipulation should thus be more temporally precise in this region. The earlier timecourse of late inhibition for spoken words may be attributed to it taking longer for people to process spoken words than naturalistic sounds in order for their meaning to interfere with the visual decision. This hypothesis is consistent with the notion that pictures and naturalistic sounds access their semantic representations more rapidly while spoken words may access their meaning mediated by the lexicon (see Chapter 7; Cummings, Ceponiene, Koyama, Saygin, Townsend, & Dick, 2006; Saygin et al., 2005).

A unique inhibitory effect specific to the presentation of semantically-congruent spoken words was observed at the 399 ms SOA. Considering that the overall average RT for the spoken word condition was 466 ms, the onset of the word would have occurred at about the time when the participants were going to press a response key. It is likely that under such conditions, the participants heard a word that happened to be the name of the picture, and they may have tried to verify their answer, so delaying their response.

In sum, both naturalistic sounds and spoken words elicited similar early crossmodal semantic congruency effects and late inhibitory effects in a visual picture categorization task. It is suggested that the former effect occurs at the level of semantic access while the latter effect occurs at the level of decision making in visual picture processing. However, the timecourse for the two types of sounds were shifted somewhat. This result can be accounted for by the notion that naturalistic sounds and spoken words may access common semantic systems via different processing pathways. This issue will be further addressed in Chapter 7. Nevertheless, the results reported in this chapter suggest that audiovisual semantic interactions rely on the time for each stimulus to access its meaning. In conclusion, it would appear that audiovisual semantic interactions are heterogeneous phenomena, and that audiovisual integration in terms of semantic congruency may be a special case (see Vatakis et al., 2008).

Chapter 7

Enhancing visual sensitivity by means of crossmodal semantic priming

7.1 Introduction

Most objects and events in our daily lives give rise to information in several sensory modalities simultaneously. Accordingly, it would make sense for us to learn the associations between them in order to correctly represent the multisensory objects and events in the external world. So, for example, when walking on the pavement, a naturalistic 'woofing' sound should be linked to a dog rather than to the nearby dog-walker, while a 'vroom-vroom' sound should be linked with the motorcycle racing down the street instead. In addition to the association between these natural auditory and visual inputs, we have also learned that words, either in their spoken or written form, refer to specific objects. Naturalistic sounds and spoken words therefore provide two types of auditory stimulus that are associated with common objects. In the experiment reported in this chapter, the aim was to investigate whether hearing a

naturalistic sound or a spoken word would prime the visual perception of an associated picture. That is, crossmodal priming in terms of the semantic congruency between visual and auditory stimuli was tested.

The modulatory effect of a meaningful auditory stimulus on visual performance in humans has been demonstrated in various different experimental paradigms. For example, the prior presentation of a naturalistic sound can reduce the latency of a participant's responses when they have to categorize a semantically-congruent picture (Schneider, Engel, & Debener, 2008). When the auditory and visual stimuli are presented simultaneously, a naturalistic sound can help a participant to identify or localize a semantically-congruent picture (see Chapter 8; Iordanescu et al., 2008, 2010). Audiovisual interactions in terms of the semantic congruency between naturalistic sounds (or spoken words) and pictures (or written words) have also been investigated, typically using either Stroop-like interference paradigms (Roelofs, 2005; Shimada, 1990; Stuart & Carrasco, 1993; Yuval-Greenberg & Deouell, 2009) or the redundant signal paradigm (Laurienti et al., 2004; Molholm et al., 2004; Suied et al., 2009). Another popular approach has been to utilize temporal order judgments (TOJs) in order to test whether the temporal window of audiovisual integration is modulated by the semantic congruency between stimuli (Mitterer & Jesse, 2010; van Wassenhove et al., 2007; Vatakis et al., 2008; Vatakis & Spence, 2007, 2008).

The first issue with regards to the above evidence is to consider how the visual

and auditory information interact as a function of their stimulus onset asynchrony (SOA). When a sound is presented either before, simultaneously with, or after the visual target, it should presumably influence visual picture processing at different stages (see Chapters 6 & 8). Examining the effect of the SOA between visual and auditory stimuli is even more complicated because the latencies needed to access the semantic representations of visual and auditory stimuli are different: The meaning of a picture can be accessed rapidly (within about 100 ms, e.g., Potter, 1975), while a sound typically takes much longer to be processed (see Ballas, 1993). This is partly due to the fact that all of the visual features in the picture are presented at the same time, whereas auditory signals typically evolve over time. Hence, in order for a sound to induce a crossmodal semantic priming effect on the processing of a visual stimulus, sufficient time would seem to be a necessary prerequisite in order to process the sound prior to the onset of visual target (e.g., the SOA was 1400 ms in Schneider, Engel, & Debener, 2008). Accordingly, when the picture and the sound are presented simultaneously, one may wonder whether any crossmodal semantic interactions occurred at the level of visual processing (e.g., by enhancing visual feature processing or semantic access; Iordanescu et al., 2008; Stuart & Carrasco, 1993), or else at a post-perceptual processing stage such as consolidation in memory (see Chapter 8), or even at the level of response execution (e.g., Laurienti et al., 2004; Roelofs, 2005; Suied et al., 2009).

The second issue concerns the way in which an auditory stimulus provides

information to a participant about a visual stimulus. The studies reviewed above have only tested audiovisual semantic interactions using either naturalistic sounds (Iordanescu et al., 2008; Molholm et al., 2004; Schneider, Engel, & Debener, 2008; Suied et al., 2009; Yuval-Greenberg & Deouell, 2009) or spoken words (Laurienti et al., 2004; Roelofs, 2005; Shimada, 1990; Stuart & Carrasco, 1993). Given that both naturalistic sounds (e.g., the 'woofing' of a barking dog) and spoken words (e.g., /dɔg/) connect to a common object or event (such as 'dog' in this example), these two types of sound should be expected to prime the associated visual picture similarly. Evidence from neuropsychological studies has demonstrated that any impairment in the understanding of naturalistic sounds and spoken words typically occurs together in aphasic patients, indicating that the processing of these two types of auditory stimuli share a common underlying neural network (e.g., Dick et al., 2007; Saygin, Dick, Wilson, Dronkers, & Bates, 2003). However, certain evidence has also demonstrated that naturalistic sounds and spoken words may be dissociated at some level of information processing. For example, naturalistic sounds and spoken words can prime sounds of the same type but not of different types in behavioral studies (e.g., Stuart & Jones, 1995, 1996). In addition, the priming effects revealed by ERPs from naturalistic sound to spoken word and vice versa are unequal (e.g., Friedman, Cycowicz, & Dziobek, 2003; van Petten & Rheinfelder, 1995). Hence, it is possible that semantically-associated naturalistic sounds and spoken words access the same

semantic representations via different processing pathways (see Cummings et al., 2006; Noppeney, Josephs, Hocking, Price, & Friston, 2008; Thierry & Price, 2006). It is therefore worth testing whether the naturalistic sound and spoken word would induce similar or different semantic priming effects on the visual detection of a picture at the same SOA.

The results of previous research suggest that different semantic priming effects on a visual picture detection task elicited by naturalistic sounds and spoken words are likely to be observed. For example, people can match a picture and a naturalistic sound in terms of their meaning more rapidly than they can match a picture and a spoken word (Saygin et al., 2005). Here the hypothesis is that the processing of naturalistic sounds versus spoken words is comparable to the processing of pictures versus written words in the visual modality: That is, naturalistic sounds and pictures may access semantic representations directly, whereas spoken and written words may access the corresponding lexical representation first (for vision, see Glaser & Glaser, 1989; Potter & Foulconer, 1975; Smith & Magee, 1980; for audition, see Cumming et al., 2006; Dick et al., 2002; for the case of word processing, see Coltheart, 2000). According to this notion, it is likely that the interaction between naturalistic sounds and pictures occurs automatically (see Werner & Noppeney, 2010b), whereas the influence from spoken words to pictures occurs only when the name of the picture is required to respond (see the models in Glaser & Glaser, 1989). Supporting evidence comes from

the results of a study by Stuart and Carrasco (1993) in which a spoken word was shown to influence participants' performance in a picture naming task but not in a picture categorization task. Hence, a task will be added that requires participants to access the name of the picture if there is no semantic priming effect on visual picture detection elicited by spoken words observed.

In the experiments reported in the present chapter, the goal was to test whether the presentation of a semantically congruent (or incongruent) sound would modulate early visual picture processing by utilizing the semantic priming paradigm (Neely, 1977). Here, the participants had to detect the presence of any picture, and their visual sensitivity and response criterion based on the hit rate in picture present condition and false alarm (FA) rate in the picture absent condition were estimated according to signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005). From Experiments 7.1 to 7.3A, the SOA for a naturalistic sound needed to enhance participants' sensitivity to the presence (vs. absence) of visually-presented pictures was assessed. Experiments 7.3 to 7.4 were designed to test whether naturalistic sounds and spoken words give rise to similar semantic priming effects on visual picture detection at the same SOA. An additional identification task was also included in Experiment 7.4 in order to boost the audiovisual semantic interactions.

7.2 Experiment 7.1

Experiment 7.1 was designed to investigate whether a participant's ability to detect the presence of a briefly-presented line-drawing picture would be modulated by the semantic congruency of a naturalistic sound (i.e., congruent or incongruent) presented before the onset of the picture. The prediction was that the presentation of a semantically congruent sound would facilitate, while the presentation of a semantically incongruent sound might impair, participants' performance when both were assessed relative to a control condition in which the picture was presented by itself.

7.2.1 Methods

7.2.1.1 Participants

Seventeen volunteers (seven males, mean age of 22 years) took part in this study in exchange for course credit or a five pound (UK Sterling) gift voucher. All of the participants had normal or corrected-to-normal vision and hearing by self-report, and all were naïve as to the purpose of the study.

7.2.1.2 Apparatus and stimuli

The visual stimuli were presented on a 15-inch CRT monitor (75 Hz refresh rate) controlled by a personal computer. The participants sat at a viewing distance of 58 cm from the monitor in a dimly-lit experimental chamber. Each trial consisted of the presentation of a target picture and a pattern mask (see Figure 7.1A). All of the stimuli

were presented in black in the center of a white background. Thirty outline-drawings (half of animals and the rest of non-animals) taken from the Snodgrass and Vanderwart (1980) and Bates et al. (2003) databases were used as visual targets. The mask consisted of random curves and covered an area of 5.32°×5.32°, sufficient to occlude all of the target pictures completely.

The congruent auditory stimuli were chosen so as to match the pictures. For example, the sound of a dog's bark was matched with the picture of a dog; the sound of a creaking door was paired with the picture of a door (The stimulus materials were exactly the same as those used in the experiments reported in Chapter 6, see Table 6.1). The sounds were edited to start from the beginning of the sound file and to last for 350 ms with equalized peak amplitudes. The sounds were presented over closed-ear headphones and ranged in loudness from 64-74 dB SPL.

7.2.1.3 Design

Three factors, semantic congruency (congruent vs. incongruent), sound (present vs. absent), and picture (present vs. absent) were manipulated. The semantic congruency factor was presented in a blocked manner. In the congruent block, the sound matched the picture. In the incongruent block, for fear of eliciting any unwanted semantically-related priming effects, the sound and picture paired in a trial belonged to different semantic categories. That is, a picture of an animal was always accompanied

by a non-animal sound, such as a tool or instrument.

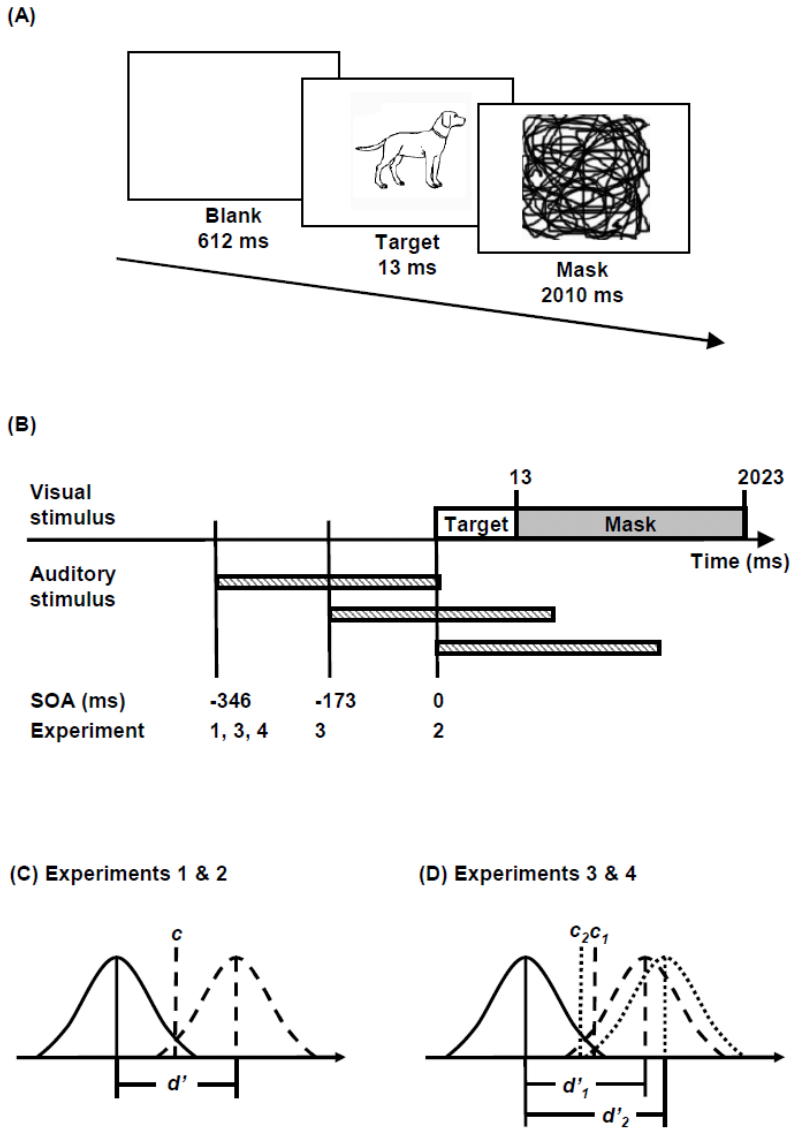


Figure 7.1. (A) The sequence of three frames presented in each trial: a blank, a target picture (e.g., dog), and a pattern mask. (B) A schematic representation of the manipulation of SOAs between the visual target picture and the task-irrelevant accessory sound across the four experiments reported in this chapter. Most of the sounds lasted for 350 ms. (C) represents the design in terms of signal detection theory in Experiments 7.1 and 7.2. The distributions of dashed line and solid line represent picture present and picture absent conditions, respectively. (D) represents the design in terms of unidirectional signal detection theory in Experiments 7.3 and 7.4. The distributions highlighted by the dotted line, the dashed line and the solid line represent the picture present/congruent sound, picture present/incongruent sound, and picture absent/sound present conditions, respectively.

In each block, all 30 pictures were presented twice, once with sound and once without. An additional 60 picture absent trials consisting of a blank frame were also presented, half with sound and the rest in silence, in order to estimate the participants' FA rate. These 120 trials were presented in a random order. Each congruent and incongruent block was tested twice, and the order of presentation of these four blocks was counterbalanced and intermixed across participants. Note that the participants were not given any information concerning the possible semantic congruency between the sound and picture prior to taking part in the study.

7.2.1.4 Procedure

Prior to the start of the experiment, a preview session including all of the pictures as well as their matched names was presented. Each picture-name pair was presented for 1012 ms. The next picture-name pair was then presented after a further 506 ms blank interval. The order of presentation of each picture-name pair was completely randomized across participants.

In the main experimental session, the participants initiated a block of trials by pressing the 'ENTER' key on the keyboard placed in front of them. In each trial, a blank frame was presented for 612 ms, followed by either a frame with a picture or another blank frame for 13 ms (1 frame at the screen refresh rate of 75 Hz). The pattern mask, used to control the presentation duration of the visual target and the difficulty of

the visual task, was presented immediately afterward. The participants had to decide whether they had seen a picture (irrespective of its identity) presented before the mask by pressing the 'SPACE' bar during the presentation of the mask (2010 ms). They were informed that they should respond to the picture rather than to the sound, and that they should try and hold a strict response criterion throughout the course of the experiment. The onset of the sound, if present, preceded that of the picture by 346 ms (i.e., the -346 ms SOA, see Figure 7.1B). The experimenter made the following unexpected inquiry after each participant had finished the experiment: *'Did you notice that in two of the blocks, the sound and the picture were semantically matched, while in the other two blocks they were not?'*

Two practice sessions consisting of 8 trials (with a picture duration of 27 ms) and 16 trials (with a picture duration of 13 ms) including all 8 conditions (created by crossing the semantic congruency, sound, and picture factors) were conducted prior to the main experiment. The stimuli in the practice session were not used in the main experiment. The experimenter made sure that the participants were following the instructions before proceeding onto the main experiment. The experiment lasted for approximately 30 minutes.

7.2.2 Results and discussion

The accuracy of participants' responding in the picture present conditions

constituted their hit rate, while the error rate in the picture absent conditions constituted their FA rate. The hit and FA rates were then transformed to *z* scores. The detectability (*d′*) and response criterion (*c* is used in the present study) for the sound present and absent conditions were calculated using the following equations (Green & Swets, 1966; Macmillan & Creelman, 2005; see Figure 7.1C):

$$d' = z(\text{Hit rate}) - z(\text{FA rate})$$

$$c = -0.5*[z(\text{Hit rate}) + z(\text{FA rate})]$$

Note that if a participant always responded to the presence of the target (i.e., if they obtained a hit rate of 100%) or always rejected the target when it was absent (0% FA rate), their value was adjusted.[24] One participant responded more often in the picture-absent trials than in the picture-present trials (i.e., his/her FA rate was higher than his/her hit rate), indicating that he/she had adopted a most unusual strategy in responding to the presence of the pictures, and therefore his/her data were excluded from further analysis.

The *d′* data (see Figure 7.2A) were submitted to a two-way ANOVA with the factors of semantic congruency (congruent vs. incongruent) and sound (present vs. absent). Neither main effect was significant (Fs < 1.26, ps > .28, η^2s < 0.08); however, their interaction was significant ($F(1,15) = 15.10$, $MSE = 0.03$, $p < .005$, $\eta^2 = 0.50$).

[24] Whenever a participant responded without error in the picture present condition, his/her hit rate was adjusted as 1-(1/2N), where N is the number of trials in that condition. Similarly, when a participant always rejected that the picture was presented in the picture absent condition, the FA rate was adjusted as 1/2N (see Macmillan & Creelman, 2005, p. 8).

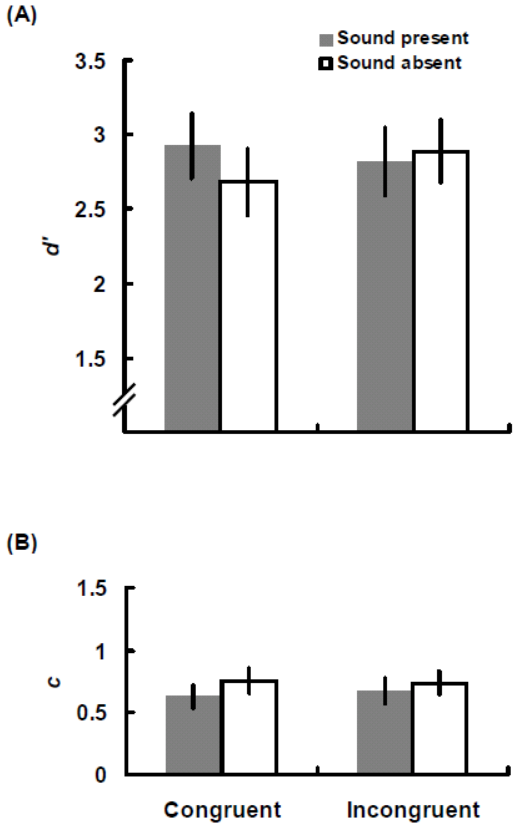


Figure 7.2. The results of (A) sensitivity (*d'*), and (B) response criterion (*c*), in Experiment 7.1. The error bars indicate ±1 standard error of the means.

The simple main effect showed that *d'* was higher in the sound present condition than in the sound absent condition when the sound was semantically congruent with the picture ($F(1,30) = 7.25$, $MSE = 0.07$, $p < .05$, $\eta^2 = 0.19$), while no such difference between the sound present and sound absent conditions was observed when the sound was incongruent with the picture ($F(1,30) = 0.45$, $p = .51$, $\eta^2 = 0.01$). The *c* data (see Figure 7.2B) were submitted to a two-way ANOVA on the factors of semantic congruency and sound. Neither of the main effects, nor their interaction, were significant (all Fs < 2.62, ps > .14, η^2s < 0.15). The results of the final inquiry revealed that 14 out of the 16 participants reported having noticed the semantic congruency

between the sound and the picture ($X^2(1) = 9.00$, $p = .003$).

The results of Experiment 7.1 demonstrated that the presentation of a semantically-congruent sound leading the onset of the masked visual target by 346 ms facilitated, while a semantically-incongruent sound (leading by the same amount) failed to impair, the detectability of the masked picture relative to a baseline no-sound condition. These results therefore indicate that the crossmodal facilitatory effects resulting from the congruent sound and the crossmodal interference effects resulting from the incongruent sound were asymmetrical in magnitude at the -346 ms SOA. It should also be noted that even though most of the participants noticed the different kinds of semantic congruency that were present in the different block types, they did not adopt different criteria in the congruent versus incongruent blocks. In the next experiment, the naturalistic sound and the picture were presented simultaneously in order to test whether a crossmodal facilitation of picture detectability induced by the presentation of a semantically-congruent sound would still be observed.

7.3 Experiment 7.2

In Experiment 7.2, the onset of the sound, if present, was simultaneous with that of the picture. In Experiment 7.2A, the sound consisted of a naturalistic sound that was either semantically congruent or else semantically incongruent with the picture (just as in Experiment 7.1). In Experiment 7.2B, the sound was either semantically congruent

with the picture or else consisted of a burst of white noise. The inclusion of a white noise enabled us to compare how a simple auditory transient (without any meaning) can modulate the participants' visual picture detection performance.

7.3.1 Experiment 7.2A

7.3.1.1 Methods

A new group of 10 volunteers (six males, mean age of 23 years) from the same pool of participants as in Experiment 7.1 took part in this experiment. Three factors, semantic congruency (congruent vs. incongruent), sound (present vs. absent), and picture (present vs. absent) were manipulated. The sound and picture in this experiment onset together. Consequently, the sound ended 337 ms after the offset of the picture. All of the other details were the same as in Experiment 7.1.

7.3.1.2 Results

The d' and c data (see Figures 7.3A & 7.3B) were submitted to a two-way ANOVA on the factors of semantic congruency and sound. In the analysis of d', neither the main effects, nor their interaction, were significant (all Fs < 2.09, ps > .19, η^2s < 0.19). In the analysis of c, there was a main effect of sound ($F(1,9) = 6.96$, $MSE = 0.17$, $p < .05$, $\eta^2 = 0.44$), indicating that the participants adopted a lower criterion (i.e., they exhibited a tendency to say 'Yes') in the sound present condition as compared to the sound absent condition. The other main effect of semantic congruency and the

interaction term were not significant (Fs < 2.25, ps > .17, η^2s < 0.20). The results of the unexpected inquiry showed that only four out of 10 participants noticed the semantic congruency between the sound and the picture in this experiment ($X^2(1) = 0.40$, $p = .53$).

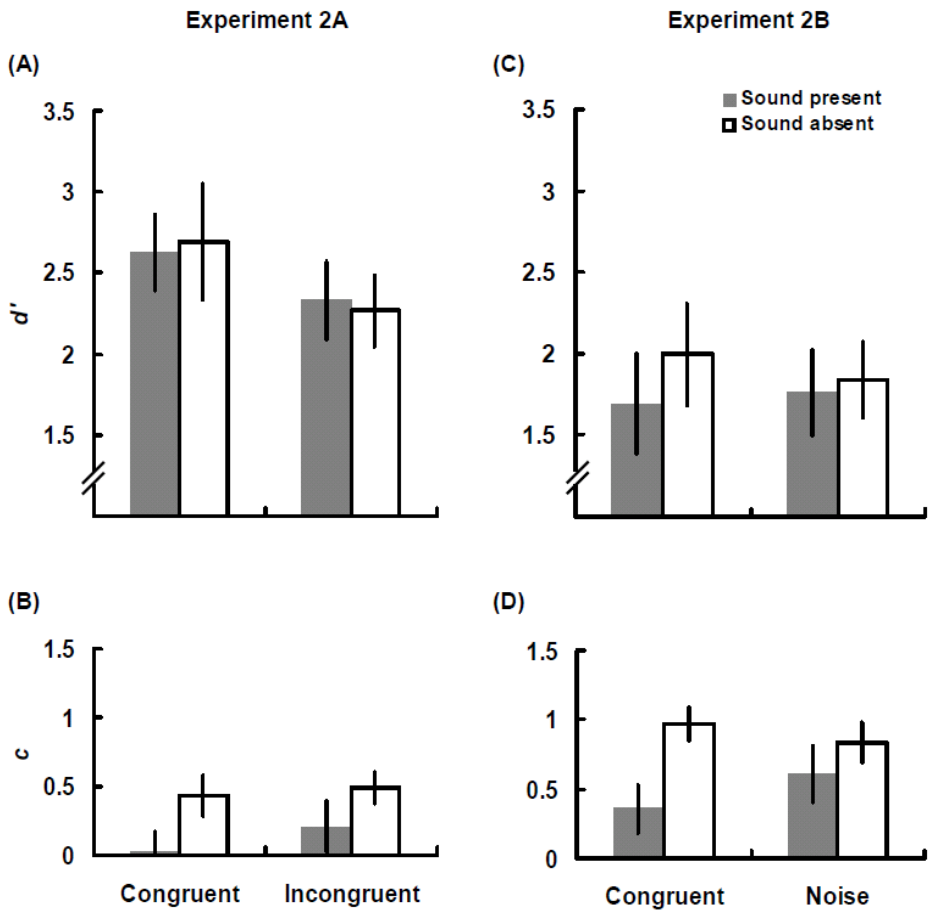


Figure 7.3. The results of sensitivity (d', in panels A and C), and response criterion (c, in panels B and D), in Experiment 7.2A and 7.2B, respectively. The error bars indicate ±1 standard error of the means.

7.3.2 Experiment 7.2B

7.3.2.1 Methods

Another group of 10 volunteers (two males, mean age of 19 years) from the same pool of participants as in Experiment 7.1 took part in this experiment. Three factors, semantic congruency (congruent vs. noise), sound (present vs. absent), and picture (present vs. absent) were manipulated. The white noise was presented at 72 dB SPL for 350 ms. The participants were informed about the sound condition before each block of trials, so there was no post-inquiry concerning whether or not they had noticed the semantic congruency between the visual and auditory stimuli. The other details were exactly the same as in Experiment 7.2A.

7.3.2.2 Results

The *d′* and *c* data (see Figures 7.3C & 7.3D) were submitted to a two-way ANOVA on the factors of semantic congruency and sound. In the analysis of *d′*, neither the main effects, nor their interaction, were significant (all Fs < 2.07, ps > .19, η^2s < 0.19). In the analysis of *c*, the participants once again adopted a lower criterion in the sound present condition than in the sound absent condition ($F(1,9) = 17.65$, $MSE = 0.10$, $p < .005$, $\eta^2 = 0.66$). The interaction between semantic congruency and sound was also significant ($F(1,9) = 18.31$, $MSE = 0.02$, $p < .005$, $\eta^2 = 0.70$). The simple main effect revealed that the presentation of the sound induced a lower criterion in

both the congruent ($F(1,18) = 31.38$, $MSE = 0.06$, $p < .0001$, $\eta^2 = 0.59$) and noise conditions ($F(1,18) = 4.14$, $p = .056$, $\eta^2 = 0.08$); critically, the presentation of the semantically-congruent naturalistic sound induced an even lower criterion than the presentation of the noise ($F(1,18) = 6.70$, $MSE = 0.05$, $p < .05$, $\eta^2 = 0.25$).

7.3.3 Discussion

The *d′* and *c* values obtained in the picture detection task when the picture was accompanied by a simultaneously-presented sound which was either semantically-congruent or incongruent were compared. The picture was accompanied by either a semantically-congruent sound or white noise (Experiments 7.2A and 7.2B, respectively). The results showed that none of the sound types gave rise to an increase in *d′* when the participants had to try and discriminate the presence (vs. absence) of a picture. Besides, in Experiment 7.2A, the participants' awareness of the semantic congruency between the sound and the picture did not rise above chance levels. However, the participants were more willing to respond 'Yes' (i.e., giving rise to a lower *c*) when a sound was presented. That said, a simultaneously-presented sound only influenced participants' response criterion rather than the detectability of the target itself (as measured by *d′*). Furthermore, a naturalistic sound triggered more of a change in the participants' response criterion than did the presentation of a burst of white noise (Experiment 7.2B), which may be due to the fact that the target consisted

of a meaningful picture. Therefore, the potential association between visual and auditory stimuli can also influence a participant's criterion for responding.

