Great expectations! How predictions about the magnitude of a forthcoming sound affect its perceptual and neural processing

Great expectations!
How predictions about the magnitude of a forthcoming sound affect its perceptual and neural processing

Author: Sol Libesman
Primary supervisor: Prof. Thomas J. Whitford
Primary supervisor: Dr Damien J. Mannion

A thesis presented for the degree of
Doctor of Philosophy

School of Psychology
Faculty of Science
September 2020

Thesis/Dissertation Sheet

Surname/Family Name	:	**Libesman**
Given Name/s	:	**Sol**
Abbreviation for degree as give in the University calendar	:	**Doctor of Philosophy**
Faculty	:	**Science**
School	:	**School of Psychology**
Thesis Title	:	**Great expectations! How predictions about the magnitude of a forthcoming sound affect its perceptual and neural processing**

Abstract 350 words maximum: (PLEASE TYPE)

Psychoacoustic research has primarily examined how the low-level properties of auditory waveforms influence perceived loudness. However, we often experience auditory events not in isolation, but accompanied by characteristic visual information. Despite this, the influence of vision on perceived loudness has rarely been considered. The intensity of an auditory signal at-the-ear depends on both the power of the sound source and the distance of the source from the listener. The aim of this thesis was to explore whether visual cues relaying information about these two factors influence perceived loudness. After providing an Introduction and Background to this thesis, Chapters 3 and 4 assessed whether visual cues that disambiguate the distance of a sound source influenced perceived loudness. These studies simulated a loudspeaker relaying sounds at different distances in either anechoic or reverberant conditions and extracted loudness judgements using a 2-interval forced-choice task. The findings of both chapters indicated that visual distance information was not exploited to disambiguate the loudness of an auditory event. Chapters 5 and 6 assessed whether visual information about the power of a sound source influenced the perception of loudness and the neurophysiological coding of auditory intensity. In both chapters visual information about the power of the sound source was provided with videos depicting a 'strong' and a 'weak' hand-clap. Chapter 5 employed a novel 2-pair forced-choice task to extract participants' loudness estimates. The results indicated that clap sounds that were paired with the 'strong' clap video were amplified in loudness relative to the 'weak' clap video. Chapter 6 used electroencaphalography (EEG) to test whether visual cues to sound source power are capable of altering the neurophysiological response to auditory stimulation. We found that when participants received identical auditory input, the 'strong' clap video evoked an increased auditory response relative to the 'weak' clap video. Chapters 5 and 6 provide evidence that the neurological and subjective coding of auditory intensity is influenced in accordance with the visually created expectation. Taken together, the findings of this thesis demonstrate that perceived loudness is dependent not only on auditory input per se, but also on higher-level predictions about the expected intensity of an auditory event.

Declaration relating to disposition of project thesis/dissertation

…………………………………………………………… ………..…………………..………

Signature Date

Abstract

Psychoacoustic research has primarily examined how the low-level properties of auditory waveforms influence perceived loudness. However, we often experience auditory events not in isolation, but accompanied by characteristic visual information. Despite this, the influence of vision on perceived loudness has rarely been considered. The intensity of an auditory signal at-the-ear depends on both the *power* of the sound source and the *distance* of the source from the listener. The aim of this thesis was to explore whether visual cues relaying information about these two factors influence perceived loudness. After providing an Introduction and Background to this thesis, Chapters 3 and 4 assessed whether visual cues that disambiguate the *distance* of a sound source influenced perceived loudness. These studies simulated a loudspeaker relaying sounds at different distances in either anechoic or reverberant conditions, and extracted loudness judgements using a 2-interval forced-choice task. The findings of both chapters indicated that visual distance information was not exploited to disambiguate the loudness of an auditory event. Chapters 5 and 6 assessed whether visual information about the *power* of a sound source influenced the perception of loudness and the neurophysiological coding of auditory intensity. In both chapters visual information about the power of the sound source was provided with videos depicting a 'strong' and a 'weak' hand-clap. Chapter 5 employed a novel 2-pair forced-choice task to extract participants' loudness estimates. The results indicated that clap sounds that were paired with the 'strong' clap video were amplified in loudness relative to the 'weak' clap video. Chapter 6 used electroencaphalography (EEG) to test whether visual cues to sound source power are capable of altering the neurophysiological response to auditory stimulation. We found that when participants received identical auditory input, the 'strong' clap video evoked an increased auditory response relative to the 'weak' clap video. Chapters 5 and 6 provide evidence that the neurological and subjective coding of auditory intensity is influenced in accordance with the visually created expectation. Taken together, the findings of this thesis demonstrate that perceived loudness is dependent not only on auditory input per se, but also on higher-level predictions about the expected intensity of an auditory event.