Given that participants' *d′* values were not influenced by the presentation of the semantically-congruent naturalistic sound when it was presented simultaneously with the picture, the next experiment thus attempted to test whether the sound can elicit a facilitatory effect on *d′* at an intermediate SOA between -346 and 0 ms. The aim of Experiment 7.3 was therefore to replicate the audiovisual facilitation effect at the -346 ms SOA (first reported in Experiment 7.1) and also to explore the semantic congruency effect at the -173 ms SOA. Furthermore, the semantic priming effect elicited by the presentation of a naturalistic sound (e.g., the sound of a dog barking) and a word presented acoustically (e.g., the spoken word 'dog') on performance in the picture detection task was also investigated. Given that both naturalistic sounds and spoken words provide semantic context, the prediction was that both kinds of sounds, with a sufficient duration to be processed, should be able to enhance a participant's visual sensitivity of the semantically-congruent picture.

7.4 Experiment 7.3

In Experiment 7.3, the onset of the sound, if present, led that of the picture by either 173 or 346 ms. In Experiment 7.3A, naturalistic sounds were presented; whereas in Experiment 7.3B, the sounds consisted of spoken words instead.

7.4.1 Experiment 7.3A

7.4.1.1 Methods

A new group of 17 volunteers (7 males, mean age of 24 years) from the same pool of participants as in Experiment 7.1 took part in this experiment. Three factors, semantic congruency (congruent vs. incongruent), SOA (-173 vs. -346 ms), and picture (present vs. absent) were manipulated. Note that in this experiment, a sound was presented on every trial, while a picture was presented in only half of the trials. Besides, in the picture-present trials, the sound and the picture were congruent in half of the trials and incongruent in the remainder of the trials. The hit rates for the congruent and incongruent conditions were therefore estimated from a quarter of the total trials, and the FA rate was estimated from the half of the trials that contained a sound but no picture (see Figure 7.1D). The other details were exactly the same as in Experiment 7.1.

7.4.1.2 Results

One participant achieved an overall accuracy of 99% (with an estimated d' greater than 5), indicating that he/she found the task too easy, so his/her data was excluded from further analysis. The d' and c data (see Figures 7.4A & 7.4B) were submitted to a two-way ANOVA with the factors of semantic congruency and SOA. In the analysis of d', none of the main effects were significant (Fs < 1.78, ps > .20, η^2s < 0.11); however,

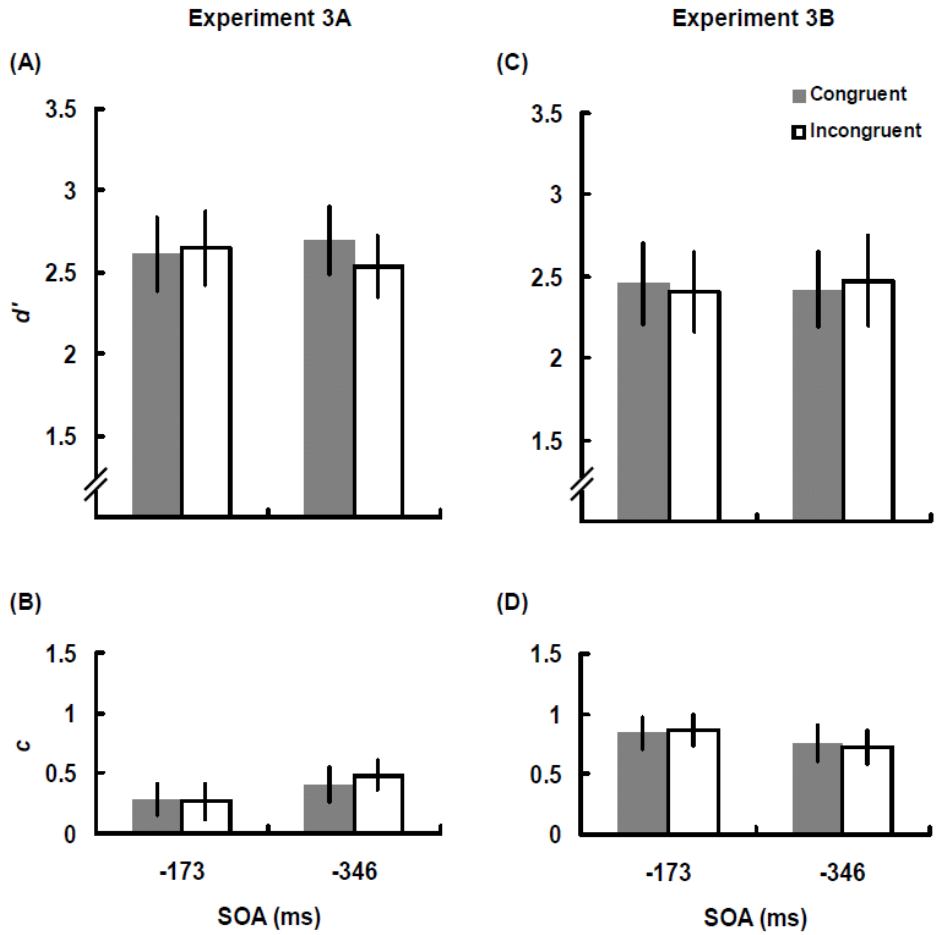


Figure 7.4. The results of sensitivity (*d'*, in panels A and C), and response criterion (*c*, in panels B and D), in Experiment 7.3A and 7.3B, respectively. Negative SOA values indicate that the auditory stimulus led the picture. The error bars indicate ±1 standard error of the means.

their interaction was significant ($F(1,15) = 4.36$, $MSE = 0.04$, $p = .05$, $\eta^2 = 0.23$). The simple main effect revealed that *d'* was higher in the congruent condition than in the incongruent condition when the SOA was -346 ms ($F(1,30) = 5.82$, $MSE = 0.04$, $p < .05$, $\eta^2 = 0.16$), while no such difference between the congruent and incongruent conditions was observed at an SOA of -173 ms ($F(1,30) = 0.26$, $p = .62$, $\eta^2 = 0.007$). In the analysis of *c*, none of the main effects were significant (Fs < 3.07, ps > .10, η^2s < 0.17); however, the interaction between semantic congruency and SOA was significant

($F(1,15) = 4.37$, $MSE = 0.01$, $p = .05$, $\eta^2 = 0.23$). The simple main effect showed that *c* was lower in the congruent condition than in the incongruent condition when the SOA was -346 ms ($F(1,30) = 5.83$, $MSE = 0.01$, $p < .05$, $\eta^2 = 0.16$), while no such difference was observed when the SOA was -173 ms ($F(1,30) = 0.26$, $p = .61$, $\eta^2 = 0.0006$). The results of the experimenter's unexpected inquiry at the end of the experiment showed that 14 out of 16 participants noticed the semantic congruency between the sound and the picture ($X^2(1) = 9.00$, $p = .003$).

7.4.2 Experiment 7.3B

7.4.2.1 Methods

Another group of 16 volunteers (four males, mean age of 20 years) from the same pool of participants as in Experiment 7.1 took part in this experiment. All of the participants were either native English speakers or else had learned English for more than 15 years. Three factors, semantic congruency (congruent vs. incongruent), SOA (-173 vs. -346 ms), and picture (present vs. absent) were manipulated. The sounds in this experiment consisted of spoken words recorded by a female native English speaker. The choice of word was based on the most commonly agreed name used to refer a given picture in the study of Snodgrass and Vanderwart (1980) and Bates et al. (2003). The sounds were presented at 62-66 dB SPL. Seventeen one-syllable words were compressed down to 350 ms, 9 two-syllable words were 450 ms, and four words with

three or above syllables were 500 ms. The other details were as in Experiment 7.3A.

7.4.2.2 Results

The *d′* and *c* data (see Figures 7.4C & 7.4D) were submitted to a two-way ANOVA with the factors of semantic congruency and SOA. In the analysis of *d′*, neither the main effects nor their interaction were significant (all Fs < 0.91, ps > .35, η^2s < 0.06). Similarly, in the analysis of *c*, neither the main effects nor their interaction were significant (all Fs < 1.47, ps > .24, η^2s < 0.09)[25]. The results of the unexpected inquiry showed that 13 out of 16 participants in this study noticed the semantic congruency between the sound and the picture ($X^2(1) = 6.25$, $p = .01$).

7.4.3 Discussion

Two critical results can be drawn from this experiment: First, the results of Experiment 7.3A demonstrate that a naturalistic sound can prime the detection of a semantically-congruent picture when the onset of the sound precedes that of the picture by 346 ms, thus replicating the results of Experiment 7.1. However, no such semantic priming effect was observed when the sound led the picture by only 173 ms. That said, a certain amount of time may be necessary to access the meaning of the sound (perhaps as much as 200~350 ms) in order to facilitate the processing of a

[25] Given that the duration of spoken words which consisted of more than two syllables lasted longer than 350 ms, one may wonder whether the inclusion of these items may have diluted the effect of semantic priming elicited by one-syllable words. The *d′* and *c* for just the 17 one-syllable-word items were therefore calculated. The results once again showed the same pattern: That is, none of the main effects nor their interaction were significant in the analysis of *d′* (all Fs < 3.37, ps > .09, η^2s < 0.18) or *c* (all Fs < 3.37, ps > .09, η^2s < 0.18).

semantically-congruent picture crossmodally. Second, when the sound consisted of a spoken word, no semantic priming effect was observed at either SOA (i.e., -173 or -346 ms). These results therefore indicate that naturalistic sounds and verbally-presented words do not prime visual pictures in quite the same way. This issue was further tested in the next experiment using a within-participants experimental design.

It should be noted that with regard to the response criterion in the condition in which the naturalistic sound led the picture by 346 ms, the participants lowered their criterion when the sound that they heard was semantically congruent as compared to when it was semantically incongruent. No such difference was observed in Experiment 7.1. This result should be considered carefully in terms of the following issues: When computing *d′* and *c* in this experiment, given that the congruent and incongruent conditions shared the same FA rate, any difference in the hit rate would be expected to influence *d′* and *c* in opposite directions; that is, a higher hit rate should give rise to an increase in *d′* but to a decrease in *c*. In fact, the estimated zero-biased *c* in such an experimental design is not unique (see Macmillan & Creelman, 2005, p. 117, on this point). Therefore, one should consider the meaning of response criterion in terms of psychological functioning: That is, did the participants truly adopt two response criteria, one for the congruent, and the other for the incongruent conditions? One possibility here is that the participants should only have adopted a single criterion in

order to respond, since the congruent and incongruent trials were intermixed within a block of trials and the detection of the presence of a picture may be independent of a participant's realization of the nature of any audiovisual semantic congruency that may be present. In this case, the criterion should be located somewhere between the estimated criteria in the congruent and incongruent conditions.

The second possibility is that the participants adopted a lower criterion when responding in the congruent trials than in the incongruent trials because they were aware of the semantic congruency between the sound and the picture. If participants intentionally used a strategy of responding more frequently in the congruent condition, *c* should have been lower in both the naturalistic sound and spoken word conditions since, in both Experiments 7.3A and 7.3B, most of the participants noticed the semantic congruency between the sounds and the pictures. The last possibility here is that the participants unconsciously adopted a lower criterion when responding in the congruent as compared to the incongruent condition. If this were to have been the case, it would suggest that the semantic priming effect induced by the naturalistic sound not only facilitated the breakthrough of a semantically congruent picture into awareness, but also potentially enhanced our participants' willingness to respond on a trial-by-trial basis.

7.5 Experiment 7.4

The aim of Experiment 7.4 was to try and replicate the difference in the priming effect at the -346 ms SOA resulting from the presentation of naturalistic sounds and spoken words on picture detection within the same group of participants. In Experiment 7.4A, the participants simply had to detect the presence of the picture as in the previous experiments reported in this chapter; in Experiment 7.4B, an additional identification task was also included.

7.5.1 Experiment 7.4A

7.5.1.1 Methods

A new group of 14 volunteers (four males, mean age of 25 years) from the same pool of participants as in Experiment 7.3B took part in this experiment. Three factors, semantic congruency (congruent vs. incongruent), sound type (naturalistic sound or spoken word), and picture (present or absent) were manipulated. The experimental blocks with the SOA of -346 ms used in Experiments 7.3A and 7.3B (containing naturalistic sounds and spoken words, respectively) were tested for each participant. Both the naturalistic sound and spoken word blocks were presented twice, and the order of presentation of these blocks was counterbalanced and intermixed across participants. The other details were exactly the same as in Experiment 7.3.

7.5.1.2 Results

The *d′* and *c* data (see Figures 7.5A & 7.5B) were submitted to a two-way ANOVA with the factors of semantic congruency and sound type. In the analysis of *d′*, the main effect of semantic congruency was significant ($F(1,13) = 4.73$, $MSE = 0.07$, $p < .05$, $\eta^2 = 0.27$). However, the interaction between semantic congruency and sound type was not significant ($F(1,13) = 0.12$, $MSE = 0.03$, $p = .74$, $\eta^2 = 0.01$). The latter null result may reflect the consistent trend in the data from the naturalistic sound and spoken word conditions. However, planned simple main effects were conducted since

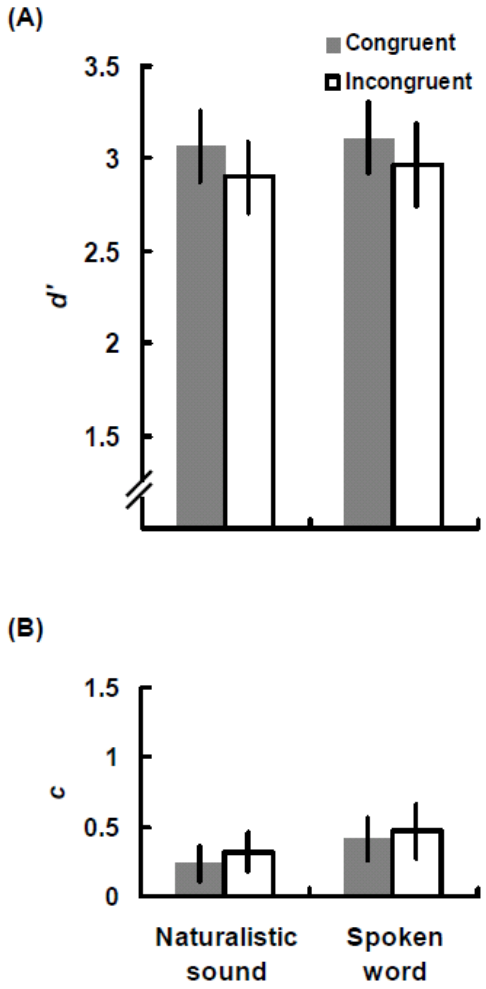


Figure 7.5. The results of (A) sensitivity (*d′*), and (B) response criterion (*c*), in Experiment 7.4A. The error bars indicate ±1 standard error of the means.

it is critical to compare the results in the present experiment to those observed in Experiment 7.3 (see Kirk, 1995, p. 118). The results showed that the semantic congruency factor was only significant when the participants heard the naturalistic sounds ($F(1,26) = 4.15$, $MSE = 0.05$, $p = .05$, $\eta^2 = 0.13$), but not when they listened to the spoken words ($F(1,26) = 2.84$, $p = .10$, $\eta^2 = 0.09$). Even though the difference in *d'* between the congruent and incongruent condition was similar in size in the naturalistic sound and spoken word conditions (0.17 and 0.14, respectively), more of the participants showed a negative semantic priming effect (i.e., *d'* was lower in the congruent than in the incongruent condition) in the spoken word condition (see Table 7.1).

Table 7.1. The number of participants whose performance exhibited positive semantic priming (*d'* in congruent > incongruent), null (*d'* in congruent = incongruent), or negative semantic priming (*d'* in congruent < incongruent) effects in Experiments 7.4A and 7.4B

Experiment		Semantic priming effect				
		Positive	Null	Negative	$X^2(2)$	*P*
7.4A	Naturalistic sound	8	4	2	4.00	.14
	Spoken word	8	1	5	5.29	.07
7.4B	Naturalistic sound	8	3	3	3.57	.17
	Spoken word	11	1	2	13.00	.002

In the analysis of *c*, neither the main effects nor their interaction were significant

(all Fs < 3.50, ps > .08, η^2s < 0.21). The results of the unexpected inquiry showed that 10 out of 14 participants noticed the semantic congruency between the sound and the picture ($X^2(1) = 2.57$, $p = .11$).

7.5.2 Experiment 7.4B

7.5.2.1 Methods

A new group of 14 volunteers (six males, mean age of 23 years) from the same pool of participants as in Experiment 7.3B took part in this experiment. The only difference from Experiment 7.4A was that a dual detection/identification task was used. Whenever the participants pressed the 'SPACE' bar in order to indicate that they detected a picture, the identification task followed: A frame with the question 'What was the picture?' was presented, and the participants were prompted to enter the name of the picture using the keyboard (free report) and allowed to press the 'SPACE' bar if they have no ideas. After answering regarding the identity of the picture, the participants pressed 'ENTER' to continue the detection task. The inclusion of the identification task in this experiment should have made the participants respond more conservatively in the detection task; however, any difference in the estimated d' in each condition is independent of the overall shift of criterion. The other details were the same as in Experiment 7.4A.

7.5.2.2 Results

The *d′* and *c* data (see Figures 7.6A & 7.6B) were submitted to a two-way ANOVA with the factors of semantic congruency and sound type. In the analysis of *d′*, there was a significant main effect of semantic congruency ($F(1,13) = 25.39$, $MSE = 0.02$, $p < .0005$, $\eta^2 = 0.66$). Neither the other main effect, nor the interaction term, were significant (Fs < 0.34, ps > .57, η^2s < 0.03). Planned simple main effects showed that the semantic congruency factor was significant both when the participants heard the naturalistic sounds ($F(1,26) = 9.99$, $MSE = 0.02$, $p < .005$, $\eta^2 = 0.20$), and when they heard the spoken words ($F(1,26) = 12.86$, $p < .005$, $\eta^2 = 0.26$). See Table 7.1 for the distribution of positive, null, and negative semantic priming effects. Note that more participants exhibited a positive priming effect in the spoken word condition than in the naturalistic sound condition. The analysis of *c* revealed a very similar pattern of results. That is, the main effect of semantic congruency was significant ($F(1,13) = 25.40$, $MSE = 0.005$, $p < .0005$, $\eta^2 = 0.66$), while none of the other terms were (Fs < 0.10, ps > .76, η^2s < 0.01). The results of the unexpected inquiry showed that 12 out of 14 participants noticed the semantic congruency between the sound and the picture ($X^2(1) = 7.14$, $p = .008$).

The accuracy of participants' performance in the identification task was generally low, presumably because the exposure duration was too short for participants to identify the picture easily (see Chapter 8). Note that when the participants accidentally entered the name associated with the sound in the congruent condition, the accuracy of

their visual identification performance would have been artificially overestimated. By contrast, such 'auditory' responses would not have increased the accuracy of their

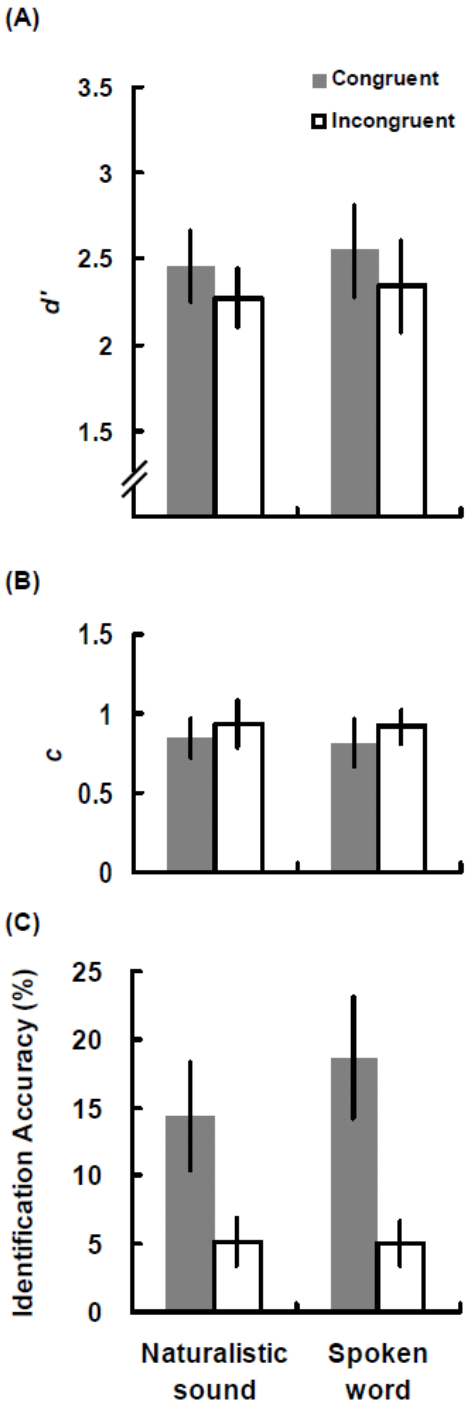


Figure 7.6. The results of (A) sensitivity (*d'*), (B) response criterion (*c*), and (C) accuracy of identification task in Experiment 7.4B. The error bars indicate ±1 standard error of the means.

responding in the incongruent condition. In order to correct for any such guessing bias, the frequency with which participants *entered the name of the sound* in both the incongruent and picture absent conditions was calculated and served as the FA rate for the congruent condition. The accuracy data from the congruent condition and the FA rate were entered into the following formula (Green & Swets, 1966; Macmillan & Creelman, 2005; Chapter 8):

(Accuracy - FA) / (1 - FA)

The adjusted accuracies of participants' visual target identification performance in each condition (see Figure 7.6C) were subjected to arc-sine transfer using the equation $X' = 2 * \sin^{-1}\sqrt{X}$ (Kirk, 1995) in order to linearize the percentage data. The transferred data were then submitted to a two-way ANOVA with the factors of semantic congruency and sound type. Both main effects were significant: Accuracy was higher in the congruent condition then in the incongruent condition (16.5% vs. 5.1%; $F(1,13) = 32.95$, $MSE = 0.06$, $p < .0005$, $\eta^2 = 0.72$), and the participants responded more accurately in the spoken word condition than in the naturalistic sound condition (11.8% vs. 9.7%; $F(1,13) = 5.75$, $MSE = 0.02$, $p < .05$, $\eta^2 = 0.31$). Even though the interaction was not significant ($F(1,13) = 1.84$, $MSE = 0.02$, $p = .20$, $\eta^2 = 0.12$), planned simple main effects showed that between the two sound types, only the accuracy in the congruent condition differed ($F(1,26) = 7.10$, $MSE = 0.02$, $p < .05$, $\eta^2 = 0.21$), but not in the incongruent condition ($F(1,26) = 0.60$, $p = .45$, $\eta^2 = 0.02$). The FA

rates in the naturalistic sound and the spoken word conditions were not significantly different (2.4% vs. 3.1%; $F(1,13) = 2.95$, $MSE = 0.02$, $p = .11$, $\eta^2 = 0.19$).

7.5.3 Discussion

The results of Experiment 7.4A are consistent with the results of Experiments 7.3A and 7.3B in that, at the -346 ms SOA, the naturalistic sounds induced a stronger priming effect than the spoken words when the participants simply had to detect the presence of a picture regardless of its identity. When a detection, as well as an identification, task was included in Experiment 7.4B, however, both the naturalistic sounds and spoken words gave rise to a similar semantic priming effect on performance in the picture detection task. Furthermore, the spoken word induced a stronger crossmodal priming effect on the picture identification task than the naturalistic sound in the congruent condition. This conclusion, however, should be qualified since the participants always heard a word that happened to be the correct answer that they had to respond with. In any case, these results suggest that the inclusion of an identification task boosted audiovisual semantic interactions between spoken words and pictures as compared to the results reported in Experiment 7.4A. In sum, the results of Experiment 7.4B suggest that when the participants had to perform a dual detection/identification task, naturalistic sounds and spoken words induced similar semantic priming effects.

7.6 General discussion

The results of the experiments reported in this chapter showed that the presentation of a naturalistic sound enhanced participants' sensitivity to a visually-presented semantically-congruent picture only when the onset of the sound occurred 346 ms before that of the picture (see Experiments 7.1 & 7.3A). On the other hand, when the auditory and visual stimuli were presented simultaneously, the presentation of a naturalistic sound simply reduced participants' response criterion to the visually-presented picture (see Experiments 7.2A & 7.2B). When the onset of the sound led by 346 ms, only naturalistic sounds, but not spoken words, enhanced the sensitivity of participants' responses in the picture detection task (Experiments 7.3A, 7.3B, & 7.4A). However, when the participants had to detect and identify the visual pictures, both naturalistic sounds and spoken words elicited similar semantic priming effects (Experiment 7.4B).

In the experiments reported in this chapter, signal detection theory was utilized to estimate participants' d' and c in order to investigate the outcome of the audiovisual interaction in terms of semantic congruency. This method allows one to segregate the modulatory effect of auditory stimuli on visual perception versus on responding. Two main conclusions can be drawn from the results reported here: First, the sound should lead the picture by about 350 ms in order for the meaning of the sound to be fully

accessed and, as a result, enhance the sensitivity of participants' perception of the semantically-congruent picture. Second, in terms of the semantic priming effects elicited by auditory stimuli on visual picture detection, the association between naturalistic sounds and pictures would appear to be stronger than that between spoken words and pictures. However, when the task also involves visual picture identification, both kinds of sound seem to be similarly effective in terms of facilitating participants' visual picture detection and identification performance. That said, the impact of naturalistic sounds and spoken words can be modulated by the task at hand. These two issues are addressed below.

7.6.1 The timecourse of the auditory modulation of visual perception

The results outlined here demonstrate that sufficient time is needed in order for a sound to access the associated meaning and thus induce genuine visual perceptual enhancement (see also Schneider, Engel, & Debener, 2008). The results here suggest that about 200-350 ms may be necessary in order for an auditory stimulus to access the semantic representations associated with it. This is much longer than the time required for a visual stimulus to access its meaning (Potter, 1975, 1999). Considering that, if the semantic representation of a visual stimulus is accessed while that of an auditory stimulus is still being processed, visual information should be expected to influence early auditory processing (see Jesse & Massaro, 2010; van Wassenhove, Grant, &

Poeppel, 2005). Hence, when a meaningful sound and picture are presented simultaneously, the crossmodal semantic modulation should be larger from visual to auditory processing than vice versa (for supporting evidence, see Yuval-Greenberg & Deouell, 2009). Nevertheless, the possibility should still be considered that visual and auditory stimuli facilitate each other interactively during stimulus processing to enrich information to each other mutually, especially when visual information is degraded or unfamiliar (see Chapter 8; Noppeney et al., 2010; Yuval-Greenberg & Deouell, 2007, 2009).

That said, the audiovisual semantic congruency effects demonstrated using two different approaches in other studies should be considered carefully. When adopting the redundant signals paradigm, the semantically congruent visual and auditory signals often map onto the same response. In this case, the cumulative distribution function of response times (RTs) for the audiovisual condition can violate the predictions of the race model (see Miller, 1982), thus indicating that the semantically-congruent visual and auditory signals can be summed to trigger a response (Laurienti et al., 2004; Molholm et al., 2004; Suied et al., 2009). However, when only one of the semantically-congruent stimuli (i.e., either the visual or the auditory stimulus) requires a response (e.g., the picture of a phone), the presentation of non-target stimuli in the other sensory modality (e.g., the sound of the phone ringing) did not shorten the response latency (see the Appendix in Suied et al., 2009). Hence, the facilitatory effect

in the audiovisual condition in terms of cumulative RTs in the redundant signal paradigm may be attributed to the visual and auditory stimuli both being linked to the same response (thus resulting in response redundancy), rather than genuinely due to their semantic congruency (see Experiment 3 in Miller, 1982, on this point). Similarly, in Experiments 7.2A and 7.2B, when a sound was presented simultaneously with a picture, the sound could only lower the response criterion, rather than increase sensitivity, for the picture irrespective of their semantic congruency. That said, the results obtained when the visual and auditory stimuli were presented simultaneously should be explained carefully, since the observed effects may be attributed to the processing at the decisional stage.