Contents

Thesis abstract ii
List of figures v
Acknowledgments vii

1 Introduction 1
1.1 Background to the thesis 1
1.2 Aim and scope 2
1.3 Thesis overview 2

2 Background 5
2.1 Why should we be interested in auditory intensity? 5
2.2 The causal influences of auditory intensity 7
2.3 Signal intensity and audition 9
2.4 The effect of sound source distance and reverberation on perceived loudness 21
2.5 The role of visual signals in audition 24
2.6 Visual signals carrying intensity information 30
2.7 Research Motivation 39

3 Visual Cues to Source Distance & Loudness (anechoic conditions) 41
3.1 Abstract 42
3.2 Experiment 1 47
3.3 Experiment 2 54
3.4 Experiment 3 57
3.5 Experiment 4 60
3.6 Experiment 5 63
3.7 Experiment 6 65
3.8 General Discussion 67
3.9 Conclusion 70

4 Visual Cues to Source Distance & Loudness (reverberant conditions) 71

4.1 Abstract 72

4.2 Introduction 73

4.3 Methods 78

4.4 Results 85

4.5 Discussion 87

5 Visual Cues to Source Power & Loudness (behavioural experiment) 92

5.1 Abstract 93

5.2 Introduction 93

5.3 Methods 99

5.4 Results 105

5.5 Discussion 107

6 Visual Cues to Source Power & Loudness (EEG experiment) 112

6.1 Abstract 113

6.2 Introduction 113

6.3 Methods 114

6.4 Results 119

6.5 Discussion 122

7 General Discussion 127

7.1 Summary 127

7.2 Implications 130

7.3 Future research 139

7.4 Conclusion 143

List of Figures

2.1 Influence graph 7
2.2 Influence graph - intensity 9
2.3 Schematic of a soundwave's physical features 11
2.4 N1 component - intensity dependence 20
2.5 Influence graph - power, distance and sound field 21
2.6 Influence graph - visual signals 30

3.1 Anechoic field Experiments 1-2 - visual stimuli 46
3.2 Anechoic field Experiments - example of the psychometric functions . . . 52
3.3 Anechoic field Experiment 1 - posterior results 53
3.4 Anechoic field Experiment 2 - posterior results 56
3.5 Anechoic field Experiments 3-6 - visual stimuli 57
3.6 Anechoic field Experiment 3 - posterior results 59
3.7 Anechoic field Experiment 4 - posterior results 62
3.8 Anechoic field Experiment 5 - posterior results 64
3.9 Anechoic field Experiment 6 - posterior results 67
3.10 Ponzo illusion 69

4.1 Reverberant hall Experiment - soundwave properties 79
4.2 Reverberant hall Experiment - visual stimuli 80
4.3 Reverberant hall Experiment - mean PSEs 87

5.1 Clap Experiment (behavioural) - hypothesised PSEs 98
5.2 Clap Experiment (behavioural) - visual stimuli 101
5.3 Clap Experiment (behavioural) - the structure of a single trial 102
5.4 Clap Experiment (behavioural) - mean PSEs 106

6.1 Clap Experiment (EEG) - stimuli 117
6.2 Clap Experiment (EEG) - waveforms short timewindow 119
6.3 Clap Experiment (EEG) - waveforms long timewindow 120

6.4 Clap Experiment (EEG) - results: amplitude of the N1 121

Acknowledgments

First, to Thomas Whitford and Damien Mannion, salt and pepper, UNSW's Hammer and Sickle, Science's Starsky and Hutch, Academia's McNulty and Bunk - I could not have wished for a better duo to mentor me. While many people feel like universities are impersonal and isolating places, my experience has been one that is in complete contrast to this. I thank both Tom and Damien for making all the journal clubs, research discussions, breakfasts, lunches, meet-ups, and pub crawls feel less like team meetings and more like family gatherings. The effort you have both invested into sewing a tight knit community around our groups has left us not only feeling inspired to dig into our research but also excited to go to work. At a personal level I am grateful for the way in which you have unconditionally dedicated generous amounts of time to relay your knowledge and feedback, *regardless* of how flanked you have been with other commitments. Destiny waited for this final draft, it was certainly less patient than both of you.