The second approach follows the 'unity assumption' which holds that semantic congruency serves as a cognitive factor to bind the information from different sensory modalities (see Spence, 2007; Welch & Warren, 1980). So, for example, observers can sometimes bind visual and auditory stimuli with larger spatiotemporal disparities when the pair of stimuli happen to be semantically congruent than when they are incongruent (Jackson, 1953; van Wassenhove et al., 2007; Vatakis et al., 2008; Vatakis & Spence, 2007). However, it is important to note that the unity effect is not observed for all combinations of stimuli (see Vatakis et al., 2008; Vatakis & Spence, 2008). One possible factor modulating human performance here is how well the observers have learned the associations between semantically-congruent visual and auditory stimuli

(see Mitterer & Jesse, 2010). Furthermore, recent evidence has demonstrated that the integration of visual and auditory stimuli resulting from their temporal coincidence can be dissociated from the integration due to stimulus attributes in terms of temporal limits of binding flickering audiovisual stimuli (Fujisaki & Nishida, 2010) or in terms of temporal window for multisensory integration (Soto-Faraco & Alsius, 2009). Hence, an experimental paradigm that requires the participant to access the features or identity of visual and auditory stimuli, rather than simply probing the spatiotemporal properties of the stimuli, may prove more appropriate when investigating the audiovisual semantic congruency effect.

7.6.2 The effect of naturalistic sounds vs. spoken words on the priming of visual pictures

The results reported in this chapter demonstrate that at the same SOA (when the sound led by 346 ms), only the presentation of naturalistic sounds induced crossmodal semantic priming in the visual detection task, while both naturalistic sounds and spoken words elicited crossmodal semantic priming effects in the dual detection/identification task. In order to explain the difference in the results elicited by the presentation of naturalistic sounds and spoken words, one need not necessarily assume that naturalistic sounds and spoken words access distinct semantic systems (see Dick et al., 2007; Noppeney et al., 2008; Saygin et al., 2003). Rather, it might be

argued instead that the processing required to access representations stored in semantic systems for naturalistic sounds and spoken words is the critical difference (see below).

The general reading-naming interference model proposed by Glaser and Glaser (1989) is expanded to a crossmodal situation in order to try and explain the interaction between naturalistic sounds and pictures versus between spoken words and pictures (for the latter case, see also Stuart & Carrasco, 1993). Figure 7.7 provides a simplified example of the stimuli associated with 'dog'. The visual picture input should go through a stage of visual feature processing before accessing its semantic representation (Potter, 1975; Potter & Faulconer, 1975), and perhaps the next step involves access to its associated name in the lexical systems. The naturalistic sound and spoken word inputs have to go through auditory feature analysis first; then, however, these two kinds of stimuli may be processed quite differently. Naturalistic sounds may, like visually-presented pictures, access their associated semantic representations first and only then the associated name; on the other hand, spoken words may, like visual words, access their associated lexical representations first and then the associated meaning (Cummings et al., 2006; Saygin et al., 2005).

Note that according to this model, not only the information processing pathways, but also the representations required by the task at hand (i.e., the output), determine whether any interaction would occur between the various input information (Glaser & Glaser, 1989; McLeod & Posner, 1984). So, for example, if the participants were to

complete a simple picture detection task, they should rely on the information provided by either visual features (e.g., 'I see four legs') or semantic representations (e.g., 'I see a dog'). In the case of presenting a naturalistic sound as the prime, the sound directly activates the common representations of the picture in the semantic system, and perhaps, the representation of visual features can be further enhanced by the top-down feedback modulation (e.g., Iordanescu et al., 2008). Therefore, the detection of the visual picture can be enhanced by the earlier-presented naturalistic sound. On the other hand, the semantic representation activated by the prior presentation of the spoken word may be too slow, or else too weak, to facilitate visual performance.

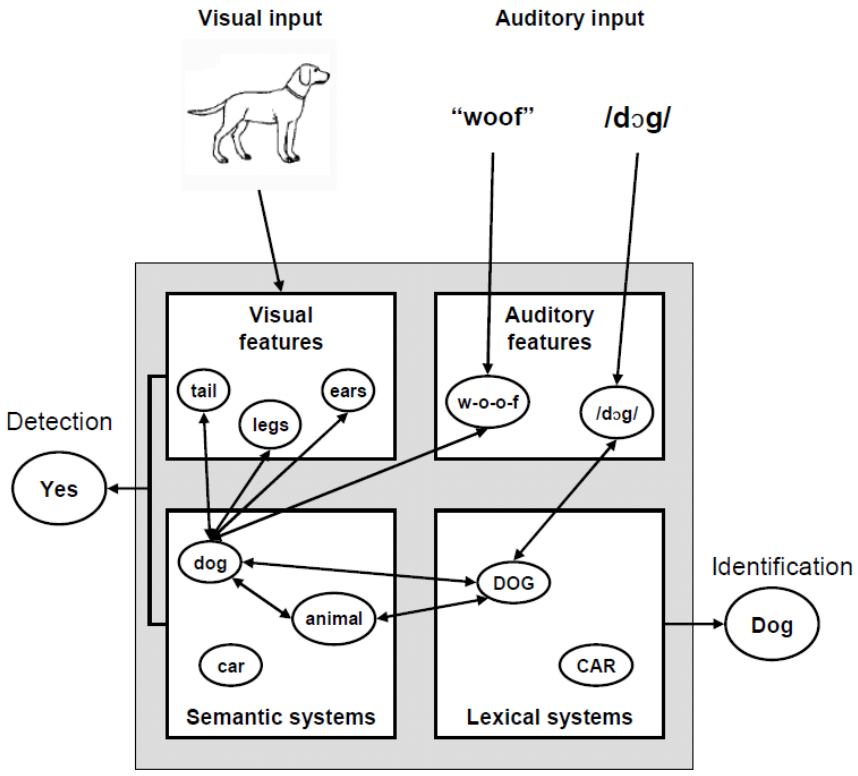


Figure 7.7. A cognitive structure of the processing of visual picture, naturalistic sound, and spoken word. This framework is adopted from Glaser and Glaser (1989).

In the dual detection/identification task, the accessing of all representations (including visual features, any semantic representation, and the associated name of the picture) is necessary, and so all representations in the semantic and lexical systems associated with the visual picture are likely to be activated (e.g., Smith & Magee, 1980). The earlier-presented sound, consisting of either naturalistic sounds or spoken words, should be able to activate this network starting from either the semantic or lexical representation. As a result, both types of sound should induce similar facilitatory effects on a dual visual picture detection and identification task.

Three further issues need to be tested in terms of the framework outlined here: First, the facilitatory effect of the congruent audiovisual stimuli in the semantic priming paradigm (as used in the experiments reported in this chapter) should be stronger than the interfering effect elicited by the presentation of incongruent audiovisual stimuli. This is because the inferference between semantically-incongruent audiovisual stimuli likely involves response competition (see La Heij, van der Heijden, & Schreuder, 1985). It may therefore be worth testing audiovisual semantic interactions by adopting typical Stroop-like interference paradigms in order to compare the interference of naturalistic sounds and spoken words on a visual stimulus in different tasks (e.g., Roelofs, 2005; Shimada, 1990; Stuart & Carrasco, 1993).

The second issue concerns whether crossmodal semantic interactions between written words and two types of auditory stimuli can be explained within a unitary

framework. In the present study, only audiovisual semantic interactions on visual picture perception were tested, and the framework proposed in Figure 7.7 therefore focuses on this. When comparing the semantic interactions by utilizing other types of visual and auditory stimulus combination, Iordanescu et al. (2008) showed that the presentation of a naturalistic sound reduced search latencies for a semantically-congruent target when the search display consisted of pictures but not when it consisted of written words. It is possible to explain these results according to the framework highlighted in Figure 7.7 if one assumes that the written word would access its lexical representation first and only then its semantic representation. Furthermore, the extended framework should be able to account for the results of Stroop-like interference between spoken and written words (e.g., Roelofs, 2005; Shimada, 1990).

Third, in the present framework, each linkage between representations is depicted with equal strength and a bidirectional arrow; however, it would certainly be worth testing whether each linkage is activated automatically and/or modulated by cognitive factors such as task demands. For example, Werner and Noppeney (2010b) have recently demonstrated superadditive activation in the STS (see below) for picture and naturalistic sound inputs irrespective of whether they are task-relevant or not, indicating that multisensory object integration may occur automatically (see also Experiments 7.3A and 7.4A in this chapter). In addition, in ERP studies, Orgs, Lange,

Dombrowski, and Heil (2007, 2008) have shown that the semantic priming from written words to naturalistic sounds was obligatory as indexed by the magnitude of N400 (an index of semantic processing); moreover, they also demonstrated that the size of priming effects can be modulated by task demands. Such results support the idea that the associated linkages are automatically activated while the magnitude of activation can be increased according to the goal at hand.

7.6.3 Human brain networks for auditory processing and audiovisual semantic interactions

The assumption here is that naturalistic sounds and spoken words go through different processing pathways before finally converging on a common semantic system. Neuropsychological studies provide evidence that is consistent with this view. For example, ERP studies have demonstrated that naturalistic sounds elicit an earlier N400 peak than spoken words (Cummings et al., 2006). Meanwhile, neuroimaging studies have shown that the right superior temporal gyrus is more sensitive to the processing of naturalistic sounds, and that the left STS, although a common area for the processing of naturalistic sounds and spoken words, more sensitive to spoken words (Thierry, Giraud, & Price, 2003; Thierry & Price, 2006; though see Meyer et al., in press). Finally, patient studies have demonstrated that recognition deficits for both naturalistic sounds and spoken words can result from the impairment of the left posterior middle

and superior temporal gyri (pMTG and pSTG, corresponding to Wernicke's area), which has been suggested to be the afferent area for auditory semantic processing (Dick et al., 2007; Saygin et al., 2003).

STS is an area that has frequently been shown to be involved in audiovisual semantic interactions (Beauchamp, Lee, Argall, & Martin, 2004; Stevenson & James, 2009; Werner & Noppeney, 2010b; for a review, see Doehrmann & Naumer, 2008). Using dynamic causal modeling, Noppeney et al. (2008) demonstrated that the audiovisual semantic congruency effect in STS is elicited in a bottom-up, rather than a top-down, manner. That is, visual and auditory stimuli are semantically congruent or not can be detected as soon as the information is sent to STS and need not rely on higher-level modulation (such as anterior cingulate or inferior frontal sulcus, see Noppeney et al., 2008). Note though that this neural mechanism may not be sufficient in-and-of-itself to explain why the congruent sound enhances visual sensitivity. We therefore suggest that a feedback enhancement from STS to visual areas should be necessary (e.g., Account *'critical role of feedback circuitry'* for multisensory integration in Driver & Noesselt, 2008).

7.7 Conclusions

The results reported in this chapter demonstrate that the presentation of a semantically congruent sound can give rise to a genuine enhancement in the sensitivity

of participants' visual picture identification in terms of the measurement of *d′* provided that two conditions are met: 1) The onset of the sound needs to occur prior to the presentation of the picture in order, presumably, for the observer to have accessed the semantic representation of the sound by the time that the picture is presented; and 2) The sound should strongly and rapidly enhance the representations of the visual picture that are required to make a response. Therefore, the influence of sounds on visual performance is flexible, rather than constant, and thus will vary according to the demands of task at hand. Here, a crossmodal framework is proposed, which is based on the general reading-naming interference model in terms of separate semantic and lexical systems (Glaser & Glaser, 1989) in order to account for the different audiovisual interactions between naturalistic sounds and pictures and between spoken words and pictures.

Chapter 8

When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures

8.1 Introduction

Researchers have been studying the processing of visually-presented pictures and how it can be influenced by the presentation of relevant visual information (such as the visual background) for nearly 40 years now (e.g., Biederman, 1972; Biederman, Mezzanotte, & Rabinowitz, 1982; Boyce & Pollastek, 1992; Boyce, Pollastek, & Rayner, 1989; Davenport & Potter, 2004; Glaser & Glaser, 1989; Henderson & Hollingworth, 1999; Hollingworth & Henderson, 1998; Intraub, 1981, 1984; Palmer, 1975; Potter, 1975, 1976). For example, researchers have shown that the theme of a picture can be detected easily in an on-line detection task while being difficult to recall in a subsequently-presented recognition task (Intraub, 1981; Potter, 1975, 1976). In addition, picture processing can be influenced by the semantic congruency of any

supplementary visual information that happens to be presented at the same time. For example, it has been shown that it is easier to detect or report a visual target when it is presented against a semantically-congruent background than when presented against a semantically-incongruent background instead (Biederman et al., 1982; Boyce et al., 1989; Davenport & Potter, 2004; Palmer, 1975; though see Hollingworth & Henderson, 1998). Similarly, when a word is superimposed on top of a picture, picture identification can either be facilitated by the presentation of a semantically-congruent word or else inhibited by the presentation of a semantically-incongruent word (Glaser & Dungelhoff, 1984). Furthermore, the identification or recognition of a picture can either be improved, when the subsequently-presented item (either a picture or a word) is semantically congruent with it (Glaser & Glaser, 1989; O'Conner & Potter, 2002), or else impaired when the subsequent item happens to be semantically incongruent (Glaser & Glaser, 1989; Intraub, 1984; Loftus & Ginn, 1984). Potter (1976) labeled the latter phenomenon *conceptual masking*.

These results have three important implications for the understanding of the processes underlying the identification of visually-presented pictures. First, the semantic information associated with a picture can be accessed within 100 ms of stimulus onset, maintained for a further 300 ms, and then either consolidated in memory or else rapidly forgotten (Intraub, 1999; Potter, 1975). Second, irrespective of the nature of the visual stimuli (e.g., picture or word), a shared abstract representation

of the semantic information may be accessed if the stimuli are associated with a common concept (Potter & Faulconer, 1975). Therefore, the interactions that are seen at the level of semantic congruency between pictures and words presumably rely on a common semantic system (see Glaser & Glaser, 1989). Third, the effects of supplementary visual information on participants' picture identification performance are not constrained by the need for strict temporal coincidence; instead, a subsequently-presented stimulus (e.g., lagging by as much as 350 ms in Loftus & Ginn's, 1984, study) can also influence performance. Potter (1993, 1999) proposed a module, called *conceptual short-term memory* (CSTM) in order to account for the semantic processing of visual stimuli. Potter visualized CSTM as consisting of a short-term buffer that serves as the interface between perception and memory. Once a visual stimulus has been identified, its meaning is rapidly accessed and maintained temporarily in CSTM. Furthermore, the maintained semantic representations are structured into a coherent context in order to be consolidated in memory; otherwise, unstructured (usually semantically incongruent) information would be rapidly forgotten (Potter, 1999).

The aim of the experiments reported in this chapter was to investigate semantic processing for the case of multisensory interactions between visual and auditory stimuli. The hypothesis was that a common semantic system can be accessed by a given picture and a semantically-congruent stimulus presented in another sensory

modality, such as audition (see Barsalou, 1999; Versace, Labeye, Badard, & Rose, 2009, on this issue). If this hypothesis were to be correct, the semantic congruency of an auditory stimulus might be expected to influence the identification of a visually-presented picture. Furthermore, audiovisual semantic interactions likely still occur even when the presentation of the sound lags the picture by a certain interval (around 300 ms).

8.1.2 Previous research on audiovisual semantic congruency effects

Jackson (1953) was perhaps the first to demonstrate the effect of semantic congruency on human information processing in a multisensory setting. He showed that people can tolerate a larger spatial (or for that matter temporal) separation between congruent (as compared to incongruent or unrelated) pairs of audiovisual stimuli and still bind them together. However, this explicit measure of audiovisual integration (namely, pointing to the apparent source of a sound) may unfortunately have resulted from the particular decision strategies adopted by participants (see Bertelson & Aschersleben, 1998; Choe, Welch, Gilford, & Juola, 1975; Vatakis & Spence, 2007). This is probably the reason why, in subsequent studies where the possible influence of such biases on performance have been ruled out, similar effects of audiovisual semantic congruency have not always been observed (e.g., Vatakis et al., 2008; Vatakis & Spence, 2008; though see Parise & Spence, 2009; Vatakis & Spence, 2007).

In addition, Koppen et al. (2008) have recently shown that audiovisual semantic congruency does not modulate the Colavita visual dominance effect (Colavita, 1974). In this paradigm, participants respond to visual and auditory targets accurately when they are presented alone; however, they sometimes fail to respond to the auditory targets when they are presented together with visual targets (see also Hecht & Reiner, 2009; Spence et al., in press). Across three experiments, Koppen et al. demonstrated that the semantic congruency between visual and auditory stimuli (congruent vs. incongruent, such as the picture of a dog accompanied by the sound of a dog 'woofing' or a cat 'meowing', respectively) had no effect on the magnitude of the Colavita visual dominance effect. That is, congruent sounds were detected no more accurately than incongruent sounds on those trials in which they were accompanied by a visual target.

On the other hand, there is some evidence from other experimental paradigms to show that the semantic congruency of visual and auditory stimuli can influence the latency of a participant's behavioral responses. For example, target detection latencies tend to be shorter when a visual target is accompanied by a semantically-congruent sound than when it is accompanied by a semantically-incongruent sound instead (Laurienti et al., 2004; Molholm et al., 2004; Suied et al., 2009). Performance in visual search tasks accompanied by a sound that happens to be semantically congruent with the visual target has also been shown to be facilitated, as compared to when search displays are accompanied by a target-incongruent sound instead (see Iordanescu et al.,

2008, 2010).

It should be noted here that those recent studies that have demonstrated an audiovisual semantic congruency effect have always done so in terms of the speed of participants' behavioral responses. As a consequence, it is unclear whether the audiovisual semantic congruency effect reported in these previous studies should be attributed primarily to a perceptual or to a decisional level effect. In the studies of Laurienti et al. (2004), Molholm et al. (2004), and Suied et al. (2009), the congruent visual and auditory stimuli corresponded to the same response key. It is therefore possible that, when information was provided both visually and auditorily (i.e., in the bimodal condition), the participant's response might have been triggered by the first stimulus that they happened to have perceived (i.e., a statistical facilitation of responses might have been observed, see Raab, 1962). All three of these studies have successfully ruled out statistical facilitation as the sole account of their data by comparing the cumulative distribution function of response times (RTs) for the bimodal condition to the predictions of the race model (see Miller, 1982). A further concern though is whether the facilitatory effect reported in the bimodal condition in these studies was truly due to semantic congruency, or alternatively, due to the visual and auditory stimuli both being linked with the same response (i.e., response redundancy, see Experiment 3 in Miller, 1982, on this point; Gondan, 2010; see also Section 1.3.1).

In the Appendix of their 2009 paper, Suied et al., for example, tried to disentangle

semantic congruency from response redundancy by mapping the semantically-incongruent stimuli (e.g., the picture of a phone and the 'ribbit' sound made by a frog) to the same response key. However, using this manipulation, the presentation of a telephone ringing sound did not speed-up the participants' responses to the picture of the telephone (i.e., no semantic congruency effect was observed). Evidence supporting the view that response redundancy is critical to the shortened RTs that have been observed in speeded tasks comes from the work of Sinnett et al. (2008). In their study, even though the visual and auditory targets were always semantically incongruent (e.g., a picture of 'traffic light' paired with the sound of a cat 'meowing'), when they were mapped to the same response key, RTs in the bimodal condition were nevertheless still significantly faster than the predictions of the race model. That is, given that the visual and auditory inputs were processed independently, they were still able to sum together to reach the criterion for initiating a response (this result is in line with the *independent coactivation model*, Miller, 1991). Therefore, one needs to exercise caution before explaining the RT data in speeded tasks in terms of semantic congruency effects since the results are easily confounded with response redundancy (i.e., a decision-level processing) effect (Santee & Egeth, 1982; Watt, 1991).

Another problem for the above-mentioned studies is that the visual target that participants had to detect, or search for, was clearly defined beforehand (Iordanescu et al., 2008; Koppen et al., 2008; Laurienti et al., 2004; Molholm et al., 2004; Suied et al.,

2009). What this means in practice is that the participants in these studies may have been able to accomplish their task simply by recognizing the salient features of the target picture (e.g., four legs in the case of the picture of dog and then inferring the rest). Besides, as the target had been well-defined in advance, a clear goal-directed control mechanism may have been active, which might have guided participants' attention to select the contingent stimulus (e.g., Folk, Remington, & Johnston, 1992). That is, the presentation of the semantically-congruent sound may have been selected for further processing, whereas the semantically-incongruent sound was ignored. The prediction here is that the crossmodal facilitation elicited by the presentation of a semantically-congruent sound can be observed, whereas the interference elicited by a semantically incongruent sound should be reduced, or even eliminated, consistent with the results of both Iordanescu et al.'s (2008) and Molholm et al.'s (2004) studies.

In the experiments reported in this chapter, a picture identification task was used to test for the existence of an audiovisual semantic congruency effect. This task is suggested to be less sensitive to the influence of response-level processing and goal-directed control than the speeded tasks used in previous research. In each trial, a picture from a corpus of at least 36 pictures was presented briefly and then immediately masked, and the participants had to try and identify the visual target. The accuracy of participants' unspeeded identification responses was recorded as the dependent variable. One of the advantages of this task is that the measurement of the

accuracy of participants' performance for briefly-presented targets provides a more sensitive measure of perceptual processing than does the measurement of RTs for targets presented for a longer time (e.g., Norman & Bobrow, 1975; Prinzmetal, McCool, & Park, 2005; Santee & Egeth, 1982). This is because any guessing bias can be estimated and separated from participants' performance, and hence the actual sensitivity of participants' responding can be determined (see Green & Swets, 1966). Another advantage of using a target identification task is that participants do not know the identity of the upcoming target on each and every trial. The increased uncertainty regarding the identity of the target should help to reduce any attentional selection resulting from goal-directed control that may have been taking place in previous studies (e.g., Yeh & Liao, 2008).

The primary goal of the experiments reported in this chapter was therefore to provide convergent evidence regarding the influence of audiovisual semantic congruency on the identification of masked visual targets, and further to examine whether the source of any congruency effect was due to the facilitation of performance in the congruent condition, the impairment of performance in the incongruent condition, or to both effects operating simultaneously. Furthermore, in previous studies of audiovisual semantic congruency, the visual and auditory stimuli were always presented simultaneously (Iordanescu et al., 2008; Koppen et al., 2008; Laurienti et al., 2004; Molholm et al., 2004; Suied et al., 2009). Given that the semantic congruency

effect that has been reported within the visual modality can be induced by the presentation of both simultaneous and asynchronous stimuli, it seemed worthwhile to investigate whether the same was true in a crossmodal (i.e., audiovisual) setting as well. The SOA between the picture and the sound was therefore manipulated; however, because the goals of this chapter were primarily to understand the processing of a picture after it had been presented, rather than the semantic priming of a picture by sound (e.g., Schneider, Engel, & Debener, 2008; see also Chapters 6 & 7),[26] the auditory stimulus was only presented simultaneously or else delayed with respect to the onset of the picture. In Experiments 8.1 and 8.2, the onset of the sound occurred simultaneously with that of the picture. In Experiments 8.3 and 8.4, the onset of the sound was delayed by 307 and 533 ms with respect to that of the picture. In Experiment 8.5, the SOAs (0, 333, 533 ms) were randomized on a trial-by-trial basis and tested within the same group of participants. Based on the results of previous semantic congruency studies in visual perception, the prediction was that any audiovisual semantic congruency effect should be observed at SOAs shorter than around 300 ms, but not at SOAs longer than 500 ms.

[26] A pilot experiment was conducted to test the semantic priming effect in the picture identification task using the same stimulus materials as for the experiments reported in this chapter. A group of 16 naïve volunteers (three males, mean age of 20 years) took part in this experiment. The experimental design was the same as in Experiment 8.3 (see Section 8.4) except that the onset of the sound led the picture by 358 ms. The results revealed a significant main effect of the sound condition ($F_1(2,30) = 24.76$, $MSE_1 = 0.009$, $p_1 < .0001$, $\eta^2_1 = 0.62$; $F_2(2,30) = 20.51$, $MSE_2 = 0.008$, $p_2 < .0001$, $\eta^2_2 = 0.58$). A *post-hoc* Fisher-Hayter test revealed that the adjusted accuracy of participants' performance was significantly higher in the congruent condition than in the incongruent condition ($ps < .01$; $q(30,2).01 = 3.89$, $MSE_1 = 0.009$; $MSE_2 = 0.008$) and the noise condition ($ps < .01$). The difference between the accuracy of participants' performance in the incongruent and noise conditions did not reach statistical significance. This result is consistent with the results of the picture detection task reported in Experiment 7.1.

8.2 Experiment 8.1

Experiment 8.1 was designed to investigate whether the semantic congruency (congruent vs. incongruent) of a sound would influence the ability of participants to identify briefly-presented line-drawing pictures. The prediction was that the presentation of a semantically-congruent sound would facilitate, whereas the presentation of a semantically-incongruent sound would impair, participants' performance relative to a control condition in which the picture was accompanied by a burst of white noise.

8.2.1 Methods

8.2.1.1 Participants

Seventeen volunteers (six males, mean age 26 years) took part in this study in exchange for course credit or a five pound (UK Sterling) gift voucher. All of the participants had normal or corrected-to-normal vision and hearing by self-report, and were naïve as to the purpose of the study.

8.2.1.2 Apparatus and stimuli

The stimuli were presented on a 15-inch CRT monitor (75 Hz refresh rate) controlled by a personal computer. The participants sat at a viewing distance of 58 cm from the monitor in a dimly-lit experimental chamber. Each trial consisted of the presentation of a fixation stimulus (a 0.69°×0.69° cross), a target picture, and a pattern

mask (see Figure 8.1A). All of the stimuli were presented in black in the center of a white background. Forty-two outline-drawing figures taken from the Snodgrass and Vanderwart (1980) database were used as the visual target stimuli. The mask consisted of random curves and extended an area of 5.32°×5.32°, sufficient to cover the target pictures completely.

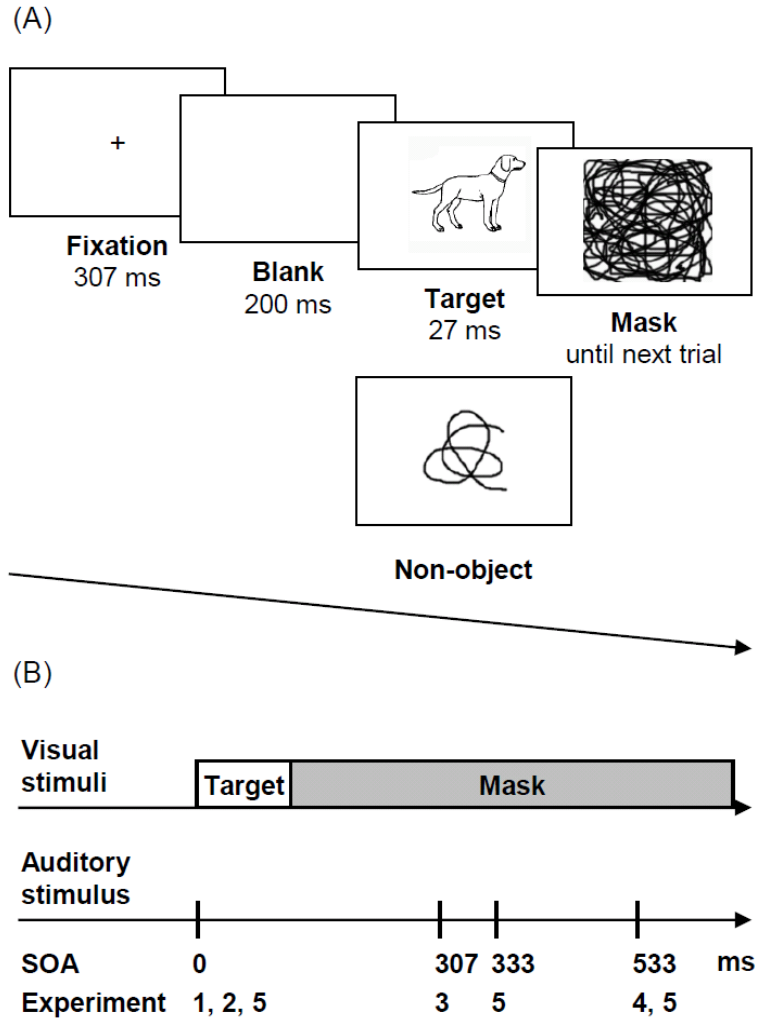


Figure 8.1. (A) The sequence of four frames presented in each trial: a fixation stimulus, a blank, a target picture (e.g., dog), and a pattern mask. An example of one of the non-object patterns used in Experiments 8.2 through 8.5 is also shown. (B) A schematic representation of the manipulation of SOAs between the visual target picture and the task-irrelevant accessory sound across the five experiments reported in this study.