I was lucky to be embedded within two sensational labs and surrounded by a close group of associates. Within this core network I thank Anthony Harrison, Nathan Han, Gabrielle Rudman, Kin Hei Lawrence Chung, Wadim Vodovozov, Oren Griffiths, Bradley Jack, Nathan Mifsud, Ruth Elijah, Lindsay Peterson, Kevin Tsang, Hua-Chun Sun, Mike Le Pelley, Poppy Watson, Daniel Pearson, Colin Clifford and Tarryn Balsdon, for the friendship, support and providing an endless springboard of inspiration. I thank the rest of the PhD cohort, in particular level 14, you guys made coming in a pleasure. I thank Ruth and Shanta Dey, you two were like the older sisters I never had. If you two didn't hold my hand through the first two years of my postgraduate life I'm sure I would still be under my desk in fetal position right now. The Ping pong crew, Lachlan Ferguson, Alexie Dawes and James Peak, you three were integral to my sanity. Lunch time never looked so lush.

Beyond the boundaries of UNSW's Mathews building, I thank my mum (Terri Libesman) for the endless love, support and guidance. The shadow supervisor of this project and the true MVP. I could not have done it without you. My dad (Vince Wall) and brothers (Ewan Wall, Ruben Wall) for being the spirit animals of this thesis. Last, the division of (emotional) labour has been split across a mass of gorgeous and pretty special people in my friendship circles. In particular, I have to single out my flatmates who have been on the front lines with me, Alex Walsh-Rossi, Anna-Luisa Ruther, George (Yiorgos) Vlotis, James White, Leo Cornish, Oscar Nimmo, Simon Comensoli, and surrogate flatmates Ned Hunter-Gilson, Thuso Lekwape, Ellen Moore, Anna Lene

Seidler and Will Edmonds - you are all legends, thank you for providing scorching warmth every time I retreated home!

Inclusion of Publication Statement

☒ The candidate has declared that some of the work described in their thesis has been published and has been documented in the relevant Chapters with acknowledgement.

A short statement on where this work appears in the thesis and how this work is acknowledged within chapter/s:

> Chapter 6 is comprised of an experimental paper which has been published in the Journal of Cognitive Neuroscience. Acknowledgment of the authors' contributions has been provided at the beginning of the Chapter. A reference to the publication has also been provided at the beginning of the Chapter.

Originality statement

☒ I hereby declare that this submission is my own work and that to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation, and linguistic expression is acknowledged.

Sol Libesman

Copyright Statement

☒ I hereby grant the University of New South Wales or its agents a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

For any substantial portions of copyright material used in this thesis, written permission for use has been obtained, or the copyright material is removed from the final public version of the thesis.

Authenticity Statement

☒ I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

Sol Libesman

1 Introduction

1.1 Background to the thesis

Transforming sensory input into useful representations of the external environment is the task of our perceptual system. Whilst we may feel that our percepts are the result of our receptors passively relaying what is 'out there' in the world, what we often do not notice is the active role our brain plays in interpreting sensory input. A signal arriving at our sensory receptors may be caused by multiple different environmental generators. In turn, the same sensory input may give rise to a variety of possible representations of the external environment. This creates the potential for uncertainty. A popular theory of perception is that our brain mitigates the uncertainty of representation (i.e., uncertainty as to what is 'out there' generating the sensory signal) by actively meshing predictive models of the world with sensory input. With regards to the auditory system, psychoacoustic research has primarily focused on the biological and behavioral consequences of lower order features of sounds (for example, by manipulating the frequency of sinusoidal tones to isolate how this basic property of sound influences its perceived loudness). But are conclusions based on experiments using simple and abstract listening conditions applicable to understanding how our auditory system navigates environments out in the world?