The congruent auditory stimuli were chosen to match the pictures: e.g., the sound of a dog's barking was matched with the picture of a dog; the sound of a creaking door was paired with the picture of a door (see Table 8.1). All of the sound files were downloaded from www.findsounds.com (on 08/08/2008). The sounds were edited to start from the beginning of the sound file and to last for 350 ms with equalized peak amplitudes. The sounds were presented over closed-ear headphones and ranged in loudness from 66 to 76 dB SPL. The white noise in the control condition was presented at 72 dB SPL.

8.2.1.3 Design

On each trial, a picture was accompanied by a sound that was semantically congruent, semantically incongruent, or else neutral (i.e., white noise) with respect to the visual target. Semantic congruency was manipulated as a within-participants factor. In the congruent condition, the sound matched the picture. In the incongruent condition, for fear of any unwanted semantically-related priming effects occurring, the paired sound and picture belonged to different semantic categories. In other words, a picture of an animal was always accompanied by the sound of a tool, or instrument, and so on. In the noise condition, the picture was accompanied by a 350 ms burst of white noise. The onsets of the picture and sound were synchronized.

Table 8.1. Stimulus materials used in the experiments reported in Chapter 8.

Picture	**Sound**	
	Congruent	**Incongruent**
Accordion	Sound of accordion	Scissors
Axe	Chopping sound	Frog
Bat	Bat hitting ball	Rooster
Bee	Humming bee	Hammer
Bell	Bell ringing	Cat
Bird	Bird chirping	Drum
Bottle	Opening a bottle	Horse
Cannon	Cannon fire	Fly
Car	Car ignition sound	Cow
Cat	'Meow'	Lock
Cow	'Moo'	Gun
Dog	'Woof'	Telephone
Door	Creek	Eagle
Drum	Drumstick hitting drum	whistle
Duck	'Quack'	Helicopter
Eagle	Eagle call	Axe
Elephant	Elephant trumpet	Trumpet
Flute	Sound of Flute	Goat
Fly	Buzzing fly	Well
Frog	'Ribbit'	Piano
Goat	'Baa-baa'	Cannon
Grasshopper	Grasshopper chirping	Door
Guitar	Sound of guitar	Light switch
Gun	Gun shot sound	Owl
Hammer	Sound of hammer hitting	Elephant
Helicopter	Spinning helicopter motor	Bat
Horse	Naying of horse	Guitar
Light switch	Clicking of switch	Grasshopper
Lock	Snap of lock	Bee
Motorcycle	'Vroom-vroom'	Tiger
Mouse	Squeak	Bottle
Owl	'Hoot'	Bell
Piano	Sound of piano	Motorcycle
Pig	'Oink'	Flute
Rooster	Crowing sound	Accordion
Scissors	Clipping scissors	Pig
Telephone	Telephone ringing	Duck
Tiger	Roaring sound	Windmill
Trumpet	Sound of trumpet	Bird
Well	Water dripping	Dog
Whistle	Whistling	Car
Windmill	Spinning windmill	Mouse

Three blocks of experimental trials were presented. All 42 pictures were presented once in each block of trials. A third of the pictures were assigned to each of the three sound conditions. The sound condition factor was manipulated according to a Latin square design across the three blocks of trials. The 42 pictures were presented in a completely randomized order within each block of trials, and the order in which the three blocks were presented was counterbalanced across participants.

8.2.1.4 Procedure

Prior to the start of the experiment, two preview sessions including all of the pictures and sounds were presented. In the visual session, all of the pictures were shown together with their names. Each picture-name pair was presented for 1012 ms. The next picture-name pair was then presented after a further 506 ms blank period. Similarly, the auditory session involved the presentation of all of the sounds and their matching names. Each sound was presented for 350 ms while its name was presented for 1012 ms. The next sound-name pair was then presented after a further 506 ms blank interval. The order of presentation of each picture-name pair or sound-name pair was completely randomized within each session across participants.

In the main experimental session, the participants initiated a trial by pressing the 'ENTER' key on the keyboard placed in front of them. The fixation cross was presented for 307 ms, followed by a blank screen for 200 ms. Next, the target picture

and the sound were presented together. The picture was presented for 27 ms (two frames at 75 Hz refresh rate) and masked by the pattern mask immediately afterward (see Figure 8.1B). The mask was used to control the presentation duration of the visual target and the difficulty of the visual task. The participants were instructed to enter the name of the picture on the keyboard. They were also informed that they should respond to the picture rather than to the sound, and to guess if they were uncertain.

A practice session consisting of five trials (including all three sound conditions) was conducted prior to the main experiment. The stimuli in the practice session were not used in the main experiment. The participants had to repeat the practice session if they failed to report the name of the picture correctly in at least two of these five trials. The adoption of this criterion resulted in approximately 40% of the participants repeating the practice session. However, all of the participants achieved this criterion the second time round.

8.2.2 Results

When participants entered the name of the picture or a word with a similar meaning (e.g., ‘chicken’ for the picture of a rooster) their response was counted as correct. The criterion was held constant across all five experiments reported in this study. The data from one participant (female) was excluded because her overall adjusted accuracy (see below) fell below 10% correct, leaving the data from 16

participants for further analysis.

If participants accidentally entered the name associated with the sound in the congruent condition (i.e., without having seen the visual target), the accuracy of their visual object identification performance would have been artificially overestimated. By contrast, such 'auditory' responses would not have increased the accuracy of their responding in either the incongruent or noise conditions. In order to correct for any such guessing bias, the frequency with which participants *entered the name of the sound* in the incongruent condition was calculated and served as the false alarm (FA) rate for the congruent condition. The accuracy data from the congruent condition and the FA rate were entered into the formula reported by Intraub (1984):

$$(\text{Accuracy} - \text{FA}) / (1 - \text{FA})$$[27]

The adjusted accuracies of participants' visual target identification performance in each condition (see Table 8.2) were subjected to arc-sine transfer using the equation $X' = 2 * \sin^{-1}\sqrt{X}$ (Kirk, 1995) in order to linearize the percentage data[28]. The transferred data were then submitted to a one-way ANOVA with the factor of sound condition. The analysis revealed a significant main effect ($F(2,30) = 33.44$, $MSE = 0.013$, $p < .0001$,

[27] It was hypothesized that, in the picture identification task, responding correctly to the picture should depend on the activation of the appropriate stimulus representation above a certain threshold level, while the criterion set for responding to a target should not change during the course of the experiment. Hence, the response to the identity of the sound should reflect pure guessing rather than misperceiving the picture. Therefore, in line with the *high threshold model*, sensitivity was calculated according to the equation (Accuracy - FA) / (1 - FA) (Green & Swets, 1966; Macmillan & Creelman, 2005). In the congruent condition, accuracy scores should be adjusted by an estimated FA rate; in the incongruent and noise conditions, the FA rate was considered as 0.

[28] Statistical analysis of the raw data in all five experiments revealed exactly same pattern of results.

η^2 = 0.69). A *post-hoc* Fisher-Hayter test (Kirk, 1995) revealed that the adjusted accuracy of participants' performance was significantly higher in the congruent condition than in either the incongruent or noise conditions (both *p*s < .01, *q*(30,2).01 = 3.89, *MSE* = 0.013). Furthermore, participants' responses were also significantly more accurate in the noise condition than in the incongruent condition (*p* < .05, *q*(30,2).05 = 2.89). See Figure 8.2 for the magnitude of the semantic congruency effects compared to the noise control condition reported in Experiment 8.1.

Table 8.2. Mean accuracy (%) and standard error (in parentheses) of participants' responses in Experiments 8.1 to 8.4.

Experiment	Sound Condition						
	Congruent					Incongruent	Control
Experiment	% Correct	FA1	Adjusted 1	FA2	Adjusted 2		
8.1	70.7 (3.4)	10.7 (2.7)	65.8 (4.3)	-	-	50.9 (5.7)	55.8 (4.8)
8.2	69.7 (4.3)	2.8 (1.2)	69.7 (4.7)	7.0 (2.0)	68.0 (5.0)	61.9 (5.5)	63.8 (5.1)
8.3	76.8 (4.4)	0.9 (0.3)	76.5 (4.6)	5.1 (1.8)	75.9 (4.4)	71.0 (5.1)	72.9 (4.8)
8.4	76.6 (2.7)	0.4 (0.2)	76.5 (2.8)	2.4 (0.9)	75.8 (3.0)	73.2 (2.8)	74.3 (3.4)

FA1: The false alarm (FA) rate was estimated from the frequency with which the participants entered the name of the object indicated by the sound as the visual target in the incongruent condition; Adjusted 1: The accuracy of congruent condition was adjusted by FA1.

FA2: FA rate was estimated from the frequency with which the participants entered the name of the object indicated by the sound as the visual target in the non-object condition; Adjusted 2: The accuracy of congruent condition was adjusted by FA2.

Control: In Experiments 8.1, 8.3, and 8.4, control was the noise condition; in Experiment 8.2, control was the no-sound condition.

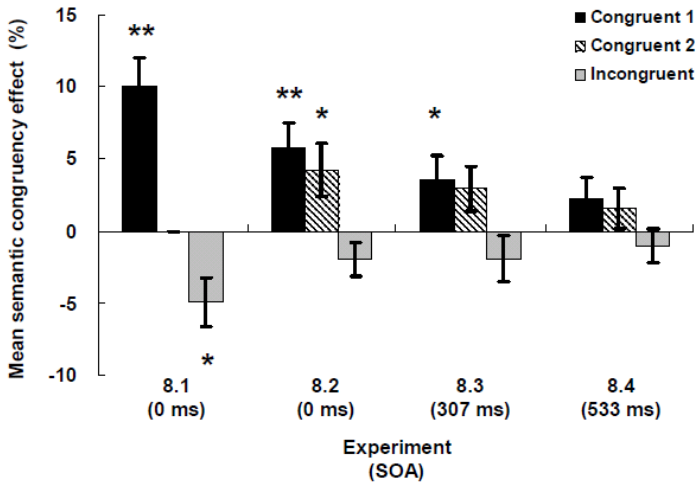


Figure 8.2. Graph highlighting the semantic congruency effects (and ±1 standard error of means) reported in Experiment 8.1 to 8.4. The magnitude of the effect was calculated by subtracting the accuracy of participants' performance in the semantically congruent condition (or semantically incongruent condition) from that in the noise condition (Experiments 8.1, 8.3, and 8.4) or the no-sound condition (Experiment 8.2). Congruent 1 and 2 refers to the accuracy adjusted from FA1 and FA2, respectively (see Table 8.2). Positive values indicate the improvement in performance in the semantically congruent condition, whereas negative values indicate the impairment in the semantically incongruent condition relative to the noise (or no-sound) condition. * denotes a significant difference (*: $p < .05$; **: $p < .01$).

8.2.3 Discussion

The results of Experiment 8.1 highlighted the existence of an audiovisual semantic congruency effect: That is, the participants were better able to identify the masked pictures when a semantically-congruent sound was presented at the same time than when a semantically-incongruent sound was presented instead. This result is consistent with the results of other studies in which the presentation of a semantically-congruent sound has been shown to facilitate the ability of participants to rapidly detect, or search for, a given visual target (Iordanescu et al., 2008; Laurienti et

al., 2004; Molholm et al., 2004; Suied et al., 2009). Taken together, it would therefore appear that audiovisual semantic congruency effects can affect both the speed and accuracy of participants' behavioral performance.

The results of Experiment 8.1 also showed that the presentation of a semantically-congruent sound facilitated, whereas the presentation of a semantically-incongruent sound impaired, participants' picture identification performance when compared to their performance in a control condition involving the presentation of white noise. Similar results (in terms of the RT data) were reported by both Laurienti et al. (2004)[29] and Suied et al. (2009), while only a facilitatory effect in the semantically-congruent condition was observed in the studies of Iordanescu et al. (2008) and Molholm et al. (2004). This inconsistency in terms of whether or not impaired performance is observed in the semantically-incongruent condition may be attributable to the specific control conditions used in the various studies; in particular, they were unimodal in these four studies (i.e., respond to either a visual or auditory target which is presented alone).

The next experiment was therefore designed to investigate whether both facilitatory and inhibitory effects of audiovisual semantic congruency would still be

[29] In Laurienti et al.'s (2004) study, RTs in the incongruent condition were slower than those reported in the unimodal (both visual and auditory) conditions. However, the stimuli used in the incongruent condition consisted of the pairing of a target in one modality with a task-irrelevant stimulus in the other modality. A similar manipulation was actually called the *single target with distractor* condition in Miller's (1991) study. The RT for this condition in Miller's study was even slower than that observed in the *incongruent redundant targets* condition in which the stimuli were both task-relevant but incongruent for the visual and auditory modalities.

observed when the participants' performance in the congruent and incongruent trials was compared to that in just such a unimodal control condition, here involving picture identification in the absence of any sound. The prediction regarding participants' performance in this condition is not entirely obvious: On the one hand, the no-sound condition might be expected to give rise to performance that is no different from that seen in the incongruent conditions, as in the studies of Iordanescu et al. (2008) and Molholm et al. (2004). Alternatively, however, the no-sound condition might also be expected to give rise to a result lying between that seen in the incongruent and congruent conditions, just as reported in the noise condition of Experiment 8.1, given that in neither case is any useful semantic information provided via the auditory modality. A final possibility is that when the visual picture is presented in isolation, performance is even better than in the congruent condition, because participants can fully focus their attentional resources on the processing of the target picture. Previous research has documented the deterioration of participants' visual target identification performance when additional information is presented within either the same modality (e.g., Davenport & Potter, 2004; Kahneman, Treisman, & Burkell, 1983) or else in a different modality (e.g., Shimojo, Watanabe, & Scheier, 2001; Spence, Ranson, & Driver, 2000). Researchers have attributed this performance decrement to the need for participants to engage their limited attentional resources in order to filter out the irrelevant non-target information (Lavie, 2005), or even to the fact that the non-target

information may itself serve to divert a participant's attention away from the target (see Dalton & Spence, 2007).

8.3 Experiment 8.2

There were two principal goals in Experiment 8.2: First, to try and replicate the audiovisual semantic congruency effect in the masked picture identification task reported in Experiment 8.1. The second aim was to compare the semantic congruency effect elicited by the presentation of a sound to a unimodal (i.e., visual only) baseline condition.

8.3.1 Methods

8.3.1.1 Participants

Eighteen naïve volunteers (two males, mean age 24 years) from the same pool of volunteers as in Experiment 8.1 took part in this experiment.

8.3.1.2 Apparatus, stimuli, design, and procedure

The three levels of the sound condition factor (consistent, inconsistent, and no-sound) were manipulated on a within-participants basis. The former two conditions were identical to those presented in Experiment 8.1. In the no-sound condition, the visual stimulus was presented in the absence of any auditory stimulation. A novel condition, in which a non-object picture was presented, was also included in this

experiment. In this non-object condition, 14 non-object visual patterns (see Figure 8.1A for an example) were accompanied by naturalistic sounds associated with those pictures used in the no-sound condition. The participants had to respond by pressing the 'SPACE' bar in the non-object trials. The inclusion of these non-object trials was designed to reduce the likelihood that the participants would respond by trying to guess the name of the sound in those trials in which they were unable to identify the picture. It also provided a second estimate of the FA rate, namely, the frequency with which the participants entered the name of the sound when an object picture was not presented. Fifty-six trials were presented in each block of trials, with the same number of trials being presented in the congruent, incongruent, no-sound, and non-object conditions, respectively. All of the other details were the same as in Experiment 8.1.

8.3.2 Results

The data from two participants were excluded from the data analysis: They entered the name of the sound in 64% and 45% of trials in the non-object condition, respectively (this compares to an average of only 7% for the other participants). Given that these participants failed to follow the instructions not to respond to the auditory stimulus, their data were discarded, leaving the data from 16 participants for further analysis (see Table 8.2 for the results).

The adjusted accuracy data for the three sound conditions was submitted to a

one-way ANOVA. This analysis revealed a significant main effect ($F_1(2,30) = 10.13$, $MSE_1 = 0.012$, $p_1 < .0005$, $\eta^2_1 = 0.40$; $F_2(2,30) = 5.96$, $MSE_2 = 0.012$, $p_2 < .01$, $\eta^2_2 = 0.28$)[30]. A *post-hoc* Fisher-Hayter test showed that the adjusted accuracy of participants' performance was significantly higher in the congruent condition than in either the incongruent (both *p*s < .01, $q(30,2).01 = 3.89$, $MSE_1 = 0.012$; $MSE_2 = 0.012$) or no-sound conditions ($p_1 < .01$; $p_2 < .05$; $q(30,2).05 = 2.89$). The difference in accuracy between the latter two conditions failed to reach statistical significance.

8.3.3 Discussion

The results of Experiment 8.2 once again highlighted the fact that it was easier for participants to identify masked line drawings when they were presented together with semantically-congruent sounds than when they were accompanied by semantically-incongruent sounds, thus replicating the audiovisual semantic congruency effect first reported in Experiment 8.1. Participants' performance in the no-sound condition was not significantly different from that reported in the incongruent condition, which is consistent with the findings of previous research (see Iordanescu et al., 2008; Molholm et al., 2004). A possible explanation for this finding is that in the no-sound condition, the participants were simply less alert than in those trials in which

[30] The FA rate was first estimated from the frequency with which the participants entered the name of the sound in the *incongruent* condition (i.e., using the same procedure as in Experiment 8.1). These statistical values are denoted by *F1* etc. An alternative estimation of the FA rate was derived from the frequency with which the participants entered the name of the sound in the *non-object* condition. These statistical values are denoted by *F2* etc. Generally speaking, the same pattern of results was obtained when the FA rates were estimated from the incongruent and the non-object conditions. Exceptions are specified in the text.

a sound was presented (e.g., Fiebelkorn, Foxe, & Molholm, 2010; Nickerson, 1973; Robertson et al., 1998; Spence & Driver, 1997). On the other hand, this result may also imply that the presentation of an auditory stimulus did not in-and-of-itself interfere with the participants' processing of the visual target. It has been argued elsewhere that crossmodal interference is typically weaker than intramodal interference, at least at the earliest stages of human information processing (e.g., Alais et al., 2006; Laurienti et al., 2004). These results are consistent with previous claims that perceptual/attentional processing resources may be, at least to a certain extent, independent for different sensory modalities (Duncan, Martens, & Ward, 1997; Hancock, Oron-Gilad, & Szalma, 2007; Rees et al., 2001; Soto-Faraco & Spence, 2002; Wickens, 1980, 1984; though see Arnell & Larson, 2002; Santangelo et al., 2007; Spence et al., 2001). By contrast, central processing resources are thought to constitute a relatively unitary pool of resources regardless of the sensory modality of input information (e.g., as when loading working memory or executive control functions; Brand-D'Abrescia & Lavie, 2008; Klemen et al., 2009; Potter, Chun, Banks, & Muckenhoupt, 1998; Spence & Soto-Faraco, 2009; Treisman & Davies, 1973). Semantic processing would thus appear to represent a case of central processing in which the information from the different senses is pooled and hence can interact crossmodally.

It is important to note that while the pattern mask used in Experiments 8.1 and 8.2 should have blocked visual feature processing, it should not have interfered with the

conceptual processing of the picture by participants (see Kiefer, 2002; Marcel, 1983). Therefore, even though the picture was only presented for 27 ms and then masked, and the identification of the sound may well have occurred long after the picture had been masked, the semantic information contained in the auditory and visual stimuli nevertheless still interacted. Given that conceptual processing takes several hundred milliseconds to complete (Potter, 1975), it is presumably possible for the later-identified sound to influence the processing of the picture in a retrospective manner. To further test this hypothesis, the onset time of the sound was delayed with respect to that of the picture in the next experiment.

8.4 Experiment 8.3

In this experiment, the onset of the sound occurred 307 ms after that of the picture. The prediction was that the semantically-congruent sound would still likely facilitate, while the semantically-incongruent sound would still inhibit, participants' picture identification performance, as compared to the performance seen in the control noise condition.

8.4.1 Methods

8.4.1.1 Participants

Sixteen naïve volunteers (three males, mean age of 18 years) from the same pool of volunteers took part in this experiment.

8.4.1.2 Stimuli, apparatus, design, and procedure

Three levels of the sound condition factor (congruent, incongruent, and noise) were manipulated on a within-participants basis, just as in Experiment 8.1. The non-object trials, first introduced in Experiment 8.2, were included here as well. The sound was presented 280 ms after the onset of the pattern mask (i.e., 307 ms after the onset of the picture, see Figure 8.1B). All other details were the same as for Experiment 8.2.

8.4.2 Results and discussion

The adjusted accuracy of participants' picture identification performance is shown in Table 8.2. The main effect of sound condition was again significant ($F_1(2,30) = 6.10$, $MSE_1 = 0.013$, $p_1 < .01$, $\eta^2_1 = 0.29$; $F_2(2,30) = 4.30$, $MSE_2 = 0.013$, $p_2 < .05$, $\eta^2_2 = 0.22$). A *post-hoc* Fisher-Hayter test revealed that the adjusted accuracy of participants' performance was also significantly higher in the congruent condition than in the incongruent condition (ps $< .01$; $q(30,2).01 = 3.89$, $MSE_1 = 0.013$; $MSE_2 = 0.013$). The adjusted accuracy was significantly higher in the congruent condition than in the noise condition when the FA rate was estimated from the incongruent condition ($p_1 < .05$; $q(30,2).01 = 2.89$). The difference between the accuracy of participants' performance in the incongruent and noise conditions did not reach statistical significance.

Although the onset of the sound was delayed by 307 ms with respect to that of the

picture in this experiment, the results nevertheless still exhibited the facilitatory effect of the semantically-congruent sound on participants' performance. However, the interfering effect of the semantically-incongruent sound was numerically much smaller in this experiment and failed to reach statistical significance. These results therefore imply that the facilitatory and interfering effects elicited by the presentation of semantically congruent versus incongruent sounds likely have somewhat different time-courses.

Contrary to the results of previous unimodal visual studies showing that a subsequently-presented semantically-incongruent picture can induce a conceptual masking effect (Intraub, 1984; Loftus & Ginn, 1984), a subsequently-presented auditory stimulus cannot. In the present experiment, the SOA between the picture and sound was 307 ms (note that the SOA between the target and mask pictures was 112 ms in Intraub's, 1984, study and 350 ms in Loftus and Ginn's, 1984, study), which should have been within the time window known to give rise to conceptual masking. However, it is important to bear in mind that it may take longer to access the meaning of a sound than to access the meaning of a picture, since the meaning of the auditory signal evolves over time, whereas all of the visual features in the picture are presented at once (e.g., Ballas, 1993). The fact that the meaning of the sound and picture may be accessed at different points in time makes it difficult to compare the conceptual masking effects induced by subsequently-presented sounds versus

subsequently-presented pictures. Given this concern, an additional experiment with a shorter SOA between the target picture and the sound (147 ms) was conducted. However, the results still demonstrated that only the facilitatory effect in the congruent condition was observed, but not the interference effect in the incongruent condition[31]. This result therefore decreases the likelihood that access to the meaning of the sound might simply have occurred outside of the time window in which conceptual masking effects can be demonstrated.

A second possibility here is that the pictures used by Intraub (1984) and Loftus and Ginn (1984) were more complicated (e.g., they included more color and detail) than the black-and-white line-drawings used in the present study. If it is true that complex pictures require more time to be consolidated, then the conceptual masking effect would be expected to be more pronounced for complex pictures than for simple pictures. A final possibility here is that intramodal competition is not only stronger than crossmodal competition at the early perceptual level of information processing, but also at the later semantic level. It is likely that the incoming information will overwrite any representations that have been presented earlier within the same modality in order

[31] An additional group of 16 naïve volunteers (three males, mean age of 18 years) took part in this experiment. The only difference from Experiment 8.3 was that the SOA between the picture and the sound was reduced to 147 ms. The results revealed a significant main effect of sound condition ($F_1(2,30) = 9.31$, $MSE_1 = 0.022$, $p_1 < .001$, $\eta^2_1 = 0.38$; $F_2(2,30) = 7.35$, $MSE_2 = 0.021$, $p_2 < .005$, $\eta^2_2 = 0.33$). A *post-hoc* Fisher-Hayter test revealed that the adjusted accuracy of participants' performance was significantly higher in the congruent condition than in the incongruent condition ($ps < .01$; $q(30,2).01 = 3.89$, $MSE_1 = 0.022$; $MSE_2 = 0.021$) and the noise condition ($p_1 < .01$; $p_2 < .05$, $q(30,2).05 = 2.89$). The difference between the accuracy of participants' performance in the incongruent and noise conditions did not reach statistical significance.

to update for the latest changes taking place in the environment, whilst it may only serve to provide supplementary information to the old representations in another modality. However, this account depends *a priori* on the modality source of the sensory input still being represented at the semantic level of information processing (e.g., Barsalou, 1999; McCarthy & Warrington, 1988; Stanfield & Zwaan, 2001; though see Caramazza & Mahon, 2003; Pylyshyn, 1984).

It seemed worth further testing whether the audiovisual semantic congruency effect reported in the three experiments described so far in this chapter would disappear if the SOA between picture and sound were to be increased still further. Addressing this question should help to rule out the possibility that the facilitatory effect simply represents some form of decisional-level facilitation effect. That is, the presentation of the semantically-congruent sound might simply have served to reduce any uncertainty that the participants may have had when trying to identify the picture. For example, if participants only identified the picture as a four-legged animal, say, and then heard a barking dog, they may have entered 'dog' as their answer. In order to rule out this uncertainty-reduction hypothesis as a possible explanation for the findings reported in this chapter, the SOA was prolonged to 533 ms in the next experiment. According to the uncertainty-reduction hypothesis, semantically-congruent sounds should always facilitate participants' performance providing that they appear prior to the decision of a participant's response (i.e., regardless of the SOA). On the other hand,

if the sound arrived after the semantic representation of the picture has either been consolidated, or else had decayed (see Potter, 1999), no such crossmodal facilitation should result from the presentation of a semantically consistent sound.

8.5 Experiment 8.4

In this experiment, the onset of the sound was delayed by 533 ms with respect to that of the picture. The prediction was that the semantic congruency of the sound would no longer have any influence on participants' picture identification performance.

8.5.1 Methods

8.5.1.1 Participants

Sixteen naïve volunteers (three males, mean age of 19 years) from the same pool of volunteers took part in this experiment.

8.5.1.2 Stimuli, apparatus, design, and procedure

Three levels of the sound condition factor (congruent, incongruent, and noise) were manipulated on a within-participants basis. The sound was presented 506 ms after the onset of the pattern mask (i.e., 533 ms after the onset of the picture, see Figure 8.1B). All of the other details were the same as in Experiment 8.3.

8.5.2 Results and discussion

The adjusted accuracy of participants' picture identification performance is shown

in Table 8.2. There was no main effect of sound condition ($F_1(2,30) = 2.36$, $MSE_1 = 0.011$, $p_1 = .11$, $\eta^2_1 = 0.14$; $F_2(2,30) = 1.67$, $MSE_2 = 0.010$, $p_2 = .22$, $\eta^2_2 = 0.10$). This result is consistent with the prediction that when the onset of a sound is delayed sufficiently (i.e., by 533 ms in this experiment), the semantic congruency of the sound no longer influences participants' picture identification performance.

The null effect of the semantic congruency manipulation in the present experiment did not simply result from participants initiating their manual responses before they had actually perceived the sound, as evidenced by the following observation: Analysis of the RT (the time to enter the first letter of answer) data recorded for four of the participants revealed that the participants started to respond 434 ms (average of the fastest RT) after the presentation of the sound (mean was 1,924 ms, standard deviation was 1035 ms). Hence, the sound would already have been presented in its entirety by the time that the participants initiated their response. Given that the data still showed that performance in the congruent condition was slightly better than in the incongruent condition, the semantically-congruent sound may have reduced the uncertainty of participants' responses somewhat. However, it should be noted that this form of uncertainty reduction was clearly not the determining factor modulating the facilitation seen in the congruent condition, since the difference from the incongruent condition failed to reach statistical significance.

Before outlining a possible explanation regarding the mechanisms underlying the

audiovisual semantic congruency effect, two further possibilities need to be ruled out. First, it could be argued that participants may have adapted to the constant SOA of 307 ms presented in Experiment 8.3. That is, it has been shown that people can (to a certain extent) recalibrate to asynchronously-presented audiovisual stimuli (i.e., and bind them into a single multisensory event; see Fujisaki, Shimojo, Kashino, & Nishida, 2004; Navarra et al., 2005; Vroomen & Keetels, 2010, for a review). However, the 533 ms SOA used in Experiment 8.4 likely fell well outside of the temporal window for audiovisual recalibration, so the visual and auditory stimuli would not have been bound together (Fujisaki et al., 2004). Second, different participants took part in Experiments 8.1 through 8.4, so it is prudent to examine the time course of the audiovisual semantic congruency effect within the same group of participants. Therefore, in the last experiment in this chapter, the three SOAs (0, 333, and 533 ms) were randomized on a trial-by-trial basis and tested on a within-participants basis in order to replicate the results in Experiment 8.1, 8.3, and 8.4.