If our perception of loudness is related to the way in which we represent the environmental events that give rise to sounds, causal inferences may play an important role in this process. We often do not only hear an auditory signal, but we also see the physical objects that are causing it. Viewing the movement of objects that give rise to a sound (e.g., viewing a car crashing into a telegraph pole) can provide an indication as to the likely magnitude of the resulting sound. We are regularly presented with environmental regularities between visual events and auditory input: a powerful collision will typically produce auditory input of higher intensity than a weak collision occurring at the same location. Likewise, a sound source that is far away will produce auditory input of lower intensity than an identical sound source positioned nearby. It is the premise of this thesis that if the brain functions as a causal inference machine, then visually-perceived events may influence our processing of auditory input.

1.2 Aim and scope

The aim of this thesis is to establish whether visual information regarding the physical causes of sounds may influence (a) subjective loudness judgements, and (b) neurophysiological indices of auditory intensity. I look to investigate two types of visual cues that can potentially disambiguate source intensity: firstly, cues to the *distance* of a sound source and secondly, cues as to the *physical movements* that produce the sound. In this thesis, subjective judgements of loudness were quantified using behavioral psychophysics techniques which required the relative judgements of stimuli. I also quantified the activation of the auditory cortex using electroencephalography (EEG). Specifically, I extracted event-related potentials (ERPs) evoked by auditory stimuli paired with visual information that predicted different source intensities. Whilst I will place these findings within the broader context of active perceptual inferences, the specific focus of this thesis is to determine whether visual signals carrying information about the likely intensity of sounds have any influence on auditory processing.

1.3 Thesis overview

To address the above aims, I conducted a series of experiments in which an auditory signal was paired with visual information that was predictive of its intensity. This thesis begins by providing a background (chapter 2) to the broader context of these studies. The background section starts by examining the current evidence for how the intensity of sound waves relate to both (1) subjective judgements of loudness and (2) the magnitude of the auditory evoked potential. Next, I develop a broader framework in which to conceptualise the influence of higher order causal information on auditory processing. Finally, I look to synthesise these two areas and discuss why predictive visual signals could plausibly influence the processing of auditory intensity.

Chapters 3, 4, 5 and 6 form the novel research components of this thesis. Chapters 3 and 4 describe experiments that explore whether viewing the distance of a sound source alters a person's subjective loudness judgements. In both chapters, a computer-generated loudspeaker played sounds at different depths in a scene. Images of the loudspeaker were depicted on a computer monitor and sounds were relayed over headphones. Both chapters implemented a 2-interval-forced-choice task in order to extract estimates of relative loudness. Chapter 3 explores loudness in anechoic conditions using a speaker relaying sounds in an open field at three distances. This study consisted of 6 experiments in which the presented sounds were either a 250 Hz pure tone (exp 1-3), pink noise (exp 4), the utterance ‘ba’ (exp 5), or a 250 Hz pure tone with an ecological time delay based on the distance of the speaker (exp 6). The results revealed that the apparent distance of the speaker did not influence participants’ loudness judgments. In light of evidence suggesting that sounds generated in anechoic environments can fail to externalize (i.e., are perceived as being generated from within the head), it is possible that some reverberation is needed to facilitate the externalisation of sounds to visual objects. Chapter 4 explores loudness estimates of a loudspeaker which delivered sounds within a reverberant hall. While Chapter 3 required participants to estimate the apparent loudness of sounds, Chapter 4 required participants to estimate the distal loudness of the sound at-its-source. The visual environment was captured with real-life photos of a loudspeaker in a concert hall, with the loudspeaker presented at five distances. Sounds were convolved with impulse response recordings taken from the speaker playing sounds at the respective distances in the same hall. In this study there were four conditions: an audio-only condition, an audiovisual condition, and two conditions where the visual cues were systematically shifted to be either closer or farther than the paired auditory recording. We again find that loudness judgements were primarily determined by the auditory signal itself, and that visual distance information did not increase the accuracy with which the intensity of an auditory event may be estimated.

Chapter 5 and 6 looked to determine whether pairing a sound with a video depicting physical movements that would be expected to produce a high or low intensity sound would alter either the (a) perceived loudness of the sound or, (b) the auditory-evoked potential elicited by the sound. In both chapters, sound source intensity predictions were generated using videos of ‘strong’ (i.e., vigorous) or ‘weak’ (i.e., non-vigorous) handclaps. These visual cues were paired with auditory stimuli (i.e., sounds of handclaps) at different sound intensities. Chapter 5 used a two-pair forced choice psychophysics design in order to gauge perceived loudness. In this experiment, it was found that visual cues to sound source intensity influenced participants’ loudness estimates. Specifically, sounds paired with the ‘strong’ handclap were perceived as being relatively louder than those paired with the ‘weak’ handclap. Chapter 6 looked at the neurophysiological processing of sound source intensity using EEG. The auditory N1 component was used as an index of primary auditory cortex activation. In this experiment, it was found that the amplitude of the N1 increased when an identical auditory signal was paired with videos of a ‘strong’ handclap compared to a ‘weak’ handclap. This result indicated that the activation of the auditory cortex was shifted in accordance with the visually-created expectation.

In Chapter 7, I summarise the findings of this thesis. I discuss the potential mechanisms by which loudness may be modulated by predictive cues, discuss the limitations to the current experiments, and make suggestions for future research.

2 Background

In this Chapter I lay out some of the concepts that are relevant to my research program, which investigates the influence of expectations on perceived loudness. I begin by discussing the importance of research into audition generally, and into sound intensity more specifically (2.1). Next, I establish a causal model of the primary streams of information carried from a sound source (2.2), which will provide the organisational scaffolding on which the subsequent sections will build. I will start by reviewing how a sound's physical intensity relates to (a) the perception of loudness and (b) the elicited neural response (2.3). I will then describe how both the distance a sound travels and the environment through which it travels, may affect the auditory signal reaching our ear, which in turn affects our perception of loudness (2.4). Following this, I will review the general conditions under which visual signals have previously been shown to influence auditory perception (2.5). Finally, I synthesise this knowledge and demonstrate that the perceived intensity of sound sources may be influenced by visual signals carrying information about source intensity and source distance (2.6). This finding is the primary motivation for the research component of this thesis (2.7).

2.1 Why should we be interested in auditory intensity?

Intensity is one of the basic features of an auditory signal (Bizley & Cohen, 2013). Auditory intensity is closely related to loudness, which can be defined as the *subjective* intensity of an auditory signal (Scharf, 1978). At its most basic level, the capacity to make use of auditory signals through auditory organs emerged in the process of evolution in order to help animals anticipate and respond to danger (Tumarkin, 1968). High intensity sound is a primary auditory cue that can alert us to potential threats. Consequently, the inner ear has direct connections with neural mechanisms in the autonomic nervous system that underlie the fight-or-flight response (May, Little, & Saylor, 2009; Pfingst, Hienz, Kimm, & Miller, 1975; Stebbins & Miller, 1964; Wagner, Florentine, Buus, & McCormack, 2004; Yeomans & Frankland, 1995). This defensive function cannot be turned off; sounds will even register in the brain during sleep (Westman & Walters, 1981).

While it is well established that auditory organs initially evolved to facilitate basic survival responses, it has also been established that auditory intensity plays a functional role in aiding auditory object recognition (Norman-Haignere & McDermott, 2018, September 16). In this process, what is less well established is how auditory intensity is represented and, more fundamentally, how auditory objects are decoded from the physical signals we receive. To date, most psychoacoustic research has focused on mapping out how loudness judgements change when manipulating simple low-order stimuli (e.g., varying the frequency and intensity of sinusoidal sound waves) (Florentine & Zwicker, 1979; Moore, 2012; Moore & Glasberg, 2004; Moore, Glasberg, & Baer, 1997; Moore, Glasberg, & Stone, 2010). These stimuli have often been synthesized in a way abstracted from ecological experience, with an ambiguous relationship with how the sound at-the-ear would relate to an external sound source. Current conceptualisations of loudness do not adequately account for the influence of higher order information, such as that which can be provided by visual cues (M. Epstein & Florentine, 2009, 2012; Moore, 2014). This is an important limitation as our perceptual systems often do not aim to capture the exact features of sensory input, but rather to relate sensory input to useful features of causal events and objects out in the world (e.g., Knill & Richards, 1996; Körding et al., 2007; Schutz & Kubovy, 2009; Shams & Beierholm, 2010). This idea was clearly articulated by one of the pioneers in perception, von Helmholtz (quoted in Warren, 1981):