8.6 Experiment 8.5

In this experiment, the goal was to attempt to replicate the results of the previous experiments reported in this chapter but now using a within-participants experimental design. The SOAs (0, 333, or 533 ms, see Figure 8.1B) were randomized on a trial-by-trial basis. This randomization procedure would have ensured that the

participants were unable to predict (and, more importantly, to adapt to) the constantly-varying SOA between the picture and the sound.

8.6.1 Methods

8.6.1.1 Participants

Twenty-one naïve volunteers (9 males, mean age of 21 years) from the same pool of volunteers took part in this experiment.

8.6.1.2 Stimuli, apparatus, design, and procedure

Six of the pictures used in previous experiments were excluded from this experiment. They were three pictures with the highest accuracy (bottle, car, and flute, with accuracies of 98%, 92%, and 92%, respectively) and the three with the lowest accuracy (bee, fly, and grasshopper, with accuracies of 4%, 15%, and 15%, respectively) in Experiment 8.1. Hence, a total of 36 pictures were used in this experiment.

Two factors, sound condition (congruent, incongruent, or noise) and SOA (0, 333, or 533 ms) were manipulated. Each sound x SOA condition was tested with each one of the 36 pictures. These trials were divided into three blocks, each having balanced sound conditions and SOAs. In each block, 12 non-object patterns were presented three times accompanied by three different naturalistic sounds on each occasion, while the SOA of each non-object-and-sound pair was counterbalanced across blocks. There

were 144 trials in each block and the order of presentation of these trials was randomized within each block. The order of presentation of these three blocks was counterbalanced across participants.

The procedure was modified in order to be sure that the sound had been presented in its entirety before the participants made their response. A visual cue 'Answer' was presented on the computer monitor 998 ms after the offset of target picture. The participants were instructed to withhold their response until this cue had been presented. They were warned that their response would not be recorded if they did not follow this instruction.

8.6.2 Results

Three participants were excluded from the data analysis: Two of these participant responded at above 95% correct, indicating that the task may have been too easy for them, while the other participant did not wait until the cue to respond, and so a number of his/her responses were not recorded. Eighteen participants therefore remained for further analysis (see Table 8.3 for the results). There was a main effect of sound condition ($F_1(2,34) = 23.90$, $MSE_1 = 0.024$, $p_1 < .0001$, $\eta^2_1 = 0.58$; $F_2(2,34) = 20.73$, $MSE_2 = 0.020$, $p_2 < .0001$, $\eta^2_2 = 0.55$). A *post-hoc* Fisher-Hayter test showed that the adjusted accuracy of participants' performance was significantly higher in the congruent condition than in either the incongruent sound or noise conditions (all *p*s

< .01; q(34,2).01 = 3.86, MSE_1 = 0.024; MSE_2 = 0.020). There was also a significant interaction between sound condition and SOA (F_1(4,68) = 2.90, MSE_1 = 0.023, p_1 < .05, η^2_1 = 0.15; F_2(4,68) = 2.80, MSE_2 = 0.024, p_2 < .05, η^2_2 = 0.14). The simple main effect showed that the sound condition factor had a significant effect on performance at the 0 ms SOA (F_1(2,102) = 15.61, MSE_1 = 0.023, p_1 < .0001, η^2_1 = 0.19; F_2(2,102) = 12.68, MSE_2 = 0.023, p_2 < .0001, η^2_2 = 0.17), and at the 333 ms SOA (F_1(2,102) = 12.15, p_1 < .0001, η^2_1 = 0.15; F_2(2,102) = 10.07, p_2 < .0005, η^2_2 = 0.13), but crucially not at the 533 ms SOA (F_1(2,102) = 2.57, p_1 = .08, η^2_1 = 0.03; F_2(2,102) = 1.57, p_2 = .23, η^2_2 = 0.02).

At the 0 ms SOA, a *post-hoc* Fisher-Hayter test revealed that the adjusted accuracy of participants' performance was significantly better in the congruent condition than in either the incongruent or noise conditions (all ps < .01; q(102,2).01 = 3.72, MSE_1 = 0.023; MSE_2 = 0.023), and the accuracy of participants' performance in the noise condition was significantly better than in the incongruent condition (both ps < .05; q(102,2).05 = 2.81). At the 333 ms SOA, a *post-hoc* Fisher-Hayter test showed that the adjusted accuracy of participants' performance was significantly better in the congruent condition than in either the incongruent or noise conditions (all ps < .01), while the difference between accuracy levels reported in the noise and incongruent conditions did not reach statistical significance (see Figure 8.3).

Table 8.3. Mean accuracy (%) and standard error (in parentheses) of participants' responses in Experiments 8.5.

SOA (ms)	Sound Condition						
	Congruent					Incongruent	Noise
	% Correct	FA1	Adjusted 1	FA2	Adjusted 2		
0	79.3 (3.5)	2.3 (0.8)	78.8 (3.5)	6.5 (2.0)	77.7 (3.6)	67.4 (4.3)	71.8 (3.6)
333	80.6 (2.7)	1.4 (0.7)	80.3 (2.7)	4.8 (1.4)	79.5 (2.8)	70.5 (3.7)	71.6 (4.1)
533	76.7 (3.0)	0.5 (0.3)	76.6 (3.0)	4.2 (1.4)	75.4 (3.2)	73.6 (3.6)	71.5 (3.6)

FA1: The false alarm (FA) rate was estimated from the frequency with which the participants entered the name of the object indicated by the sound as the visual target in the incongruent condition; Adjusted 1: The accuracy of congruent condition was adjusted by FA1.

FA2: FA rate was estimated from the frequency with which the participants entered the name of the object indicated by the sound as the visual target in the non-object condition; Adjusted 2: The accuracy of congruent condition was adjusted by FA2.

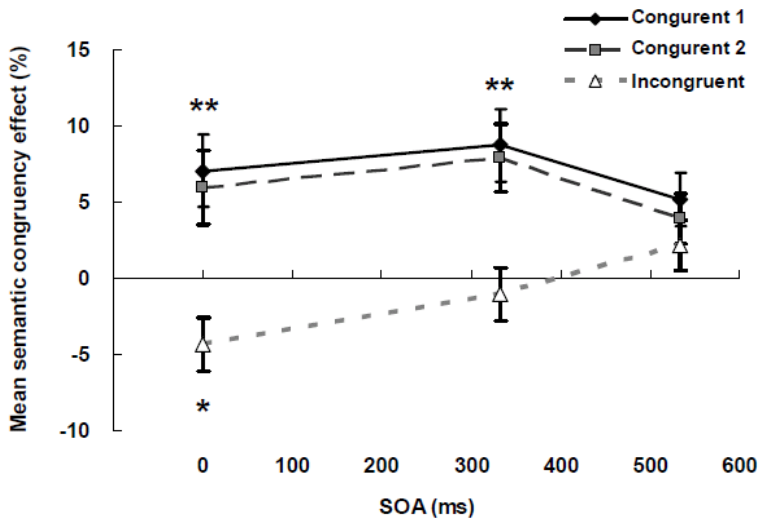


Figure 8.3. Graph highlighting the semantic congruency effects (and ±1 standard error of means) reported in Experiment 8.5. The magnitude of the effect was calculated by subtracting the accuracy of participants' performance in the semantically congruent condition (or semantically incongruent condition) from that in the noise condition. Congruent 1 and 2 reflect the accuracy adjusted from FA1 and FA2, respectively (see Table 8.2). Positive values indicate the improvement in the semantically congruent condition, whereas negative values indicate the impairment in the semantically incongruent condition relative to the noise condition. * denotes a significant difference (*: $p < .05$; **: $p < .01$).

8.6.3 Discussion

The results of Experiment 8.5 therefore replicated those of Experiments 8.1, 8.3, and 8.4. When the onset of the sound coincided with that of the picture, the presentation of a semantically-congruent sound facilitated participants' performance whereas the presentation of a semantically-incongruent sound impaired performance (as compared to that seen in the noise condition) on the picture identification task. When the onset of the sound was delayed by 333 ms with respect to that of the picture, the crossmodal facilitatory effect of the congruent sound remained intact. However, when the sound was delayed by 533 ms, the semantic congruency of the sound no longer influenced participants' picture identification performance.

Three main conclusions can be drawn from the results of this final experiment: First, the audiovisual semantic congruency effects observed across the two shortest SOAs (0 and 333ms), and the null effect of semantic congruency reported at the longest SOA (533 ms) are robust, no matter whether tested on a between- or within-participants basis. Second, since the manipulation of the SOA between picture and sound was randomized on a trial-by-trial basis in this experiment, the participants would have had little opportunity to adapt (and recalibrate) to the changing intervals between the picture and sound. The difference between the results in the 333 and 533 ms SOAs in the present experiment therefore cannot be attributed to any form of audiovisual temporal recalibration. Third, even though the participants in Experiment

8.5 had to respond about 1000 ms after the offset of the picture, the semantic congruency of the sound nevertheless still influenced their performance at the 0 ms and 333 ms SOAs, but not when the SOA was increased to 533 ms (for a similar result at the 0 ms SOA condition, see Yuval-Greenberg & Deouell, 2009). This result therefore suggests that the representation of the picture was undergoing some form of dynamic processing across these three intervals, and that the meaning of the sound only influenced the processing of the picture within a certain time window (which was smaller than 533 ms).

8.7 General discussion

The results of the five experiments reported in this chapter demonstrate that the semantic congruency of a sound can influence participants' identification of a masked line-drawing picture. Specifically, when the picture and the sound were presented at the same time, participants' picture identification performance was facilitated by the presentation of the semantically-congruent sound, as compared to when the picture was presented together with a burst of white noise, or in the absence of any sound (see Experiments 8.1, 8.2, and 8.5). On the other hand, performance was impaired when a semantically-incongruent sound was presented at the same time as the picture, as compared to when a burst of white noise was presented instead (see Experiments 8.1 and 8.5). When the onset of the sound was delayed by a little more than 300 ms with

respect to that of the picture, facilitation by the semantically-congruent sound was still observed (see Experiments 8.3 and 8.5). However, there was no longer any evidence that audiovisual semantic congruency effects influenced performance once the sound had been delayed by 533 ms (see Experiments 8.4 and 8.5).

These results therefore provide the first empirical evidence that audiovisual semantic congruency effects can influence participants' performance on a masked visual picture identification task. Given that the *accuracy* of participants' unspeeded responses was adjusted to take account of their estimated guessing bias, this index should closely approximate the true sensitivity of participants' picture identification abilities. It is easier to account for the congruency effects in terms of a genuine perceptual effect (rather than in terms of response redundancy; see Suied et al., 2009) given the use of a sensitivity measure rather than the RT measure used in several previous studies (see Davenport & Potter, 2004; McDonald et al., 2000; Prinzmetal et al., 2005; Santee & Egeth, 1982, on this issue). The audiovisual semantic congruency effect observed in the present study therefore suggests that the abstract semantic representations that are associated with the visual and auditory stimuli are processed by the same semantic system. Furthermore, the novel observation reported here was that not only did the semantically-congruent sounds facilitate participants' picture identification performance (when presented simultaneously), but those that had been delayed as much as 333 ms did so as well. On the other hand, the

semantically-incongruent sound only impaired participants' performance when its onset coincided with that of the picture.

The observation of a crossmodal semantic congruency effect in the experiments reported in the present chapter implies that the semantic systems for different senses should certainly not be considered as being completely independent (see McCarthy & Warrington, 1988). On the other hand, since the modality source of sensory input is suggested to be still represented at the semantic level (see the Discussion of Experiment 8.3), the semantic system cannot be considered as being completely amodal either (e.g., Pylyshyn, 1984). Some form of intermediate representation between these two extreme hypotheses may therefore be more plausible in terms of accounting for the semantic system in humans, such as a analogue system that retains some access to modality information (Barsalou, 1999), a interconnected multisensory network (Shallice, 1988), or else a hierarchical modal-to-amodal system (Plaut, 2002).

It could be argued that the participants in the experiments reported in this chapter may have strategically attended to the identity of the sound, since the auditory stimulus provided potentially useful information with regards to the identity of the visual target. Note that the auditory and visual stimuli were matched on one-third (Experiment 8.1) or one-fourth (Experiments 8.2 to 8.5) of all trials. Thus, the sound correctly indicated the identity of the target picture at a level that is far higher than chance. Consequently, it is hard to assert unequivocally that the meaning of the sound was processed

'automatically' by participants in the present study (cf. Schneider et al., 1984). On the other hand, given that the information from different sensory modalities often helps to identify an object, as in the ball bouncing example discussed in the Introduction (see Section 8.1), the strategy of attending to multisensory inputs is certainly appropriate in daily life (and not just in an experimental setting). It would therefore seem that, even though the auditory input may first undergo some form of gating by selective attention at an early stage of information processing (e.g., Treisman & Riley, 1969), once it has been selected and processed, its meaning may then unavoidably interact with any relevant visual information (see Lavie, 2005; Lavie et al., 2004; Treisman & Davies, 1973; Yuval-Greenberg & Deouell, 2009).

Potter's idea regarding CSTM (see Potter, 1993, 1999) may provide a useful theoretical framework with which to explain the audiovisual interactions in human information processing reported here. Potter suggested that the meaning of a picture is accessed rapidly (within the first 100 ms or so), and the activated semantic representation is then retained in CSTM for a further 300 ms. At the retention stage, the semantic representation of the picture, as well as other relevant semantic information, are then consolidated in order to be encoded into memory (for a review, see Intraub, 1999). The results reported in this chapter therefore suggest that the processing of a visually-presented picture prior to, and in, CSTM can also be influenced by the meaning of any sound that happens to be presented at around the

same time. While a picture is being processed and a sound perceived, access to the meaning of that picture can either be enhanced or interfered with as a function of the meaning of the sound being either congruent or incongruent with it (i.e., at the SOA of 0 ms). When the representation of a picture passes on to the retention stage (i.e., once its meaning has been accessed), the congruent sound can only help with the consolidation of the picture in CSTM (see also Murray et al., 2004). It is important to note that, at the retention stage, those activated semantic representations referring to the input information are dynamically structured in order to achieve the participants' current task goal (Potter, 1999; Potter, Stiefbold, & Moryadas, 1998). That is, when this goal happens to involve identifying a visually-presented picture, as in the experiments reported in this chapter, the presentation of a semantically-congruent sound would therefore provide coherent information, and so enhance the consolidation of the picture representation. On the other hand, the presentation of the semantically-incongruent sound was certainly not consistent with the representation of the target picture (e.g., the incongruent sound always belonged to another category in the experiments reported in this chapter), and hence, being useless to the task at hand, the sound should rapidly have been forgotten (see Potter, 1999). These results therefore not only provide evidence concerning audiovisual interactions taking place at the semantic level but, in turn, suggest that CSTM is *not* a unimodal visual structure.

Potter's (1993, 1999) CSTM construct therefore provides a plausible explanation

for why the representation of a picture would be influenced by a simultaneously- or subsequently-presented (by up to around 300 ms) sound in the semantic domain. That is, the mechanism of CSTM is not constrained by the need for strict temporal coincidence between the interacting stimuli. Indeed, this interval is more than sufficient for participants to correctly judge the order of (or perceive as asynchronous) visual and auditory stimuli (e.g., Stone et al., 2001; Vatakis & Spence, 2008; Zampini et al., 2005). Soto-Faraco and Alsius (2009) recently demonstrated that the McGurk effect (McGurk & MacDonald, 1976) occurs over a range of asynchronies (from sound leading by 230 ms to sound lagging by 480 ms) that is much wider than the temporal window in which the stimuli are perceived as being simultaneous (from sound leading by 70 ms to sound lagging by 234 ms). That is, the McGurk effect also occurs at SOAs at which participants can accurately judge that the heard phonemes and seen lip movements are not synchronized.

It is possible that when the heard syllable is temporarily retained in the phonological loop (in STM), it can still be modulated by the subsequently-arriving seen lips movements (i.e., for those cases where the sound leads). Hence, combining the results of this chapter with those of Soto-Faraco and Alsius's (2009) recent work, it would seem that there may, in fact, be multiple audiovisual interaction mechanisms in different domains, such as the CSTM for semantic representations, and the STM for phonological representations (see also Driver & Noesselt, 2008; Soto-Faraco & Alsius,

2009). These short-term buffers may retain and integrate the information from different sensory modalities in order to construct coherent multisensory representations suitable for further processing. Consequently, these short-term buffers (for more complex stimuli) may serve to widen the temporal window of multisensory integration beyond that seen in previous behavioral studies that have so often used simple light and sound stimuli.

8.7.1 A possible framework for multisensory integration

Previously, researchers have categorized the variables that influence audiovisual interactions into *structural* and *cognitive* factors (Radeau & Bertelson, 1977; Spence, 2007; Welch & Warren, 1980), which are thought to conjointly determine the final percept (see Welch, 1999; Welch & Warren, 1980). Spence (2007) has argued that structural and cognitive factors may not be as easily distinguished as has often been asserted previously. For example, although spatial and temporal coincidence are generally considered to be structural factors, it is important to note that they can also serve as cognitive cues to bind visual and auditory events as well (e.g., Arnold, Johnston, & Nishida, 2005). Furthermore, as the temporal windows in which simultaneity is perceived and in which the semantic congruency of audiovisual stimuli has been shown to influence multisensory integration can be dissociated (this chapter; Soto-Faraco & Alsius, 2009), the underlying mechanisms should therefore not

necessarily be considered as being unified (see also Driver & Noesselt, 2008, on this point). A new classification of factors and framework for crossmodal integration is therefore needed. Thus, the dissociation between spatiotemporal and identity processing may provide two basic interdependent mechanisms capable of subserving multisensory interactions in human information processing (see Section 4.8.3 and Figure 4.10 for the type-token binding framework).

Concerning the physiological basis of multisensory interactions, several subcortical and cortical areas, such as the superior colliculus, the superior temporal sulcus (STS), and the middle temporal gyrus (MTG) have all been shown to be involved in constraining the integration of crossmodal events (see Driver & Noesselt, 2008; Stein & Meredith, 1993). Recent studies have further probed the multisensory interactions taking place in those brain areas traditionally considered to be 'sensory-specific' (e.g., Noesselt et al., 2007; see Schroeder & Foxe, 2005, for a review). By measuring neural activity, researchers have been able to show that multisensory interactions start very soon after stimulus onset (i.e., within 100 ms in certain cases, see Giard & Peronnet, 1999; Murray et al., 2005).

The above-mentioned studies, however, have tended to focus on the spatiotemporal aspects of multisensory interactions. To date, only a limited number of such studies have attempted to investigate crossmodal semantic congruency effects as in behavioral studies (see Doehrmann & Naumer, 2008, for a recent review). For

instance, Molholm et al. (2004) provided ERP data to suggest that crossmodal semantic congruency effects occur around 400 ms after stimulus onset (and are associated with the N400 component; see also Schneider, Debener, Oostenveld, & Engel, 2008). To date, neuroimaging studies have found some candidate brain areas in which audiovisual semantic interaction have been shown to occur. For example, STS and MTG are both involved in audiovisual object integration; however they do not appear to be sensitive to audiovisual semantic congruency (see Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Beauchamp et al., 2004; Hein et al., 2007; Noppeney et al., 2008; Werner & Noppeney, 2010b; though see Belardinelli et al., 2004; Schneider, Debener et al., 2008). A few studies have reported that activation in the inferior frontal cortex (IFC) is more pronounced under conditions of semantically-incongruent stimulus presentation than under conditions of semantically congruent presentation (Belardinelli et al., 2004; Hein et al., 2007; Naumer et al., 2009); however, the activation of IFC may reflect the cognitive control induced by perceiving conflicting information (see Doehrmann & Naumer, 2008).

One of the most plausible brain areas discussed to date that may be involved in the processing of at least certain kinds of audiovisual semantic congruency is the perirhinal cortex. Neural activation in this area has been shown to be sensitive to categories of living versus non-living things (see Murray & Richmond, 2001; Taylor, Moss, Stamatakis, & Tyler, 2006; Tyler et al., 2004), which is a basic organizational

dimension for concepts (Caramazza & Mahon, 2003; McCarthy & Warrington, 1988). Given that this structure receives information from unimodal sensory regions that represent object features (see Taylor et al., 2006), it is worth considering whether it may serve as the afferent site from the 'what' pathways of each sensory modality that facilitates the identification of multisensory objects in the world around us. Furthermore, it is also worth investigating in future research that the functional role of the neural circuit of perirhinal cortex and associated regions; that is, whether the activation of such a connectivity can last for hundreds of milliseconds, which may serve as the neural basis of a buffer such as CSTM (e.g., Gazzaley, Rissman, & D'Esposito, 2004; Postle, 2006).

8.8 Conclusions

In conclusion, the results of the experiments reported in this chapter have demonstrated that the identification of a picture can be modulated by the semantic congruency of any sound that happens to be presented at around the same time. The effect of audiovisual semantic congruency on picture identification performance occurs within a temporal window that extends for at least 300 ms after the presentation of the picture. This audiovisual semantic congruency effect likely depends on a short-term buffer for semantic processing (e.g., CSTM), which enables a person to access the meaning of a picture rapidly and to retain it for a short period of time thereafter.

Chapter 9

General discussion

The experiments reported in this thesis were designed to investigate the conditions under which the presentation of a task-irrelevant sound would modulate visual perception in normal adult human participants, and the possible mechanisms underlying these phenomena. The auditory stimulus in the studies reported here was typically informative in terms of its temporal coincidence (Chapters 3 & 4) and/or semantic congruency (Chapters 5-8) with respect to the visual target that participants were attempting to respond to.

The visual performance of neurologically-normal human participants was estimated using psychophysical methods in Chapter 3 in order to investigate the influence of temporal coincidence on human audiovisual interactions. In Experiment 3.1, the sensitivity of participants' visual detection and discrimination performance were both enhanced by the presentation of a simultaneous sound. However, this crossmodal facilitation effect was only observed at an intermediate noise level. In Chapter 4, the

backward masking paradigm was used: Two alphabetic letters were presented sequentially at the same spatial location at various ISIs. The participants had to report the first letter (i.e., the target) while trying to ignore the second one (i.e., the mask). The presentation of a sound at the same time as the visual target enhanced the participants' identification performance when the ISI was 27-40ms. This ISI range happened to lie just beyond the minimum ISI at which the participants were able to correctly segregate the presentation of two letters as constituting distinct perceptual events (see Experiment 4.1). Besides, the ISI required to demonstrate crossmodal facilitation was different from that required to observe facilitation by making the target brighter instead (Experiment 4.2). When the sound was presented at the same time as the visual mask, the multisensory event did not induce a larger masking effect on the unimodal visual target; instead, a facilitatory effect across ISIs was still observed (Experiment 4.3). Further results showed that the crossmodal facilitation effect elicited by the presentation of a simultaneous sound with the visual target occurred only when the two stimuli reliably onset together, irrespective of whether their co-occurrence with respect to the onset of the fixation was fixed or jittered within a certain range (Experiment 4.4). However, the sound did not need to be spatially coincident with the visual target in order for it to give rise to crossmodal facilitation (Experiment 4.5). A crossmodal type-token binding framework was proposed in order to account for the fact that the multisensory event is less likely to be interfered with by the other unimodal event (see Figure 4.10). The

suggestion put forward here was that crossmodal facilitation may occur at the level of object representation.

In Chapters 5 through 8, the effect of semantic congruency on audiovisual interactions was investigated. The experiments reported in Chapter 5 were designed to use the binocular rivalry paradigm in order to investigate whether a person's visual awareness could be modulated by the semantic context provided by a concurrently-presented auditory soundtrack. The results revealed that the dominance duration of a given percept was longer when it happened to be semantically-congruent with the sound that the participant was listening to than when it was not (see Experiments 5.1-5.3). In addition, the modulatory effect of auditory semantic context was additive with the intramodal modulation resulting either from visual selective attention or from visual luminance contrast in terms of the dominance duration measure (Experiments 5.2 & 5.3). The modulation by auditory semantic context was robust even when the presentation of the sound could not directly bias participants' reporting of the dominance of the visual target (Experiment 5.4).

In Experiment 6.1, crossmodal semantic interactions between visually-presented pictures and both naturalistic sounds and spoken words were assessed using a picture categorization task. Crossmodal semantic congruency effects due to priming were observed when the onset of the sound preceded that of the picture by 240 ms or more. When the sound and picture were presented within about 106 ms of one another, all

meaningful sounds (i.e., no matter whether they were congruent or incongruent) gave rise to crossmodal inhibition, as compared to a burst of white noise.

In Chapter 7, crossmodal semantic interactions between naturalistic sounds (or spoken words) and visually-presented pictures were again assessed while now using a picture detection task in order to clearly separate perceptual from decisional effects. The results of the experiments reported in this chapter revealed that the presentation of a naturalistic sound genuinely enhanced participants' sensitivity to a visually-presented semantically-congruent picture only when the onset of the sound occurred 346 ms before that of the picture (see Experiments 7.1 & 7.3A). On the other hand, when the auditory and visual stimuli were presented simultaneously, the presentation of a naturalistic sound simply reduced participants' response criterion to the visually-presented picture (see Experiments 7.2A & 7.2B). When the onset of the sound led the onset of the picture by 346 ms, only naturalistic sounds, but not spoken words, enhanced the sensitivity of participants' responses in the picture detection task (see Experiments 7.3A, 7.3B, & 7.4A). However, when the participants had to detect and identify the visual pictures, both naturalistic sounds and spoken words elicited similar semantic priming effects (Experiment 7.4B). A crossmodal version of the general reading-naming interference model (Glaser & Glaser, 1989) was proposed in order to explain the interactions taking place between visual pictures and verbal and nonverbal sounds.

The results of the five experiments reported in Chapter 8 demonstrated that the semantic congruency of a sound can influence participants' identification of a masked picture. Specifically, when the picture and the sound were presented at the same time, participants' picture identification performance was modulated by the presentation of the semantically-congruent or -incongruent sound, as compared to a burst of white noise (see Experiments 8.1 and 8.5). When the onset of the sound was delayed by a little more than 300 ms with respect to the onset of the picture, crossmodal facilitation by the semantically-congruent sound was still observed (see Experiments 8.3 and 8.5). However, it was no longer observed once the sound had been delayed by 533 ms (see Experiments 8.4 and 8.5). These results were explained in terms of a short-term buffer for semantic processing (e.g., CSTM) which may enable a person to access the meaning of a picture rapidly and to retain it for a short period of time thereafter.

9.1 The role of temporal coincidence in audiovisual interactions

The crossmodal facilitation elicited by the presentation of a simultaneous sound on participants' visual target perception was only observed under certain conditions in the experiments reported in this thesis. For example, it was observed when the noise contrast level was just strong enough to interfere with visual processing (Experiment 3.1). Besides, in Chapter 4, crossmodal facilitation was observed when the visual target constituted a distinct event from any other subsequently-presented visual stimulus

(Experiment 4.1), and when the sound was reliably presented together with the visual target (Experiments 4.4A & 4.4B). In sum, these results therefore suggest that temporal coincidence provides a necessary, but by no means a sufficient, condition for an auditory stimulus to induce crossmodal facilitatory effects. Here, two further conditions, the unity assumption regarding the common origin of visual and auditory stimuli, and the role of spatial coincidence, are discussed.

9.1.1 Temporal coincidence and the unity assumption

The results of Experiment 4.4 are consistent with the findings of Lippert et al.'s (2007) earlier study showing that a sound has to be presented not only simultaneously but also in reliable temporal alignment with a visual target in order for crossmodal facilitation to be observed. Hence, the weak crossmodal facilitatory effects observed in Chapter 3 may be partly attributable to the fact that only one of the two sounds presented in each sound-present trial co-occurred with the visual target. The necessity for the the visual and auditory stimuli to reliably co-occur together in order for robust crossmodal effects to be observed can be linked to the notion of the *unity assumption*. This is the term used to describe an observer's belief with regards to whether the stimuli presented in different sensory modalities come from the same or different events (Spence, 2007; Vatakis & Spence, 2007; Welch & Warren, 1980). Therefore, even though the crossmodal facilitation attributable to the temporal coincidence of stimuli

was formerly considered to be a bottom-up effect, this factor should also induce a certain degree of top-down modulation (see also Lippert et al., 2007).