> We are exceedingly well trained in finding out by our sensations the objective nature of the objects around us, but we are completely unskilled in observing these sensations per se; and the practice of associating them with things outside of us actually prevents us from being distinctly conscious of the pure sensations.

If perceptual systems function less to capture a veridical representation of input and more to estimate the properties of the external environment that give rise to these signals, how do they achieve this goal? When estimating the external environment our sensory system has to handle a causal inference problem: there is no direct, objective link between input and representation. That is, the same signal at our sensory receptor may be produced by multiple environmental generators. For example, a plane in the sky can produce a visual signal that is the same size on the retina as a toy plane in one's hand; likewise, a whisper in the ear can produce auditory input of the same intensity as a shout coming from down the road. As a result, there is inherently a degree of uncertainty involved in transforming input arriving at our sensory receptors into a representation of an external source. The present thesis aims to elucidate whether a person's causal models of the world can influence their perception of loudness, and their neurophysiological response to sound. If our perceptual system engages in causal inference of this nature, a supplementary aim of this thesis will be to examine whether our perception of auditory intensity is based on an estimation of the magnitude of a sound wave at our receptors, or an estimation of the magnitude of the sound wave at-its-source.

2.2 The causal influences of auditory intensity

In order to conceptualise how inferential processes could influence auditory perception, it is useful to establish a model relating the physical signals we receive at-our-ears to the events in the external world that generate them.

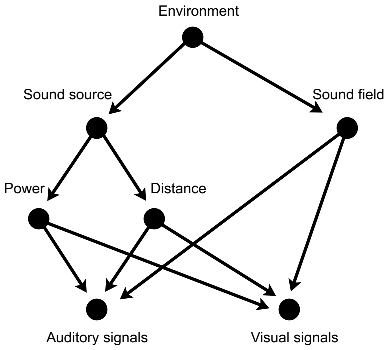

Figure 2.1. Influence graph for an example scenario. Nodes (circles) represent variables, and directed arrows represent causal relationships. Note that any perceptual inferences about signal-generating events go in the opposite direction to the arrows, from the bottom of the graph upwards. (after Kersten et al., 2004).

Information carried in a signal is physically related to a generative source. The relationship between a signal and a source can be broken down into a number of influences, as shown in Figure 2.1. The key point is this: *the intensity of an auditory signal at a given location may be determined by many combinations of source powers and source distances.* This is due to the fact that as a sound wave travels outwards, its intensity attenuates and thus, a higher intensity signal (at-the-ear) may be the result of either the sound source having more power *or* the sound source being closer (Bronkhorst & Houtgast, 1999; Coleman, 1962). Furthermore, the environment a sound source is situated in forms a sound field which may also influence the auditory signal we receive at-the-ear (Zahorik, 2002a, 2009). As a soundwave travels outwards it will interact with the physical properties of the space it moves through. Surfaces will reflect, refract and, to some degree, absorb soundwave energy. Therefore, in the presence of reflective sound fields, our ears will not only receive the direct sound wave energy, but also indirect sound energy that has been reflected off other surfaces before arriving at-the-ear. These indirect elements are commonly dubbed 'reverberation', and arrive after the direct waveform (Zahorik, 2002b). Source power, source distance and the sound field are the three primary variables that influence the intensity of an auditory signal at-the-ear. It is also important to note that events in the environment typically do not only produce auditory signals, but also produce *visual* signals. That is to say, when an auditory event occurs, visual signals also often carry information about the distance and likely intensity of the sound source, as well as the sound field in which the sound source exists.