The unity assumption for crossmodal sensory inputs can be modeled in terms of a *prior* according to the Bayesian modeling approach to multisensory perception (e.g., Beierholm, Quartz, & Shams, 2009; Ernst, 2007). The prior is defined as '*A probability distribution, p(s), representing the prior knowledge or expectation about the hidden variable, s*' (Shams & Beierholm, 2010, p. 425; see also Beierholm et al., 2009), where *p(s)* can be referred to the probability regarding the co-occurrence (or same origin) of a particular visual and an auditory stimulus (i.e., the *s*). That is, a high prior value (*p(s)*) indicates that the observer hold a high unity assumption so he/she would tend to represent two stimuli as a unitary event. An observer can establish a prior for a crossmodal stimulus pair by being explicitly taught (such as when given a specific instruction prior to encountering a certain given situation, see Chapter 4; Alais & Burr, 2004; Sugita & Suzuki, 2003) or rapidly learned based on the statistical properties of the stimuli that happened to be presented (e.g., Ernst, 2007; Lippert et al., 2007). According to recent modeling studies, the prior plays a critical role in multisensory integration (for reviews, see Ma & Pouget, 2008; Shams & Beierholm, 2010).

9.1.2 The role of spatial coincidence in the temporal coincidence situation

The results of Experiment 4.5 revealed that the presentation of both

spatially-consistent and inconsistent sounds facilitated participants' target letter identification performance as compared to the sound absent condition. The only difference was that the facilitation induced by the spatially-consistent sound occurred at a shorter ISI (i.e., the visual stimuli were changing faster) than the facilitation induced by the presentation of the spatially-inconsistent sound. This result may be attributed to the spatially-consistent sound directing the participants' attention to the visual target (see Section 4.8.2; see also Santangelo et al., 2007; Talsma et al., 2010). Nevertheless, it is unclear whether attention helped the *integration* of the visual target and the sound, or else helped the participant to *segregate* visual target and the following visual mask from each other. The former can be considered as a case of perceptual grouping occurring crossmodally for simultaneously-presented visual and auditory stimuli, while the latter, a case of perceptual grouping (or segregation) occurring for unimodal visual events (see Spence & Chen, in press).

9.1.3 The mechanism(s) underlying crossmodal facilitation following the presentation of a temporally coincident sound

Considering the five possible mechanisms proposed in Section 2.1.2 (i.e., alerting, signal enhancement, temporal uncertainty reduction, stronger object representation, and temporal attention capture), the former three hypotheses were not supported by the crossmodal facilitation effects reported in Chapters 3 and 4 (see Table 9.1 for a

summary).

The results of the experiments reported in Chapter 4 were explained in terms of the concurrent presentation of multisensory information giving rise to the construction of a more consolidated object representation than a unimodal one. So the argument put forward in this thesis goes, this multisensory object representation is therefore less likely to be interfered with by any subsequent visual event (i.e., the mask). Besides, the

Table 9.1. Summary of the results of the experiments reported in Chapters 3 and 4 regarding the five hypotheses concerning the crossmodal facilitation of visual perception by the presentation of a simultaneous sound.

Hypothesis	Experiment	Results
Alerting	3.1 & 4.1	***Rejected*** because of the results demonstrating crossmodal facilitation under only specific experimental conditions, rather than it being a general effect.
Signal enhancement	3.1	Perhaps correct, but one needs to consider not only signal enhancement but also *noise reduction* when considering this explanation.
	4.2 & 4.3	***Rejected*** because of results showing that the ISI between two visual stimuli required in order for crossmodal facilitation to occur was dissociated from the facilitation elicited by making the visual target brighter. Besides, the sound presented with the visual mask did not induce a larger visual masking effect.
Temporal uncertainty reduction	3.1	***Rejected*** because of results showing that the presentation of a simultaneous sound simply reduced the threshold for detecting and discriminating the visual signal rather than systematically modulating the slope of the psychometric function.
Stronger object representation	4.1	***Supported*** by results showing that crossmodal facilitation only occurred when the visual target could be temporally segregated from the visual mask.
Attentional capture	4.5	Perhaps correct, since the presentation of a simultaneous sound was shown to facilitate visual target perception at a shorter ISI when the sound was spatially coincident than when it was presented from the opposite side.

sound may also capture the participants' attention and hence modulate their visual perception.

The role of attention in the crossmodal facilitation reported in the experiments in Chapter 4 should be considered carefully. As discussed in Section 4.8.4, the attentional modulation elicited by the sound, if any, was more likely to result from a feedback enhancement on the visual target representation (Di Lollo et al., 2000). Besides, even though the way in which attention was captured by the sound may be considered as purely bottom-up, attention is also likely directed as a consequence of top-down factors (see Talsma et al., 2010). Remember that crossmodal facilitation was only observed when the auditory stimulus and the visual target co-occurred reliably (see Experiment 4.4), the prior relating to audiovisual stimulus concurrency may therefore play an important role in all crossmodal interactions, including those concerning mechanisms of crossmodal attention. Hence, it is plausible that the sound can direct a participant's attention to the concurrent visual target only when it becomes a part of the attentional setting for the target (this is the contingent-orienting hypothesis of attentional capture, see Becker, Folk, & Remington, in press; Folk et al., 1992). In this case, as a result, the crossmodal attention mechanism also consists of a top-down component (see Talsma et al., 2010).

9.2 The role of semantic congruency in audiovisual interactions

In Chapters 5 through 8, audiovisual semantic congruency effects were investigated. The meaning of a sound can modulate visual perception when the visual system encounters an ambiguous situation (such as the binocular rivalry utilized in the experiments reported in Chapter 5). Besides, a meaningful sound leading the onset of a semantically-congruent picture by more than 240 ms can prime participants, resulting in better picture categorization, detection, and identification performance (see the results of the experiments reported in Chapters 6 and 7). Finally, the presentation of a semantically-congruent sound that lags the onset of the picture by 300 ms can apparently still help to consolidate the representation of a picture in visual short-term memory (see Chapter 8).

9.2.1 Audiovisual semantic interactions are asymmetrical

The results of the experiments reported in Chapters 6 and 7 revealed that the onset of the meaningful sounds had to precede the onset of the picture in order to induce a semantic priming effect. Otherwise, they only influenced the participants' decisional processes regarding the picture. This is due to the fact that the time required for a sound to access its meaning is greater than that required for a picture. Yuval-Greenberg and Deouell (2009) recently demonstrated that the semantic modulation of vision by audition was larger than vice versa when a visual picture was presented at a high

contrast, while the difference was reduced when it was presented at a lower contrast instead. That said, the direction of audiovisual semantic interactions (i.e., whether from vision to audition or vice versa) depends on how reliable the semantic information provided in each sensory modality is (see also Yuval-Greenberg & Deouell, 2009). Hence, audiovisual semantic interactions are likely asymmetrical and the direction of influence should be determined by which modality provides semantic information more precisely (or accurately) to the other modality. This idea is consistent with the *modality preciseness view*.

9.2.2 Audiovisual semantic interactions and Bayesian priors

The audiovisual semantic effects reported in the experiments described in Chapters 5 through 8 may be different from the outcome of multisensory integration (or the unity effect, see Vatakis & Spence, 2007, 2008), especially when the SOA between the visual and auditory stimuli is longer than the time window of audiovisual simultaneity (about ±100 ms, see Vatakis & Spence, 2010). Nevertheless, the crossmodal effects resulting from audiovisual semantic interactions indicate that the semantically-congruent picture and sound are associated in the semantic system. Once again, these crossmodal semantic associations can be considered as the prior term in a Bayesian model. The associations between different pairs of visual and auditory stimuli should have different strengths due to their frequency of co-occurrence. This is called the '*coupling prior*'

(see Ernst, 2007, p. 4). Hence, when testing whether the temporal window for audiovisual integration can be modulated by semantic congruency between visual and auditory stimuli (e.g., Vatakis et al., 2008; Vatakis & Spence, 2007, 2008), the strength of the association (or coupling prior) between the stimuli that happened to be used should be a critical factor.[32]

9.2.3 The mechanism(s) underlying crossmodal facilitation by semantic congruency

A crossmodal framework adopted from the general reading-naming interference model put forward by Glaser and Glaser (1989) was proposed in order to account for the interactions taking place between visual pictures and both naturalistic sounds and spoken words (see Section 7.6.2 and Figure 7.7). According to this framework, a naturalistic sound and a visual picture (both of which constitute non-verbal stimuli) directly access their corresponding semantic representations (e.g., Dick et al., 2002; Glaser & Glaser, 1989). On the other hand, a spoken word (i.e., a verbal stimulus) has to access its associated semantic representation via its lexical representation (Coltheart, 2000).

This crossmodal framework can easily explain the semantic priming effect reported

[32] Note that the prior in a particular experimental setting may contain multiple components (see Beierholm et al., 2009, p. 7). For example, in Vatakis and her colleagues' TOJ studies (Vatakis et al., 2008; Vatakis & Spence, 2007, 2008), they manipulated the congruency and the SOAs between pairs of visual and auditory stimuli. Their assumption was that the temporal window of audiovisual integration (i.e., the range of SOAs over which the visual and auditory stimuli would be perceived as referring to the same event) can be modulated by the congruency of the stimuli. In this case, the prior at least consisted of two components- the congruency and temporal coincidence between the stimuli (see Ernst, 2007). As a result, the perceptual outcome of audiovisual integration for the participants should follow the stronger prior.

in Chapters 6 and 7. Furthermore, when the representation of the picture is briefly maintained in a short-term buffer, the meaning of a later-presented semantically-congruent sound may be able to help prevent the visual representation from decaying too rapidly (see the experiments reported in Chapter 8). That said, audiovisual semantic interactions can occur at different stages of visual information processing, such as semantic access and memory retention, as reported in the experiments in Chapters 6 to 8. Therefore, crossmodal semantic interactions constitute a heterogeneous group of empirical phenomena.

One may wonder whether the crossmodal facilitation of visual perception by a meaningful sound could have been mediated by some form of visual imagery. That is, the participants might have started to imagine a dog after hearing the sound of a dog barking (or the verbal sound /dɔg/). However, it is unlikely that this hypothesis can explain the results reported in Chapters 5 to 8. Previous studies have demonstrated that perceptual sensitivity to a visual target is *reduced* when a participant generates a visual mental image regarding the presentation of that stimulus at the same time and/or the same spatial location (Ishai & Sagi, 1995; Segal & Fusella, 1969, 1970). As discussed in Section 5.6.1, there was no evidence to show that the auditory semantic modulation observed in the binocular rivalry situation could be explained by using mental imagery. Besides, in Chapters 6 to 8, the SOA between the sound and the visually-presented picture was less than 400 ms, so the participants would have had to generate a mental

image very rapidly after they realized the meaning of the sound.[33] That said, it seems unlikely that the strategy of imagining the object referred to by the sound can explain the audiovisual semantic interactions reported here. However, recent neuroimaging studies, together with psychological theory in terms of the grounded cognition view of perception and mental imagery do suggest that audiovisual semantic interactions and mental imagery may actually enhance the activity of a common group of neurons representing a given visual or auditory object (see below).

Meyer et al. (2010), for example, recently demonstrated that auditory cortex was activated when their participants watched a silent video (e.g., dog barking). What is more, the activation pattern was different from that elicited by watching a musical instrument being played silently. That said, the activity seen in auditory cortex following the presentation of a silent video corresponded to the pattern relevant to a specific category of visual stimuli. It is, though, unclear whether such activities in auditory cortex are simply induced by the audiovisual interaction network (e.g., Driver & Noesselt, 2008) or are mediated by auditory mental imagery instead.[34] In a similar

[33] Fallgatter, Mueller, and Strik (1997) reported that the mental image elicited by written instruction can be observed by the index of P300 component (i.e., 250-450 ms after stimulus onset) in an ERP study. However, the mental image elicited by auditory instruction likely requires longer since it will take longer to identify the sound itself than for visual stimuli (e.g., Ballas, 1993).

[34] fMRI studies of mental imagery have demonstrated that the brain regions supporting visual mental imagery are similar to those involved in visual perception (for a review, see Kosslyn, Ganis, & Thompson, 2001; Kosslyn & Thompson, 2003). The same is true for auditory mental imagery and auditory perception (Daselaar, Porat, Huijbers, & Pennartz, 2010; for a review, see Zatorre & Halpern, 2005). That said, for the levels of brain networks representing visual or auditory objects, activity elicited by a veridical stimulus (i.e., perception) may be indistinguishable from that elicited by mental imagery (though the activity in low-level brain areas is most likely different, see Daselaar et al., 2010).

vein, the presentation of a meaningful sound may be able to elicit neural activities in visual cortex responsible for object identification, and these activities are likely to be similar to those elicited by veridical visual inputs as well (e.g., Daselaar, Porat, Huijbers, & Pennartz, 2010).

In terms of the grounded cognition view, perception, as well as other cognitive functions (such as working memory and mental imagery), share common neural representation systems, called *perceptual symbol systems* (*PSS*, see Barsalou, 1999, 2008). These perceptual symbols are used to simulate a subset of perceptual states. They are dynamic, componential, and multisensory. For example, simulating a chirping bird may include the combination of visual symbols representing the visual features of the bird and auditory symbols representing the bird's chirping. Such a simulation of a chirping bird is used to perceive a bird, to imagine a bird, or to retrieve knowledge relevant to a bird (see Barsalou, 1999, 2008). According to PSS, either the priming elicited by a semantically-congruent auditory stimulus, or mental imagery, can give rise to the activation of the perceptual symbol corresponding to that visual stimulus. Besides, the crossmodal facilitation of the maintenance of a visual representation in short-term memory can be explained by the notion that the perceptual symbol regarding the visual object continued simulating for a longer duration as a result of the presentation of the semantically-congruent sound.

9.3 The type-token binding framework

In their early research, Radeau and Bertelson (1977, p. 137) categorized the factors modulating multisensory interactions into structural and cognitive factors: '*By structural factors, we mean abstract properties of the sensory input like those cited in the Gestalt rules of figure segregation - proximity, similarity, common fate, etc.- which can be applied even to completely unfamiliar constellations of data. We use the term "cognitive" to designate features of the situation which communicate knowledge about the distal situation from which the data originate, and so indicate whether a unitary interpretation is correct or not.*' That said, according to Radeau and Bertelson's original idea, 'structural factors' actually contain certain heuristics to help organize the complex stimuli in the outside world; namely, they represent a kind of knowledge established based on previous experience. Hence, Radeau and Bertelson themselves suggested that: '*The structural-cognitive distinction is of course not always clear-cut, but it was felt to have sufficient heuristic value for an exploratory study.*' (p. 138; see also Spence, 2007; Welch & Warren, 1980).

In most cases, spatial and temporal disparities between two stimuli have been categorized in terms of structural factors (see Welch & Warren, 1980). According to Radeau and Bertelson's (1977) original idea, however, spatial and temporal coincidence are not simply properties of the structure of stimuli, but they also implicate heuristics as

to whether multisensory stimuli should be grouped or not (i.e., crossmodal versions of perceptual grouping rules due to spatial proximity and synchrony, see Usher & Donnelly, 1998; Wertheimer, 1950). This notion can be supported by recent modeling studies in terms of the Bayesian causal inference (Shams & Beierholm, 2010). That is, a pair of concurrently-presented visual and auditory (or tactile) stimuli are likely to be integrated not only because of the physical occurrence of spatiotemporal coincidence, but also because of the strong coupling prior elicited by their spatiotemporal coincidence relationship (see Ernst, 2007; Körding, Beierholm, Ma, Quartz, Tenenbaum, & Shams, 2007; see also Chapter 4). The coupling prior constitutes a form of unity assumption, which is therefore a 'cognitive factor' (Welch & Warren, 1980; see Table 1.1). At one level, the separation between structural and cognitive factors is perhaps a valid one for modulatory factors of multisensory interactions. However, this categorization cannot provide a proper separation regarding how structural and cognitive factors modulate the outcome of multisensory interactions in a stimulus-driven and knowledge-based fashion, respectively, as Radeau and Bertelson originally aimed to do.

Here, a crossmodal type-token binding framework was adopted in order to categorize the factors modulating audiovisual interactions and thus provide a possible framework for multisensory interactions (see Figure 9.1). This framework was first proposed in Chapter 4 (see Section 4.8.3 and Figure 4.10) to account for the crossmodal facilitation elicited by the presentation of a concurrent sound. Considering the three

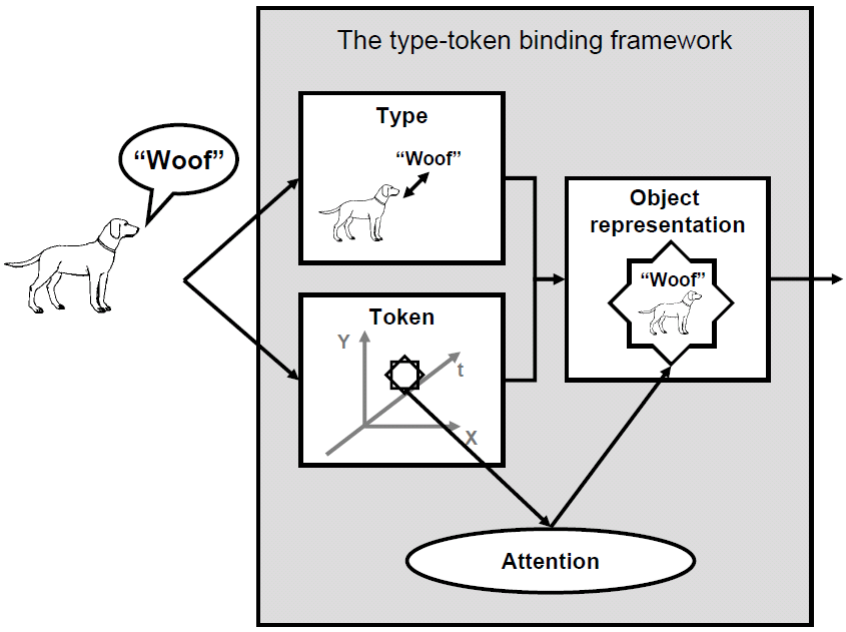


Figure 9.1. Schematic representations of the crossmodal type-token binding framework. Here an example is depicted that is congruent in all its spatial, temporal, and semantic aspects. Vtype (the dog picture) and Atype (an dog's woofing) are integrated in the type representation system. Vtoken (□) and Atoken (◇) are integrated in the token representation system. The type and token information is finally bound together in order to construct a multisensory object representation with the aid of attention (Bowman & Wyble, 2007).

basic properties of perceptual events (space, time, and identity), they can be categorized into spatiotemporal factors ('when' and 'where', or 'token') versus identity factors ('what' or 'type'; see Treisman & Kanwisher, 1998). Such a classification is well-established in cognitive models in the visual modality (Bowman & Wyble, 2007; Driver & Vuilleumier, 2001; Kanwisher, 1987, 2001; Marcel, 1983) and neuropsychological studies (DeYoe & van Essen, 1988; Grossberg, 1994; Ungerleider & Mishkin, 1982). That type and token processing occur in separate neural pathways has also been verified in the auditory modality (e.g., Alain et al., 2001; Kaas & Hackett,

1999; Kubovy & van Valkenburg, 2001; Rauschecker & Tian, 2000; Romanski et al., 1999). The type-token binding framework may therefore be suitable for use in understanding how auditory and visual information interact.

9.3.1 Separable type and token processing in audiovisual interactions

That the binding of types can be dissociated from the binding of tokens is supported by the evidence that human observers can match crossmodal stimuli simply in terms of either their onsets or their identities, and the temporal frequency threshold is higher for the former than for the latter (Fujisaki & Nishida, 2010; see also Experiment 4.1). Crossmodal facilitation in terms of temporal coincidence can be observed irrespective of the stimulus content (see Andersen & Mamassian, 2008; Noesselt et al., 2008). Similarly, multisensory interactions in terms of stimulus type can occur over a range of SOAs within which the stimuli are perceived as being asynchronous (see Chapters 6 to 8; see also Soto-Faraco & Alsius, 2009). Taken together, these results therefore seem to provide evidence that audiovisual interactions in terms of spatiotemporal factors and identity factors can be doubly-dissociated.

Type and token processes in multisensory interactions can be modulated in both a bottom-up and top-down fashion. Note the separation of bottom-up from top-down should be determined according to whether the crossmodal modulation occurred during feedforward or feedback processing (see Section 1.4.3). For example, visual and

auditory inputs that happen to be spatiotemporally coincident can mutually enhance activity in the primary visual and auditory cortices via direct neural connections and/or top-down modulation (Noesselt et al., 2007; Werner & Noppeney, 2010a). Similarly, although the associations between semantically-congruent visual and auditory stimuli are learned as a result of prior experience (e.g., Mitterer & Jesse, 2010), well-learned semantically-congruent audiovisual pairs may subsequently come to interact in feedforward processing (e.g., Noppeney et al., 2008). Hence, information in the type and token pathways, though different in nature, is likely to be processed in a similar way.

9.3.2 The role of attention in multisensory type-token binding

A complete object representation requires type and token to be bound together, a process that presumably needs attention (Bowman & Wyble, 2007; Chun & Potter, 1995; Kanwisher, 1991). In the chapters reported in this thesis, none of the experiments were specifically designed to test whether attention is necessary for audiovisual integration. However, it is worth considering what exactly is meant by the term 'attention' in the context of multisensory interactions.

To say that a process requires *'attention'* normally means that it is effortful, time-consuming, and controllable. This contrasts with an *automatic* processing (cf. Schneider, Dumais, & Shiffrin, 1984). Hence, for example, the AB is a visual deficit

that used to be attributed to the failure of the second target to undergo type-token binding processing because the first target was still though to be utilizing some form of attentional resources (Chun, 1997; Chun & Potter, 1995). However, the latest research now suggests that the AB results from a cognitive strategy, designed to enhance the episodic structure of information encoded into memory. As a consequence, following distractor items have to be inhibited in order to avoid interfering with the preceding target (see Wyble, Bowman, & Nieuwenstein, 2009). This specific top-down modulation provides a more dedicated notion than a general description in terms of an 'attention blink' to our understanding of human cognitive function. In the realm of multisensory interactions, the *unity assumption* (or the *prior*, see Sections 9.1.1 & 9.2.2) serves the cognitive modulation used to enhance the multisensory information being encoded as a unified object (or event). With regards to whether *'attention'* is required to facilitate multisensory integration (e.g., Driver, 1996) may be substituted by the notion of the unity assumption (or prior), or attention itself may be an emergent property of audiovisual integration (e.g., Driver & Vuilleumier, 2001; see Spence, Parise, & Chen, in press).

Certainly, it is hard to assert unequivocally that the influence of the sound in the experiments reported in each of the experimental chapters that constitute this thesis was automatic. For example, the participants were either informed that the sound would accompany the visual target (see the experiments reported in Chapter 4) or they had to

respond to the sound after each trial (see the experiments in Chapter 5). In other chapters, even though the participants were not explicitly informed of the relationship between the sound and the visual target, they may have suspected that the goal of the experiment related to the presentation of accessory sounds, and hence they noticed the sounds. Nevertheless, in the different tasks reported in each chapter in this thesis, attending to task-irrelevant sound was required to differing extents. So, for example, when the task loading was low (such as in the picture categorization task in Chapter 6, only one picture had to be processed), multisensory interactions may have occurred relatively involuntarily. However, when the task loading was high (such as in the binocular rivalry task reported in Chapter 5 where a dichoptic picture had to be processed and continuously monitored), the participants may have had to voluntarily attend to the sound so that they could be influenced by it (see also van Ee et al., 2009). These results therefore suggest only a certain capacity of resources shared by visual and auditory processing.

The most plausible explanation for the interplay between attentional selection and multisensory interactions is that the sound should be selected in at the early stages of human information processing (e.g., Treisman & Riley, 1969). That is, when the loading of the visual task is relatively high, it is more likely that the sound will be ignored (Chapter 5, see also Santangelo et al., 2007). Therefore, the results reported in this thesis do not support the view that the resources required to process information in each

modality are completely independent (see Section 1.4.1 and Figure 1.3). Nevertheless, after the auditory information passes the gate of early selection, it may then inevitably interact with any relevant visual information (see the experimental chapters in this thesis; Yuval-Greenberg & Deouell, 2009). In sum, these results may suggest a two-stage structure, with both levels consisting of limited resources for the processing of information (see Figure 9.2): At the perceptual stage, multisensory interactions occur when the loading of the information in each sensory modality is low. Hence, all of the information can be processed so that it will have the opportunity to interact. At the cognitive control stage, on the other hand, the multisensory interaction would primarily occur when the loading of information is high. This is because sensory inputs from different modalities have been processed to a certain level, and congruent information may fuse in order to preserve cognitive resources (cf. Brand-D'Abrescia & Lavie, 2008; Treisman & Davies, 1973). According to this notion, multisensory interactions may

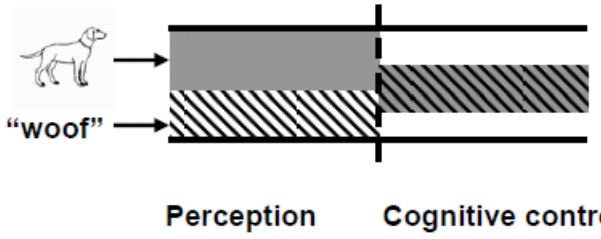


Figure 9.2. A two-stage hypothesis for crossmodal attentional resources. At both the perceptual and cognitive control levels, only limited resources are shared by visual and auditory information processing. At the perceptual stage, both sources of sensory information utilize certain resources. At the cognitive level, congruent information may merge in order to save cognitive resources.

occur not only because of the need to make adequate responses, but also to utilize cognitive resources economically.

9.3.3 Binding type and token information to construct a multisensory object representation

If visual and auditory stimuli are coherent in all spatial, temporal, and semantic regards, they will be integrated in the type and token representations first, and then bound to construct a complete multisensory object representation (see Figure 9.1). Note that a large disparity or incongruency in any one domain may result in the failure of multisensory integration, and therefore the visual and auditory stimuli will be perceived as distinct unimodal events (see Shams & Beierholm, 2010).

According to the PSS model (Barsalou, 1999, 2008), type-token binding is not only used to construct the simulation of a specific object, but it can also occur recursively to locate the spatiotemporal relationships between objects, and finally construct a complete perceptual representation (see Figure 9.3). Given that the type-token binding of an object(s) requires attention, the binding state will be maintained only when the observer pays attention to that object(s) (Barsalou, 1999).

For example, the token representation of the dog and the spatial layout of ‘the bird on the tree’ can first be established individually, and the token of the dog will then be localized to the left side of the tokens of the bird and tree. The perceptual simulation

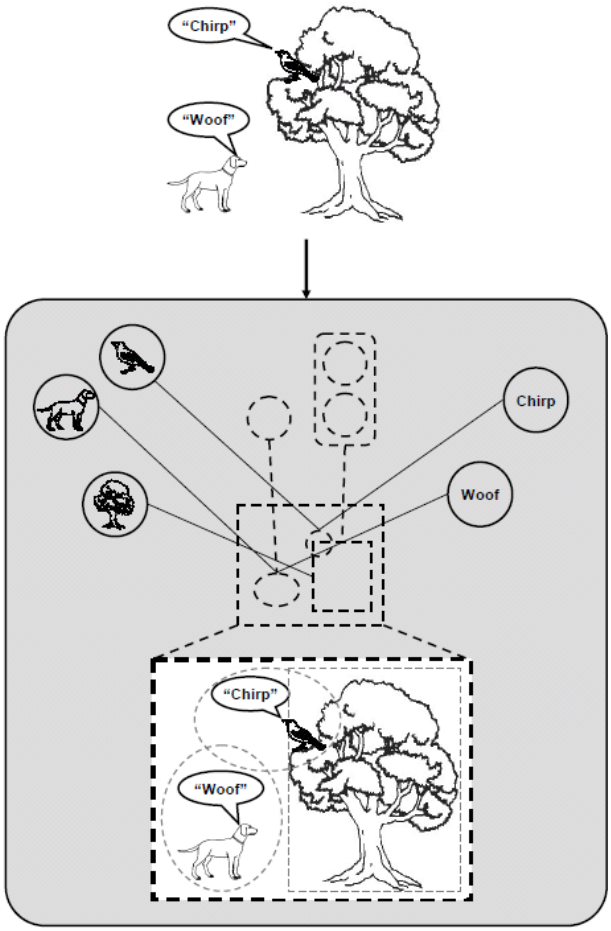


Figure 9.3. The simulation process according to the perceptual symbol system (PSS, Barsalou, 1999). The simulation is based on the type-token binding framework depicted in Figure 9.1. The solid circles represent the type, and the dashed circles and squares represent the token of visual and auditory stimuli. The token representations can recursively bind to specify the spatiotemporal relationships of these objects. After binding all the tokens and types, the complete internal representation should correspond to the outside stimuli. Note that attention is needed in order to maintain this type-token binding (Barsalou, 1999).

will then continue to bind the Vtypes and Atypes (in solid circles) to the tokens (in dashed circles and squares) in order to construct the representations corresponding to the given sensory inputs. The PSS explanation can therefore account for how the multisensory information originating from the same source merges into a single multisensory object representation, and how each object representation can be further

structured in order to correspond to the external stimulus (Barsalou, 1999). The recursion of simulation and tokenization of the sensory input in the PPS therefore provides a framework in which to conceptualize how internal representations (or simulations) are constructed to infer the outside world.

9.4 Directions for future research

An important unsolved issue concerns the interplay between multisensory interaction and attentional capture. As discussed in Section 9.1.3, it is still unclear how attention is directed to the visual target when a sound happens to be presented at the same time. It is likely the sound itself directs attention to the location of visual target (e.g., McDonald et al., 2000; see Spence, McDonald, & Driver, 2004, for a review) or the visual and auditory stimuli have been preliminarily integrated into a salient multisensory event so it can attract attention. The former possibility can be taken as a purely bottom-up crossmodal attentional capture effect; while the latter possibility, since it relies on certain levels of multisensory interactions, is likely to combine both bottom-up and top-down modulations of crossmodal attentional capture (see Talsma et al., 2010). As the long debate regarding whether visual attentional capture reflects a stimulus-driven or top-down modulation (for a recent review see Theeuwes, 2010), attentional capture in a crossmodal situation may face a very similar issue.

To date, behavioural studies have demonstrated that the unity assumption plays a

critical role in both multisensory interactions in terms of spatiotemporal coincidence (e.g., Chapter 4, Lippert et al., 2007) and semantic congruency (e.g., see Chapters 5-8; Mitterer & Jesse, 2010; Vatakis et al., 2008; Vatakis & Spence, 2007). Computational modeling studies adopting Bayesian rules (or causal inference, see Shams & Beierholm, 2010) implement such knowledge-based modulation in terms of the *prior* term. Hence, one direction for future research is perhaps to investigate how a *prior* is established in a specific situation.

Ernst (2007), for example, trained his participants to learn a new correlation between a visual dimension (e.g., luminance) and a tactile dimension (e.g., stiffness). After just one hour of training, the participants had apparently established a weak prior between two crossmodal dimensions. Such a training study may then be used to understand the formation of new priors in the spatial domain (such as in the case of visual prism adaptation modulating auditory spatial perception, see Jacquin-Courtois et al., 2010) or in the temporal domain (such as in the case of audiovisual asynchrony recalibration, see Fujisaki et al., 2004; Navarra et al., 2005; see Vroomen & Keetels, 2010, for a review).

In order to better explain, or further predict, human multisensory perception, future studies will likely need to try to estimate the prior of multisensory stimulus pair under specific experimental situations. For example, in Ma et al.'s (2009) recent study, the visual and auditory stimuli consisted of the video of lip movements and matched spoken

words, respectively, and therefore the prior can be generated from the estimated spoken word frequency (p. 7). According to their modeling results, Ma et al. successfully modeled human behavior showing that the largest crossmodal facilitatory effects for speech perception occurred at an intermediate levels of task difficulty (Ross et al., 2007). However, there is not always a statistical corpus available for the stimulus materials used in a study. Even though the prior can be similar when using the same stimuli and same task while under different levels of task difficulty (Beierholm et al., 2009; Ma et al., 2009), the prior can be changed for different stimuli and tasks (see Beierholm et al., 2009, p. 7). Hence, one of the further challenges comes from the question of how to estimate the participants' prior in order to predict their behavioural performance in a specific experimental situation. Besides, up until now, it is still unclear how the prior is represented neurophysiologically in the human brain network.

9.5 Conclusions

In this thesis, several lines of experiments designed to investigate how the presentation of a sound can modulate visual object perception in terms of the temporal coincidence and semantic congruency of the stimuli have been reported. A crossmodal type-token binding framework was proposed in order to account for how temporal and semantic properties are processed in parallel (i.e., the 'when and where' and the 'what' pathways) and ultimately combined in order to construct multisensory object

representations. In both processing pathways, the *coupling prior* concerning the probability of the crossmodal stimuli resulting from the same object (or event) plays a critical role in determining whether or not crossmodal information should be integrated. This is a critical inference process according to the view holding that perception is indirect (see Section 1.2). Consider the example of the bouncing ball with which this thesis started, combining visual and auditory inputs, we can best perceive the space, time, and force regarding each bounce. In turn, we are able to predict the trajectory of the ball so that we can catch or hit it. In sum, given that the surrounding environment provides information to multiple sensory modalities, understanding how individual perceptual systems communicate in order to correctly represent the outside world should provide a full view of human perception.

References

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., & Grady, C. L. (2001). ‘‘What’’ and ‘‘where’’ in the human auditory system. *Proceedings of the National Academy of Sciences USA, 98,* 12301-12306.

Alais, D., & Blake, R. (1999). Grouping visual features during binocular rivalry. *Vision Research, 39,* 4341-4353.

Alais, D., & Blake, R. (Eds.). (2005). *Binocular rivalry.* Cambridge, MA: MIT Press.

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

Alais, D., Lorenceau, J., Arrighi, R., & Cass, J. (2006). Contour interactions between pairs of Gabors engaged in binocular rivalry reveal a map of the association field. *Vision Research, 46,* 1473-1487.

Alais, D., Morrone, C., & Burr, D. (2006). Separate attentional resources for vision and audition. *Proceedings of the Royal Society B*, *273*, 1339-1345.

Alais, D., Newell, F. N., & Mamassian, P. (2010). Multisensory processing in review: From physiology to behaviour. *Seeing and Perceiving, 23,* 3-38.

Andersen, T. S., Tiippana, K., & Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Cognitive Brain Research, 21,* 301-308.

Anderson, T. S., & Mamassian, P. (2008). Audiovisual integration of stimulus transients. *Vision Research, 48,* 2537-2544.

Aramaki, M., Marie, C., Kronland-Martinet, R., Ystad, S., & Besson, M. (2010). Sound categorization and conceptual priming for nonlinguistic and linguistic sounds. *Journal of Cognitive Neuroscience, 22,* 2555-2569.

Arieh, Y., & Marks, L. E. (2008). Cross-modal interaction between vision and hearing: A speed-accuracy analysis. *Perception & Psychophysics, 70,* 412-421.

Arnell, K. M., & Larson, J. M. (2002). Cross-modality attentional blinks without preparatory task-set switching. *Psychonomic Bulletin & Review, 9,* 497-506.

Arnold, D. H., Johnston, A., & Nishida, S. (2005). Timing sight and sound. *Vision Research, 45,* 1275-1284.

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

Baker, D. H., & Graf, E. W. (2009). On the relation between dichoptic masking and binocular rivalry. *Vision Research, 49,* 451-459.

Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 250-267.

Ballas, J. A., & Howard, J. H., Jr. (1987). Interpreting the language of environmental sounds. *Environment and Behavior, 19,* 91-114.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22,* 577-660.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59,* 617-645.

Bartlett, J. C. (1977). Remembering environmental sounds: The role of verbalization at input. *Memory & Cognition, 5,* 404-414.

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., et al. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review, 10,* 344-380.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A, 20,* 1391-1397.

Bauer, M. (2008). Multisensory integration: A functional role for inter-area synchronization? *Current Biology, 18,* R709-R710.

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

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

Becker, S. I., Folk, C. L., & Remington, R. W. (in press). The role of relational

information in contingent capture. *Journal of Experimental Psychology: Human Perception and Performance.* doi: 10.1037/a0020370

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

Beierholm, U. R., Körding, K. P., Shams, L., & Ma, W.-J. (2008). Comparing Bayesian models of multisensory cue combination without mandatory integration. *Advances in Neural Information Processing Systems. 20,* 81-88.

Beierholm, U. R., Quartz, S. R., & Shams, L. (2009). Bayesian priors are encoded independently from likelihoods in human multisensory perception. *Journal of Vision, 9(5):23,* 1-9.

Belardinelli, M. O., Sestieri, C., Di Matteo, R., Delogu, F., Del Gratta, C., Ferretti, A., et al. (2004). Audio-visual crossmodal interactions in environmental perception: An fMRI investigation. *Cognitive Processing, 5,* 167-174.

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

Bernstein, I. H., Rose, R., & Ashe, V. M. (1970). Energy integration in intersensory facilitation. *Journal of Experimental Psychology, 86,* 196-203.

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

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177,* 77-80.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143-177.

Blake, R. (1988). Dichoptic reading: The role of meaning in binocular rivalry. *Perception & Psychophysics, 44,* 133-141.

Blake, R., & Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience, 3,* 1-11.

Bolognini, N., Frassinetti, F., Serino, A., & Ladavas, E. (2005). "Acoustical vision" of below threshold stimuli: Interaction among spatially converging audiovisual inputs. *Experimental Brain Research, 160,* 273-282.

Bonnel, A.-M., & Hafter, E. R. (1998). Divided attention between simultaneous auditory and visual signals. *Perception & Psychophysics, 60,* 179-190.

Bowman, H., & Wyble, B. (2007). The simultaneous type, serial token model of temporal attention and working memory. *Psychological Review*, *114*, 38-70.

Boyce, S. J., & Pollastek, A. (1992). Identification of objects in scenes: The role of scene background in object naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 531-543.

Boyce, S. J., Pollastek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 556-566.

Brand-D'Abrescia, M., & Lavie, N. (2008). Task coordination between and within sensory modalities: Effects on distraction. *Perception & Psychophysics*, *70*, 508-515.

Breitmeyer, B. G., & Ganz, L. (1976). Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing. *Psychological Review, 83,* 1-36.

Breitmeyer, B. G., & Ogmen, H. (2006). *Visual masking.* Oxford: Oxford University Press.

Breitmeyer, B. G., Kafaligonul, H., Ogmen, H., Mardon, L., Todd, S., & Ziegler, R. (2006). Meta- and paracontrast reveal differences between contour and brightness-processing mechanisms. *Vision Research, 46,* 2645-2658.

Broadbent, D. E. (1958). *Perception and communication.* Elmsford, NJ: Pergamon.

Bruns, P., & Getzmann, S. (2008). Audiovisual influences on the perception of visual apparent motion: Exploring the effect of a single sound. *Acta Psychologica, 129,* 273-283.

Burr, D., Banks, M. O., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research, 198,* 49-57.

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 USA, 102,* 18751-18756.

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,* 649-657.

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

Cappe, C., Rouiller, E. M., & Barone, P. (2009). Multisensory anatomical pathways. *Hearing Research, 258,* 28-36.

Caramazza, A., & Mahon, B. Z. (2003). The organization of conceptual knowledge: The evidence from category-specific semantic deficits. *Trends in Cognitive Sciences, 7,* 354-361.

Carriere, B. N., Royal, D. W., & Wallace, M. T. (2008). Spatial heterogeneity of cortical receptive fields and its impact on multisensory interactions. *Journal of Neurophysiology, 99,* 2357-2368.

Chan, J. S., & Newell, F. N. (2008). Behavioral evidence for task-dependent "what" versus "where" processing within and across modalities. *Perception & Psychophysics, 70,* 36-49.

Chen, K. M., & Yeh, S. L. (2009). Asymmetric cross-modal effects in time perception. *Acta Psychologica, 130,* 225-234.

Chen, Y.-C., & Yeh, S.-L. (2008). Visual events modulated by sound in repetition blindness. *Psychonomic Bulletin & Review, 15,* 404-408.

Chen, Y.-C., & Yeh S.-L. (2009). Catch the moment: Multisensory enhancement of rapid visual events by sound. *Experimental Brain Research, 198,* 209-219.

Choe, C. S., Welch, R. B., Gilford, R. M., & Juola, J. F. (1975). The "ventriloquist effect": Visual dominance or response bias? *Perception & Psychophysics, 18,* 55-60.

Chong, S. C., Tadin, D., & Blake, R. (2005). Endogenous attention prolongs dominance durations in binocular rivalry. *Journal of Vision, 5,* 1004-1012.

Chun, M. M. (1997). Types and tokens in visual processing: A double dissociation between the attentional blink and repetition blindness. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 738-755.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 109-127.

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics, 16*, 409-412.

Coltheart, M. (2000). Dual routes from print to speech and dual routes from print to meaning: Some theoretical issues. In A. Kennedy, R. Radach, J. Pynte, & D. Heller (Eds.), *Reading as a perceptual process* (pp. 475-490). Oxford: Elsevier.

Costello, P., Jiang, Y., Baartman, B., McGlennen, K., & He, S. (2009). Semantic and subword priming during binocular suppression. *Consciousness and Cognition, 18,* 375-382.

Crick, F. (1996). Visual perception: Rivalry and consciousness. *Nature, 379,* 485-486.

Cummings, A., Ceponiene, R., Koyama, A., Saygin, A. P., Townsend, J., & Dick, F. (2006). Auditory semantic networks for words and natural sounds. *Brain Research, 1115,* 92-107.

Dalton, P., & Spence, C. (2007). Attentional capture in serial audiovisual search tasks. *Perception & Psychophysics, 69,* 422-438.

Daselaar, S, M., Porat, Y., Huijbers, W., & Pennartz, C. M. A. (2010). Modality-specific and modality-independent components of the human imagery system. *NeuroImage, 52,* 677-685.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15,* 559-564.

de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7,* 460-467.

De Lucia, M., Clarke, S., Murray, M. M. (2010). A temporal hierarchy for conspecific vocalization discrimination in humans. *Journal of Neuroscience, 30,* 11210-11221.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18,* 193-222.

Deutsch, J. A., & Deutsch, D. (1963) Attention: Some theoretical considerations. *Psychological Review, 70,* 80-90.

DeYoe, E. A., & van Essen, D. C. (1988). Concurrent processing streams in monkey visual cortex. *Trends in Neurosciences, 11,* 214-226

Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal of Experimental Psychology: General, 129,* 481-507.

Dick, F., Bussiere, J., & Saygin, A. P. (2002). The effects of linguistic mediation on the identification of environmental sounds. *Center for Research in Language Newsletter, 14,* 3-9.

Dick, F., Saygin, A. P., Galati, G., Pitzalis, S., Bentrovato, S., D'Amico, S., et al. (2007). What is involved and what is necessary for complex linguistic and nonlinguistic auditory processing: Evidence from functional magnetic resonance imaging and

lesion data. *Journal of Cognitive Neuroscience, 19,* 799-816.

Dijkerman, H. C., & de Haan, E. H. F. (2007). Somatosensory processes subserving perception and action. *Behavioral and Brain Sciences, 30,* 189-239.

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

Doesburg, S. M., Emberson, L. L., Rahi, A., Cameron, D., & Ward L. M. (2008). Asynchrony from synchrony: Long-range gamma-band neural synchrony accompanies perception of audiovisual speech asynchrony. *Experimental Brain Research, 185,* 11-20.

Dosher, B. A., & Lu, Z.-L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences USA, 95,* 13988-13993.

Downar, J., Crawley, A. P., Mikulis, D. J., & Davis, K. D. (2000). A multimodal cortical network for the detection of changes in the sensory environment. *Nature Neuroscience, 3,* 277-283.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature, 381,* 66-68.

Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on "sensory specific" brain regions, neural responses, and judgments. *Neuron, 57,* 11-23.

Driver, J., & Vuilleumier, P. (2001). Perceptual awareness and its loss in unilateral neglect and extinction. *Cognition, 79,* 39-88.

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature*, *387*, 808-810.

Earle, D. C., & Lowe, G. (1974). Channel, temporal, and composite uncertainty in the detection and recognition of auditory and visual signals. *Perception & Psychophysics, 9,* 177-181.

Eltiti, S., Wallace, D., & Fox, E. (2005). Selective target processing: Perceptual load or distractor salience? *Perception & Psychophysics, 67,* 876-885.

Engel, E. (1956). The role of content in binocular resolution. *American Journal of Psychology, 69,* 87-91.

Engel, G. R., & Dougherty, W. G. (1971). Visual-auditory distance constancy. *Nature,*

234, 308.

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences, 4,* 345-352.

Eriksen, C. W. (1966). Temporal luminance summation effects in backward and forward masking. *Perception & Psychophysics, 1*, 87-92.

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

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

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

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

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience, 22,* 5749-5759.

Fallgatter, A. J., Mueller, T. J., & Strik, W. K. (1997). Neurophysiological correlates of mental imagery in different sensory modalities. *International Journal of Psychophysiology, 25,* 145-153.

Farah, M. J. (1988). Is visual imagery really visual? Overlooked evidence from neuropsychology. *Psychological Review*, *95,* 307-317.

Felleman, D. J., & van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1,* 1-47.

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

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

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: The MIT Press.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 1030-1044.

Fowler, C. A. (2004). Speech as a supramodal or amodal phenomenon. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processing* (pp. 189-201). Cambridge, MA: MIT Press.

Frassinetti, F., Pavani, F., & Lavadas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental Brain Research, 147,* 332-343.

Freeman, E., & Driver, J. (2008). Direction of visual apparent motion driven solely by timing of a static sound. *Current Biology, 18,* 1262-1266.

Freides, D. (1974). Human information processing and sensory modality: Cross-modal functions, information complexity, memory, and deficit. *Psychological Bulletin, 81,* 284-310.

Friedman, D., Cycowicz, Y. M., & Dziobek, I. (2003). Cross-form conceptual relations between sounds and words: Effects on the novelty P3. *Cognitive Brain Research, 18,* 58-64.

Fujisaki, W., & Nishida, S. (2005). Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Experimental Brain Research, 166,* 455-464.

Fujisaki, W., & Nishida, S. (2009). Audio-tactile superiority over visuo-tactile and audio-visual combinations in the temporal resolution of synchrony perception. *Experimental Brain Research, 198,* 245-259.

Fujisaki, W., & Nishida, S. (2010). A common perceptual temporal limit of binding synchronous inputs across different sensory attributes and modalities. *Proceedings of The Royal Society B, 277,* 2281-2290.

Fujisaki, W., Kitazawa, S., & Nishida, S. (in press). Multisensory timing. In B. E. Stein (Ed.), *The new handbook of multisensory processing.* Cambridge, MA: MIT Press.

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

Gazzaley, A., Rissman, J., & D'Esposito, M. (2004). Functional connectivity during working memory maintenance. *Cognitive, Affective, & Behavioral Neuroscience, 4,* 580-599.

Gebhard, J. W., & Mowbray, G. H. (1959). On discriminating the rate of visual flicker and auditory flutter. *American Journal of Psychology, 72*, 521-528.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory?

Trends in Cognitive Sciences, 10, 278-285.

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*, 473-490.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Giesbrecht, B., & Di Lollo, V. (1998). Beyond the attentional blink: Visual masking by object substitution. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 1454-1466.

Glaser, W. R., & Dungelhoff, F.-J. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 640-654.

Glaser, W. R., & Glaser, M. O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General, 118,* 13-42.

Gondan, M. (2010). A permutation test for the race model inequality. *Behavior Research Methods, 42,* 23-28.

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

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

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20,* 464-472.

Grossberg, S. (1994). 3-D vision and figure-ground separation by visual cortex. *Perception & Psychophysics, 55,* 48-120.

Hairol, M. I., & Waugh, S. J. (2010). Lateral interactions across space reveal links between processing streams for luminance-modulated and contrast-modulated stimuli. *Vision Research, 50,* 889-903.

Hancock, P. A., Oron-Gilad, T., & Szalma, J. L. (2007). Elaborations of the multiple-resource theory of attention. In A. F. Kramer, D. A. Wiegmann, & A. Kirlik (Eds.), *Attention: From theory to practice* (pp. 45-56). Oxford: Oxford University Press.

Hays, W. (1973). *Statistics for the social sciences.* New York: Holt, Rinehart, &

Winston.

Hecht, D., & Reiner, M. (2009). Sensory dominance in combinations of audio, visual and haptic stimuli. *Experimental Brain Research, 193,* 307-314.

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

Hellige, J. B., Walsh, D. A., Lawrence, V. W., & Prasse, M. (1979). Figural relationship effects and mechanisms of visual masking. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 88-100.

Helmholtz, H. v. (1962). *Helmholtz's treatise on physiological optics* (J. P. C. Southall, Ed.), New York: Dover Publications. (Original work published 1866).

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50,* 243-271.

Heron, J., Whitaker, D., & McGraw, P. V. (2004). Sensory uncertainty governs the extent of audio-visual interaction. *Vision Research, 44,* 2875-2884.

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology, 63,* 289-293.

Hidaka, S., Manaka, Y., Teramoto, W., Sugita, Y., Miyauchi, R., Gyoba, J., et al. (2009). Alternation of sound location induces visual motion perception of a static object. *PLoS ONE 4(12)*: e8188. doi:10.1371/journal.pone.0008188

Hinrichs, H., Heinze, H. J., & Schoenfeld, M. A. (2006). Causal visual interactions as revealed by an information theoretic measure and fMRI. *NeuroImage, 31,* 1051–1060.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition, 108,* 687-701.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127,* 398-415.

Holmes, N. P. (2009). Inverse effectiveness, multisensory integration, and the bodily self: Some statistical considerations. *Consciousness and Cognition, 18,* 762-765.

Holmes, N. P., & Spence, C. (2005). Multisensory integration: Space, time and superadditivity. *Current Biology, 15,* R762-R764.

Howard, I. P., & Templeton, W. B. (1966). *Human spatial orientation*. New York:

Wiley.

Hsiao, J. Y., Chen, Y. C., Spence, C., & Yeh, S. L. (2010). Semantic congruency, attention, and fixation position modulate conscious perception when viewing a bistable figure. *Journal of Vision, 10(7):* 867.

Huddleston, W. E., Lewis, J. W., Phinney, R. E., & DeYoe, E. A. (2008). Auditory and visual attention-based apparent motion share functional parallels. *Perception & Psychophysics, 70,* 1207-1216.

Hughes, H. C., Reuter-Lorenz, P. A., Nozawa, G., & Fendrich, R. (1994). Visual-auditory interactions in sensorimotor processing: Saccades versus manual responses. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 131-153.

Hupé, J. M., Joffo, L. M., & Pressnitzer, D. (2008). Bistability for audiovisual stimuli: Perceptual decision is modality specific. *Journal of Vision, 8(7):1,* 1-15.

Innes-Brown, H., & Crewther, D. (2009). The impact of spatial incongruence on an auditory-visual illusion. *PLoS ONE, 4,* e6450. doi:10.1371/journal.pone.0006450

Intraub, H. (1981). Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 604-610.

Intraub, H. (1984). Conceptual masking: The effects of subsequent visual events on memory for pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 115-125.

Intraub, H. (1999). Understanding and remembering briefly glimpsed pictures: Implications for visual scanning and memory. In V. Coltheart (Ed.), *Fleeting memories: Cognition of brief visual stimuli* (pp. 47-70). Cambridge, MA: MIT Press.

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., & Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics, 72,* 1736-1741.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sound facilitates visual search. *Psychonomic Bulletin & Review*, *15,* 548-554.

Ishai, A., & Sagi, D. (1995). Common mechanisms of visual imagery and perception. *Science, 268,* 1772-1774.

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

Jacquin-Courtois, S., Rode, G., Pavani, F., O'Shea, J., Giard, M. H., Boisson, D., & Rossetti, Y. (2010). Effect of prism adaptation on left dichotic listening deficit in neglect patients: Glasses to hear better? *Brain, 133,* 895-908.

James, W. (1890). Sensation. *Principles of psychology* (Vol. 2, Chapter 17). New York : Henry Holt. http://psychclassics.yorku.ca/James/Principles/index.htm

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics, 72,* 209-225.

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

Kaas, J. H., & Collins, C. E. (2004). The resurrection of multisensory cortex in primates. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processing* (pp. 285-294). Cambridge, MA: MIT Press.

Kaas, J. H., & Hackett, T. A. (1999). 'What' and 'where' processing in auditory cortex. *Nature Neuroscience, 2,* 1045-1047.

Kahneman, D. (1968). Method, findings, and theory in studies of visual masking. *Psychological Bulletin, 70,* 404-425.

Kahneman, D., Treisman, A., & Burkell, J. (1983). The cost of visual filtering. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 510-522.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24,* 174-219.

Kanwisher, N. (1987). Repetition blindness: Type recognition without token individuation. *Cognition, 27,* 117-143.

Kanwisher, N. (2001). Neural events and perceptual awareness. *Cognition, 79,* 89-113.

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

Kelly, D. H. (1974). Spatio-temporal frequency characteristics of color-vision mechanisms. *Journal of the Optical Society of America, 64,* 983-990.

Kiefer, M. (2002). The N400 is modulated by unconsciously perceived masked words: Further evidence for an automatic spreading activation account of N400 priming

effects. *Cognitive Brain Research, 13*, 27-39.

King, A. J. (2005). Multisensory integration: Strategies for synchronization. *Current Biology, 15,* R339-R341.

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

Kingdom, F. A .A., & Prins, N. (2010). *Psychophysics: A practical introduction*. London: Elsevier.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Klemen, J., Büchel, C., & Rose, M. (2009). Perceptual load interacts with stimulus processing across sensory modalities. *European Journal of Neuroscience, 29,* 2426-2434.

Klemen, J., Büchel, C., Buhler, M., Menz, M. M., & Rose, M. (2010). Auditory working memory load impairs visual ventral stream processing: Toward a unified model of attentional load. *Journal of Cognitive Neuroscience, 22,* 437-446.

Knowles, W. B. (1963). Operator loading tasks. *Human Factors, 5,* 151-161.

Kopinska, A., & Harris, L. R. (2004). Spatial representation in body coordinates: Evidence from errors in remembering positions of visual and auditory targets after active eye, head, and body movements. *Canadian Journal of Experimental Psychology, 57,* 23-37.

Koppen, C., Alsius, A., & Spence, C. (2008). Semantic congruency and the Colavita visual dominance effect. *Experimental Brain Research, 184,* 533-546.

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

Kosslyn, S. M., & Thompson, W. L. (2003). When is early visual cortex activated during visual mental imagery? *Psychological Bulletin, 129,* 723-746.

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience, 2,* 635-642.

Kovacs, I., Papathomas, T. V., Yang, M., & Feher, A. (1996). When the brain changes its mind: Interocular grouping during binocular rivalry. *Proceedings of the National Academy of Sciences USA, 93,* 15508-15511.

Kubovy, M., & van Valkenburg, D. (2001). Auditory and visual objects. *Cognition, 80,* 97-126.

La Heij, W., van der Heijden, A. H. C., & Schreuder, R. (1985). Semantic priming and Stroop-like interference in word-naming tasks. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 62-80.

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

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception & Performance, 21,* 451-468.

Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in Cognitive Sciences, 9,* 75-82.

Lavie, N., & de Fockert, J. (2005). The role of working memory in attentional capture. *Psychonomic Bulletin and Review, 12,* 669-674.

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General, 133,* 339-354.

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics, 56,* 183-197.

Leopold, D. A., & Logothetis, N. K. (1996). Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature, 379,* 549-553.

Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences, 3,* 254-264.

Levelt, W. J. M. (1965). *On binocular rivalry*. The Netherlands: Royal VanGorcum.

Lewald, J., Ehrenstein, W. H., & Guski, R. (2001). Spatio-temporal constraints for auditory-visual integration. *Behavioural Brain Research, 121,* 69-79.

Lewald, J., & Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Research, 16,* 468-478.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revisited. *Cognition, 21,* 1-36.

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

Loftus, G. R., & Ginn, M. (1984). Perceptual and conceptual masking of pictures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 435-441.

Logothetis, N. K., Leopold, D. A., & Sheinberg, D. L. (1996). What is rivalling during binocular rivalry? *Nature, 380,* 621-624.

London, I. D. (1954). Research on sensory interaction in the Soviet Union. *Psychological Bulletin, 51,* 531-568.

Lovejoy, L. P., & Krauzlis, R. J. (2010). Inactivation of primate superior colliculus impairs covert selection of signals for perceptual judgments. *Nature Neuroscience, 13,* 261-267.

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,* 447-453.

Lu, Z.-L., & Dosher, B. A. (1998). External noise distinguishes attention mechanisms. *Vision Research, 38,* 1183-1198.

Lu, Z.-L., & Dosher, B. A. (1999). Characterizing human perceptual inefficiencies with equivalent internal noise. *Journal of the Optical Society of America A, 16,* 764-778.

Lumer, E. D., Friston, K. J., & Rees, G. (1998). Neural correlates of perceptual rivalry in the human brain. *Science, 280*, 1930-1934.

Ma, W. J., & Pouget, A. (2008). Linking neurons to behavior in multisensory perception: A computational review. *Brain Research, 1242,* 4-12.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PLoS ONE, 4,* e4638. doi:10.1371/journal.pone.0004638

Macaluso, E., & Driver, J. (2005). Multisensory spatial interactions: A window onto functional integration in the human brain. *Trends in Neurosciences, 28,* 264-271.

Macmillan, N. A., & Creelman, C. D. (2005). Alternative approaches: Threshold models and choice theory. In *Detection theory: A user's guide* (pp. 81-111). Mahwah, NJ: Lawrence Erlbaum Associates.

Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology, 15*, 197-237.

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

Massaro, D. W. (2004). From multisensory integration to talking heads and language learning. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The handbook of multisensory processing* (pp. 153-176). Cambridge, MA: MIT Press.

McCarthy, R. A., & Warrington, E. K., (1988). Evidence for modality-specific meaning systems in the brain. *Nature, 334,* 428-430.

McDonald, J. J., Teder-Sälejärvi, W. A., & Hillyard, S. A. (2000) Involuntary orienting to sound improves visual perception. *Nature, 407,* 906-908.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

McLeod, P., & Posner, M. I. (1984). Privileged loops from percept to act. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X* (pp. 55-66). Hillsdale, NJ: Erlbaum.

Meng, M., & Tong, F. (2004). Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures. *Journal of Vision, 4,* 539-551.

Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science, 221,* 389-391.

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology, 56,* 640-662.

Metzger, W. (1934). Beobachtungen über phänomenale Identität (Sudies of phenomenal identity). *Psychologische Forschung, 19,* 1-60.

Meyer, G. F., Wuerger, S., & Greenlee, M. (in press). Interactions between auditory and visual semantic stimulus classes: Evidence for common processing networks for speech and body actions. Journal of Cognitive Neuroscience. doi:10.1162/jocn.2010.21593

Meyer, K., Kaplan, J. T., Essex, R., Webber, C., Damasio, H., & Damasio, A. (2010). Predicting visual stimuli on the basis of activity in auditory cortices. *Nature Neuroscience, 6,* 667-668.

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

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

Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action.* New York: Oxford University Press.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two central pathways. *Trends in Neurosciences, 6,* 414-417.

Mitchell, J. F., Stoner, G. R., & Reynolds, J. H. (2004). Object-based attention determines dominance in binocular rivalry. *Nature, 429,* 410-413.

Mitterer, H., & Jesse, A. (2010). Correlation versus causation in multisensory perception. *Psychonomic Bulletin & Review, 17,* 329-334.

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

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,* 115-128.

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

Mueller, T. J., & Blake, R. (1989). A fresh look at the temporal dynamics of binocular rivalry. *Biological Cybernetics, 61,* 223-232.

Muller, J. R., Philiastides, M. G., & Newsome, W. T. (2005). Microstimulation of superior colliculus focuses attention without moving the eyes. *Proceedings of the National Academy of Sciences USA, 102,* 524-529.

Murray, E. A., & Richmond, B. J. (2001). Role of perirhinal cortex in object perception, memory, and associations. *Current Opinion in Neurobiology, 11,* 188-193.

Murray, M. M., Camen, C., Gonzalez Andino, S. L., Bovet, P., & Clarke, S. (2006). Rapid brain discrimination of sounds of objects. *Journal of Neuroscience, 26,* 1293-1302.

Murray, M. M., Michel, C. M., Grave de Peralta, R., Ortigue, S., Brunet, D., Gonzalez

Andino, S., & Schnider, A. (2004). Rapid discrimination of visual and multisensory memories revealed by electrical neuroimagine. *NeuroImage, 21,* 125-135.

Murray, M. M., Molholm, S., Michel, C. M., Heslenfeld, D. J., Ritter, W., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2005). Grabbing your ear: Rapid auditory-somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. *Cerebral Cortex, 15,* 963-974.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research*, *25*, 499-507.

Navon, D., & Purcell, D. (1981). Does integration produce masking or protect from it? *Perception, 10,* 71-83.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106,* 226-254.

Ngo, M. K., & Spence, C. (2010a). Crossmodal facilitation of masked visual target discrimination by informative auditory cuing. *Neuroscience Letters, 479,* 102-106.

Ngo, M. K., & Spence, C. (2010b). Crossmodal facilitation of masked visual target identification. *Attention, Perception, & Psychophysics, 72, 1938-1947.*

Ngo, M. K., & Spence, C. (2010c). Auditory, tactile, and multisensory cues facilitate search for dynamic visual stimuli. *Attention, Perception, & Psychophysics, 72,* 1654-1665.

Nickerson, R. S. (1973). Intersensory facilitation of reaction time: Energy summation or preparation enhancement? *Psychological Review, 80,* 489-509.

Noesselt, T., Bergmann, D., Hake, M., Heinze, H.-J., & Fendrich, R. (2008). Sound increases the saliency of visual events. *Brain Research, 1220,* 157-163.

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,* 11431-11441.

Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex, 18,* 598-609.

Noppeney, U., Ostwald, D., & Werner, S. (2010). Perceptual decisions formed by

accumulation of audiovisual evidence in prefrontal cortex. *Journal of Neuroscience, 30,* 7434-7446.

Norman, D. A. (1968). Towards a theory of memory and attention. *Psychological Review, 75,* 522-536.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7,* 44-64.

O'Conner, K. J., & Potter, M. C. (2002). Constrained formation of object representations. *Psychological Science, 13,* 106-111.

O'Connor, N., & Hermelin, B. (1972). Seeing and hearing in space and time. *Perception & Psychophysics, 11,* 46-48.

Odgaard, E. C., Arieh, Y., & Marks, L. E. (2003). Cross-modal enhancement of perceived brightness: Sensory interaction versus response bias. *Perception & Psychophysics, 65,* 123-132.

Odgaard, E. C., Arieh, Y., & Marks, L. E. (2004). Brighter noise: Sensory enhancement of perceived loudness by concurrent visual stimulation. *Cognitive, Affective, & Behavioral Neuroscience, 4,* 127-132.

Ogmen, H., Breitmeyer, B. G., & Melvin, R. (2003). The what and where in visual masking. *Vision Research, 43,* 1337-1350.

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

Olivers, C. N. L., & van der Burg, E. (2008). Bleeping you out of the blink: Sound saves vision from oblivion. *Brain Research, 1242,* 191-199.

Orgs, G., Lange, K., Dombrowski, J.-H., & Heil, M. (2007). Is conceptual priming for environmental sounds obligatory? *International Journal of Psychophysiology, 65,* 162-166.

Orgs, G., Lange, K., Dombrowski, J.-H., & Heil, M. (2008). N400-effects to task-irrelevant environmental sounds: Further evidence for obligatory conceptual processing. *Neuroscience Letters, 436,* 133-137.

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3,* 519-526.

Parise, C., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic associations modulate multisensory integration in non-synesthetes. *PLoS ONE, 4,* e5664. doi:10.1371/journal.pone.0005664

Partan, S., & Marler, P. (1999). Communication goes multimodal. *Science, 283,* 1272-1273.

Pearson, J., Clifford, C. W. G., & Tong, F. (2008). The functional impact of mental imagery on conscious perception. *Current Biology, 18,* 982-986.

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America A, 2,* 1508-1532.

Pennartz, C. M. A. (2009). Identification and integration of sensory modalities: Neural basis and relation to consciousness. *Consciousness and Cognition, 18,* 718-739.

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,* 339-352.

Petrini, K., Russell, M., & Pollick, F. (2009). When knowing can replace seeing in audiovisual integration of actions. *Cognition, 110,* 432-439.

Pick, H. L., Jr., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Perception & Psychophysics, 6,* 203-205.

Plaut, D. C. (2002). Graded modality-specific specialization in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology, 19,* 603-639.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review, 83,* 157-171.

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience, 139,* 23-38.

Potter, M. C. (1975). Meaning in visual search. *Science, 187,* 965-966.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 509-522.

Potter, M. C. (1993). Very short-term conceptual memory. *Memory & Cognition, 21,* 156-161.

Potter, M. C. (1999). Understanding sentences and scenes: The role of conceptual short-term memory. In V. Coltheart (Ed.), *Fleeting memories: Cognition of brief visual stimuli* (pp. 13-46). Cambridge, MA: MIT Press.

Potter, M. C., Chun, M. M., Banks, B. S., & Muckenhoupt, M. (1998). Two attentional

deficits in serial target search: The visual attentional blink and an amodal task-switch deficit. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 979-992.

Potter, M. C., & Faulconer, B. A. (1975). Time to understand pictures and words. *Nature, 253,* 437-438.

Potter, M. C., Stiefbold, D. R., & Moryadas, A. (1998). Word selection in reading sentences: Preceding versus following contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 68-100.

Prinzmetal, W., McCool, C., & Park, S. (2005). Attention: Reaction time and accuracy reveal different mechanisms. *Journal of Experimental Psychology: General*, *134*, 73-92.

Pylyshyn, Z. W. (1984). *Computation and cognition.* Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial index model. *Cognition, 32,* 65-97.

Pylyshyn, Z. W. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22,* 341-423.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80,* 127-158.

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

Radeau, M., & Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Perception & Psychophysics, 22,* 137-146.

Rapp, B., & Hendel, S. K. (2003). Principles of cross-modal competition: Evidence from deficits of attention. *Psychonomic Bulletin & Review, 10,* 210-219.

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 USA, 97,* 11800-11806.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance, 18,* 849-860.

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

Reed, C. L., Klatzky, R. L., & Halgren, H. (2005). What vs. where in touch: An fMRI study. *NeuroImage, 25,* 718-726.

Rees, G., Frith, C., & Lavie, N. (2001). Processing of irrelevant visual motion during performance of an auditory attention task. *Neuropsychologia*, *39*, 937-949.

Reeves, A. (1982). Metacontrast U-shaped functions derive from two monotonic functions. *Perception, 11,* 415-426.

Renier, L. A., Anurova, I., De Volder, A. G., Carlson, S., VanMeter, J., & Rauschecker, J. P. (2009). Multisensory integration of sounds and vibrotactile stimuli in processing streams for "what" and "where". *Journal of Neuroscience, 29,* 10950-10960.

Robertson, I. H., Mattingley, J. B., Rorden, C., & Driver, J. (1998). Phasic alerting of neglect patients overcomes their spatial deficit in visual awareness. *Nature, 395,* 169-172.

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

Roelofs, A. (2005). The visual-auditory color-word Stroop asymmetry and its time course. *Memory & Cognition, 33,* 1325-1336.

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., & Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nature Neuroscience, 2,* 1131-1136.

Romei, V., Murray, M. M., Cappe, C., & Thut, G. (2009). Preperceptual and stimulus-selective enhancement of low-level human visual cortex excitability by sounds. *Current Biology, 19,* 1799-1805.

Rommetveit, R., Toch, H., & Svendsen, D. (1968). Effects of contingency and contrast contexts on the cognition of words: A study of stereoscopic rivalry. *Scandinavian Journal of Psychology, 9,* 138-144.

Roseboom, W., Nishida, S., & Arnold, D. H. (2009). The sliding window of audio-visual simultaneity. *Journal of Vision, 9(12):4,* 1-8.

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,* 1147-1153.

Sakowitz, O. W., Quian Quiroga, R., Schürmann, M., & Başar, E. (2005). Spatio-temporal frequency characteristics of intersensory components in audiovisually evoked potentials. *Brain Research Cognitive Brain Research, 23,*

316-326.

Santangelo, V., Belardinelli, M. O., & Spence, C. (2007). The suppression of reflexive visual and auditory orienting when attention is otherwise engaged. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 137-148.

Santangelo, V., & Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 1311-1321.

Santee, J. L., & Egeth, H. E. (1982). Do reaction time and accuracy measure the same aspects of letter recognition? *Journal of Experimental Psychology: Human Perception and Performance, 8,* 489-501.

Saygin, A. P., Dick, F., & Bates, E. (2005). An on-line task for contrasting auditory processing in the verbal and nonverbal domains and norms for younger and older adults. *Behavior Research Methods, 37,* 99-110.

Saygin, A. P., Dick, F., Wilson, S. W., Dronkers, N. F., & Bates, E. (2003). Neural resources for processing language and environmental sounds. *Brain, 126,* 928-945.

Schneider, T. R., Debener, S., Oostenveld, R., & Engel, A. K. (2008). Enhanced EEG gamma-band activity reflects multisensory semantic matching in visual-to-auditory object priming. *NeuroImage, 42,* 1244-1254.

Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Experimental Psychology, 55,* 121-132.

Schneider, W., Dumais, S. T., & Shiffrin, R. M. (1984). Automatic and control processing and attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 1-27). New York: Academic Press.

Schnupp, J., W. H., Dawe, K. L., & Pollack, G. L. (2005). The detection of multisensory stimuli in an orthogonal sensory space. *Experimental Brain Research, 162,* 181-190.

Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level, 'unisensory' processing. *Current Opinion in Neurobiology, 15,* 454-458.

Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 1791-1810.

Segal, S. J., & Fusella, V. (1969). Effects of imaging on signal-to-noise ratio, with varying signal conditions. *British Journal of Psychology, 60,* 459-464.

Segal, S. J., & Fusella, V. (1970). Influence of imagined pictures and sounds on detection of visual and auditory signals. *Journal of Experimental Psychology, 83,* 458-464.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385,* 308.

Senkowski, D., Molholm, S., Gomez-Ramirez, M., & Foxe, J. J. (2006). Oscillatory beta activity predicts response speed during a multisensory audiovisual reaction time task: A high-density electrical mapping study. *Cerebral Cortex, 16,* 1556-1565.

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,* 401-409.

Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C. S., & Woldorff, M. G. (2007). Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations. *Neuropsychologia, 45,* 561-571.

Shallice, T. (1988). Specialization within the semantic system. *Cognitive Neuropsychology, 5,* 133-142.

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

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear: Sound induced visual flashing. *Nature, 408,* 788.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research, 14,* 147-152.

Sheinberg, D. L., & Logothetis, N. K. (1997). The role of temporal cortical areas in perceptual organization. *Proceedings of the National Academy of Sciences USA, 94,* 3408-3413.

Sheth, B. R., & Shimojo, S. (2004). Sound-aided recovery from and persistence against visual filling-in. *Vision Research, 44,* 1907-1917.

Shi, Z., Chen, L., & Muller, H. J. (2010). Auditory temporal modulation of the visual Ternus effect: The influence of time interval. *Experimental Brain Research, 203,* 723-735.

Shimada, H. (1990). Effect of auditory presentation of words on color naming: The

intermodal Stroop effect. *Perceptual & Motor Skills, 70,* 1155-1161.

Shimojo, S., & Nakayama, K. (1990). Real world occlusion constraints and binocular rivalry. *Vision Research, 30,* 69-80.

Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: Plasticity and interactions. *Current Opinion in Neurobiology, 11,* 505-509.

Shimojo, S., Watanabe, K., & Scheier, C. (2001). The resolution of ambiguous motion: Attentional modulation and development. In J. Braun, C. Koch, & J. L. Davis (Eds.), *Visual attention and control circuits* (pp. 243-264). Cambridge, MA: MIT Press.

Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science, 145,* 1328-1330.

Shore, D. I., Barnes, M. E., & Spence, C. (2006). Temporal aspects of the visuotactile congruency effect. *Neuroscience Letters, 392,* 96-100.

Sinnett, S., Costa, A., & Soto-Faraco, S. (2006). Manipulating inattentional blindness within and across sensory modalities. *Quarterly Journal of Experimental Psychology, 59,* 1425-1442.

Sinnett, S., Soto-Faraco, S., Spence, C. (2008). The co-occurrence of multisensory competition and facilitation. *Acta Psychologica, 128,* 153-161.

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

Smith, M. C., & Magee, L. E. (1980). Tracing the time course of picture-word processing. *Journal of Experimental Psychology: General, 109,* 373-392.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 174-215.

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 580-587.

Soto-Faraco, S., Kingstone, A., & Spence, C. (2003). Multisensory contributions to the perception of motion. *Neuropsychologia, 41,* 1847-1862.

Soto-Faraco, S., & Spence, C. (2002). Modality-specific auditory and visual temporal processing deficits. *Quarterly Journal of Experimental Psychology A*, *55*, 23-40.

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science &*

Technology, *28*, 61-70.

Spence, C. (2009). Explaining the Colavita visual dominance effect. *Progress in Brain Research, 176,* 245-258.

Spence, C. (2010a). Crossmodal spatial attention. *Annals of the New York Academy of Sciences, 1191,* 182-200.

Spence, C. (2010b). Crossmodal attention. *Scholarpedia, 5(5)*:6309. doi:10.4249/scholarpedia.6309

Spence, C. (in press). Multisensory perception, cognition, and behavior: Evaluating the factors modulating multisensory integration. In B. E. Stein (Ed.), *The new handbook of multisensory processing.* Cambridge, MA: MIT Press.

Spence, C. (submitted). Crossmodal correspondences and multisensory integration: A tutorial review. *Attention, Perception, & Psychophysics.*

Spence, C., & Chen, Y.-C. (in press). Intramodal and crossmodal perceptual grouping. In B. E. Stein (Ed.), *The new handbook of multisensory processing.* Cambridge, MA: MIT Press.

Spence, C. [J.], & Driver, J. (1994). Covert spatial orienting in audition: Exogenous and endogenous mechanisms. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 555-574.

Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics*, *59,* 1-22.

Spence, C., McDonald, J., & Driver, J. (2004). Exogenous spatial cuing studies of human crossmodal attention and multisensory integration. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp. 277-320). Oxford, UK: Oxford University Press.

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

Spence, C., Nicholls, M. E. R., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics, 63,* 330-336.

Spence, C., Parise, C., & Chen, Y.-C. (in press). The Colavita visual dominance effect. In M. M. Murray & M. Wallace (Eds.), *Frontiers in the neural bases of multisensory processes.*

Spence, C., Ranson, J., & Driver, J. (2000). Cross-modal selective attention: On the

difficulty of ignoring sounds at the locus of visual attention. *Perception & Psychophysics, 62,* 410-424.

Spence, C., & Soto-Faraco, S. (2009). Auditory perception: Interactions with vision. In C. Plack (Ed.), Auditory perception (pp. 271-296). Oxford: Oxford University Press.

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

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

Spencer, T. J., & Shuntich, R. (1970). Evidence for an interruption theory of backward masking. *Journal of Experimental Psychology, 85,* 198-203.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12,* 153-156.

Stein, B. E., London, N., Wilkinson, L. K., & Price, D. D. (1996). Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. *Journal of Cognitive Neuroscience, 8,* 497-506.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*, Cambridge, MA: MIT Press.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience, 9,* 255-266.

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,* 113-126.

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *NeuroImage, 44,* 1210-1223.

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

Stone, J. V., Hunkin, N. M., Porrill, J., Wood, R., Keeler, V., Beanland, M., et al. (2001). When is now? Perception of simultaneity. *Proceedings of the Royal Society B*, *268*, 31-38.

Stuart, D. M., & Carrasco, M. (1993). Semantic component of a cross-modal Stroop-like task. *American Journal of Psychology, 106,* 383-405.

Stuart, G. P., & Jones, D. M. (1995). Priming the identification of environmental sounds. *Quarterly Journal of Experimental Psychology, 48A,* 741-761.

Stuart, G. P., & Jones, D. M. (1996). From auditory image to auditory percept: Facilitation through common processes? *Memory & Cognition, 24,* 296-304.

Sugita, Y., & Suzuki, Y. (2003). Implicit estimation of sound-arrival time. *Nature, 421,* 911.

Suied, C., Bonneel, N., & Viaud-Delmon, I. (2009). Integration of auditory and visual information in the recognition of realistic objects. *Experimental Brain Research, 194,* 91-102.

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,* 679-690.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14,* 400-410.

Tamura, H., & Tanaka, K. (2001). Visual response properties of cells in the ventral and dorsal parts of the macaque inferotemporal cortex. *Cerebral Cortex*, *11*, 384-399.

Taylor, K. I., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2006). Binding crossmodal object features in perirhinal cortex. *Proceedings of the National Academy of Sciences USA, 103,* 8239-8244.

Tellinghuisen, D. J., & Nowak, E. J. (2003). The inability to ignore auditory distractors as a function of visual task perceptual load. *Perception & Psychophysics, 65,* 817-828.

Theeuwes, J. (2010). Top-down and bottom-up control of visual selection. *Acta Psychologica, 123,* 77-99

Thierry, G., Giraud, A. L., & Price, C. J. (2003). Hemispheric dissociation in access to the human semantic system. *Neuron, 38,* 499-506.

Thierry, G., & Price, C. J. (2006). Dissociating verbal and nonverbal conceptual processing in the human brain. *Journal of Cognitive Neuroscience, 18,* 1018-1028.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520-522.

Tong, F. (2001). Competing theories of binocular rivalry: A possible resolution. *Brain and Mind, 2,* 55-83.

Tong, F., & Engel, S. A. (2001). Interocular rivalry revealed in the human cortical blind-spot representation. *Nature, 411,* 195-199.

Tong, F., Meng, M., & Blake, R. (2006). Neural bases of binocular rivalry. *Trends in Cognitive Sciences, 10,* 502-511.

Tong, F., Nakayama, K., Vaughan, J. T., & Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron, 21,* 753-759.

Treisman, A. (1962). Binocular rivalry and stereoscopic depth perception. *Quarterly Journal of Experimental Psychology, 14,* 23-37.

Treisman, A. (1969). Strategies and models of selective attention. *Psychological Review, 76,* 282-299.

Treisman, A., & Davies, A. (1973). Dividing attention to ear and eye. In S. Kornblulm (Ed.), *Attention and performance IV* (pp. 101-117). New York: Academic Press.

Treisman, A. M., & Kanwisher, N. G. (1998). Perceiving visually presented objects: Recognition, awareness, and modularity. *Current Opinion in Neurobiology, 8,* 218-226

Treisman, A., & Riley, J. G. A. (1969). Is selective attention selective perception or selective response? A further test. *Journal of Experimental Psychology, 79,* 27-34.

Tyler, L. K., Stamatakis, E. A., Bright, P., Acres, K., Abdallah, S., Rodd, J. M., & Moss, H. E. (2004). Processing objects at different levels of specificity. *Journal of Cognitive Neuroscience, 16,* 351-362.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549-580). Cambridge, MA: MIT Press.

Usher, M., & Donnelly, N. (1998). Visual synchrony affects binding and segmentation in perception. *Nature, 394,* 179-182.

van Boxtel, J. J. A., van Ee, R., & Erkelens, C. J. (2007). Dichoptic masking and binocular rivalry share common perceptual dynamics. *Journal of Vision, 7(14):3,* 1-11.

van der Burg, E,, Cass, J., Olivers, C. N. L., Theeuwes, J., & Alais, D. (2010). Efficient visual search from synchronized auditory signals requires transient audiovisual events. *PLoS ONE 5(5)*: e10664. doi:10.1371/journal.pone.0010664

van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008a). Pip and pop: Non-spatial auditory signals improve spatial visual search. *Journal of*

Experimental Psychology: Human Perception and Performance, 34, 1053-1065.

van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008b). Audiovisual events capture attention: Evidence from temporal order judgments. *Journal of Vision, 8(5):2,* 1-10.

van der Burg, E., Talsma, D., Olivers, C. N. L., Hickey, C., & Theeuwes, J. (submitted). Early multisensory interactions affect the competition among multiple visual objects.

van Ee, R., van Boxtel, J. J. A., Parker, A. L., & Alais, D. (2009). Multimodal congruency as a mechanism for willful control over perceptual awareness. *Journal of Neuroscience, 29,* 11641-11649.

van Ee, R., van Dam, L. C. J., & Brouwer, G. J. (2005). Voluntary control and the dynamics of perceptual bi-stability. *Vision Research, 45,* 41-55.

van Petten, C., & Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia, 33,* 485-508.

van Vleet, T., Robertson, L. C. (2006). Cross-modal interactions in time and space: Auditory influences on visual attention in patients with hemispatial neglect. *Journal of Cognitive Neuroscience, 18,* 1368-1379.

van Wassenhove, V., Buonomano, D. V., Shimojo, S., & Shams, L. (2008). Distortions of subjective time perception within and across senses. *PLoS ONE, 3(1),* e1437. doi:10.1371/journal.pone.0001437

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences USA*, *102*, 1181-1186.

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

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

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

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

12-23.

Vatakis, A., & Spence, C. (2010). Audiovisual temporal integration for complex speech, object-action, animal call, and musical stimuli. In M. J. Naumer & J. Kaiser (Eds.), *Multisensory object perception in the primate brain* (pp. 95-121). New York: Springer.

Versace, R., Labeye, E., Badard, G., & Rose, M. (2009). The contents of long-term memory and the emergence of knowledge. *European Journal of Cognitive Psychology, 21,* 522-560.

von Kriegstein, K., & Giraud, A.-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology, 4(10): e326.* DOI: 10.1371/journal.pbio.0040326

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

Vroomen, J., & de Gelder, B. (2000). Sound enhances visual perception: Crossmodal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 1583-1590.

Vroomen, J., & de Gelder, B. (2004). Temporal ventriloquism: Sound modulates the flash-lag effect. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 513-518.

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

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception, & Psychophysics, 72*, 871-884.

Walker, J. T., & Scott, K. J. (1981). Auditory-visual conflicts in the perceived duration of lights, tones, and gaps. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 1327-1339.

Watanabe, K., & Shimojo, S. (2001). When sound affects vision: Effects of auditory grouping on visual motion perception. *Psychological Science, 12,* 109-116.

Watkins, S., Shams, L., Josephs, O., & Rees, G. (2007). Activity in human V1 follows multisensory perception. *Neuroimage, 37,* 572-578.

Watson, A. B., & Robson, J. G. (1981). Discrimination at threshold: Labelled detectors in human vision. *Vision Research, 21,* 1115-1122.

Watson, T. L., Pearson, J., & Clifford, C. W. G. (2004). Perceptual grouping of biological motion promotes binocular rivalry. *Current Biology, 14,* 1670-1674.

Watt, R. J. (1991). Studying human vision. *Understanding vision* (pp. 191-216). Academic Press: London.

Welch, R. B. (1999). Meaning, attention, and the 'unity assumption' in the intersensory bias of spatial and temporal perceptions. in G. Ashersleben, T. Bachmann, & J. Musseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 371-387). Amsterdam: Elsevier Science.

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

Welch, R. B., DuttonHurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Perception & Psychophysics, 39,* 294-300.

Werner, S., & Noppeney, U. (2010a). Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *Journal of Neuroscience, 30,* 7434-7446.

Werner, S., & Noppeney, U. (2010b). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral Cortex, 20,* 1829-1842.

Werner, S., & Noppeney, U. (in press). The contributions of transient and sustained response codes to audiovisual integration. *Cerebral Cortex.* doi:10.1093/cercor/bhq161

Wertheimer, M. (1950). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A sourcebook of Gestalt psychology* (pp. 71-81). New York: The Humanities Press.

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

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics, 63,* 1314-1329.

Wickens, C. D. (1980). The structure of attentional resources. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 239-257). Hillsdale, NJ: Erlbaum.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 63-102). San Diego, CA: Academic Press.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomic Science, 3,* 159-177.

Widmann, A., Gruber, T., Kujala, T., Tervaniemi, M., & Schröger, E. (2007). Binding symbols and sounds: Evidence from event-related oscillatory gamma-band activity. *Cerebral Cortex, 17,* 2696-2702.

Witten, I. B., & Knudsen, E. I. (2005). Why seeing is believing: Merging auditory and visual worlds. *Neuron, 48,* 489-496.

Wyble, B., Bowman, H., & Nieuwenstein, M. (2009). The attentional blink provides episodic distinctiveness: Sparing at a cost. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 787-807.

Yeh, S. L., & Liao, H. I. (2008). On the generality of the contingent orienting hypothesis. *Acta Psychologica, 129,* 157-165.

Yu, C., Klein, S. A., & Levi, D. M. (2002). Facilitation of contrast detection by cross-oriented surround stimuli and its psychophysical mechanisms. *Journal of Vision, 2,* 243-255.

Yuval-Greenberg, S., & Deouell, L. Y. (2007). What you see is not (always) what you hear: Induced Gamma band responses reflect cross-modal interactions in familiar object recognition. *Journal of Neuroscience, 27,* 1090-1096.

Yuval-Greenberg, S., & Deouell, L. Y. (2009). The dog's meow: Asymmetrical interaction in cross-modal object recognition. *Experimental Brain Research, 193,* 603-614.

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & Psychophysics, 67,* 531-544.

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

Zatorre, R. J., & Halpern, A. R. (2005). Mental concerts: Musical imagery and auditory cortex. *Neuron, 47,* 9-12.

Zeki, S. (2003). The disunity of consciousness. *Trends in Cognitive Sciences, 7,* 214-218.

Zimba, L. D., & Blake, R. (1983). Binocular rivalry and semantic processing: Out of sight, out of mind. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 807-815.