The section above describes how an auditory event can be broken down into a number of causal influences. In the following sections of this Chapter, I will review the research examining the relationship between each causal influence, and how they have been shown to effect (a) the perception of loudness, or (b) the auditory-evoked response elicited by the auditory event. The variable that has had by far the most attention payed to it in psychoacoustic literature has been aural intensity (as influenced by varying the power of the sound source). This will be the first variable I review.

2.3 Signal intensity and audition

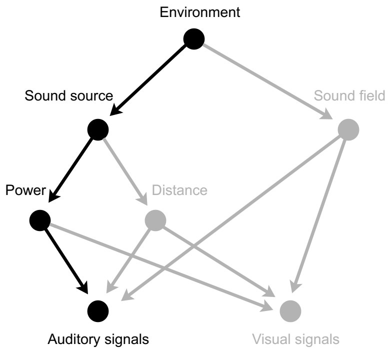

Figure 2.2. Influence graph for an example scenario. The black arrows are the causal influences that are focused on in this section.

2.3.1 Defining the relevant components of a sound wave

Previous research has investigated the perceptual and neurophysiological responses elicited by sound waves with different properties. Because the physical properties of sound waves have technical definitions and, in turn, a specific relationship with our auditory response, I will begin by providing a brief overview of these physical properties.

When an object vibrates its movement will push and pull adjacent particles in the air. This, in turn, will change the pressure in the air immediately surrounding the object. Adjacent particles in the air are pushed such that they are compressed (condensation) and pulled such that they are farther apart (rarefaction), as shown in Figure 2.3A. This condensation and rarefaction of particles causes a chain reaction of vibration that cascades outwards. Here, the pushing and pulling of particles occurs on the same axis in which the sound is travelling (longitudinal wave). There are two key features of a soundwave: (1) its amplitude, which is the change in the peak pressure of the air relative to an atmospheric baseline (measured in pascals: Pa), and (2) its wavelength, which is the length between successive peaks of the soundwave (audible wavelengths approximately range from 17mm to 17m). These features can be illustrated by graphing the pressure changes caused by simple sine waves, as shown in Figure 2.3B. The frequency of a sound wave is conventionally measured in *Hertz*, which describes the number of complete wavelength cycles over a period of 1 second. When a sound causes the movement of particles in the air, this transmits energy, quantified as watts per

Table 2.1. Key definitions extracted from Moore (2012)

Terms	Description
Sound power	The sound energy transmitted per second.
Sound intensity	The sound energy transmitted per second (i.e., power) through a unit area in a sound field.
Decibel (dB)	The dB scale is conventionally used to gauge sound intensity. It is a logarithmic scale expressing the ratio of two intensities, where the number of dB = $10 \log_{10} \frac{I_1}{I_0}$. I_1 represents the intensity of the sound and I_0 represents a reference intensity. The reference intensity most commonly used for sounds in air is 2×10^5 Pa. This intensity is chosen as it is close to the absolute threshold of humans for a 1000-Hz sinusoidal tone. When dB are specified using this reference, it is referred to as dB sound pressure level (SPL).
Phon	A phon is the unit of a sound in dB SPL that is found to be subjectively equal to a 1-kHz reference tone. For example, a sound that is perceived to be the same loudness as a 1-KHz 70 dB SPL tone is 70 phons).
Sone scale	One sone is the unit of a sound that is found to be subjectively equal in loudness to a 1-kHz 40-dB SPL reference tone. Two sones are quantified as double the loudness of the reference tone.
Loudness	The *subjective* intensity of the sound.

square meter. The energy in a sound wave is directly related to the pressure variation caused by particle movement; greater pressure variations transmit more energy. The power of a sound is the energy transmitted in a soundwave per second. Finally, the intensity of a sound is the energy transmitted per second (i.e., power) through a unit area in a sound field (Moore, 2012). A summary of these key terms is provided in table 2.1.

It is crucial to note that there is no simple association between a person's subjective judgement of sound intensity (i.e., its loudness) and objective measures of physical sound intensity. Because of the subjective nature of quantifying loudness, the challenge for psychoacoustic research is to determine what experimental design and auditory measures most 'appropriately' capture the underlying phenomenological experience of loudness. Below, I will delve into some of the hurdles that must be overcome when attempting to measure subjective loudness.

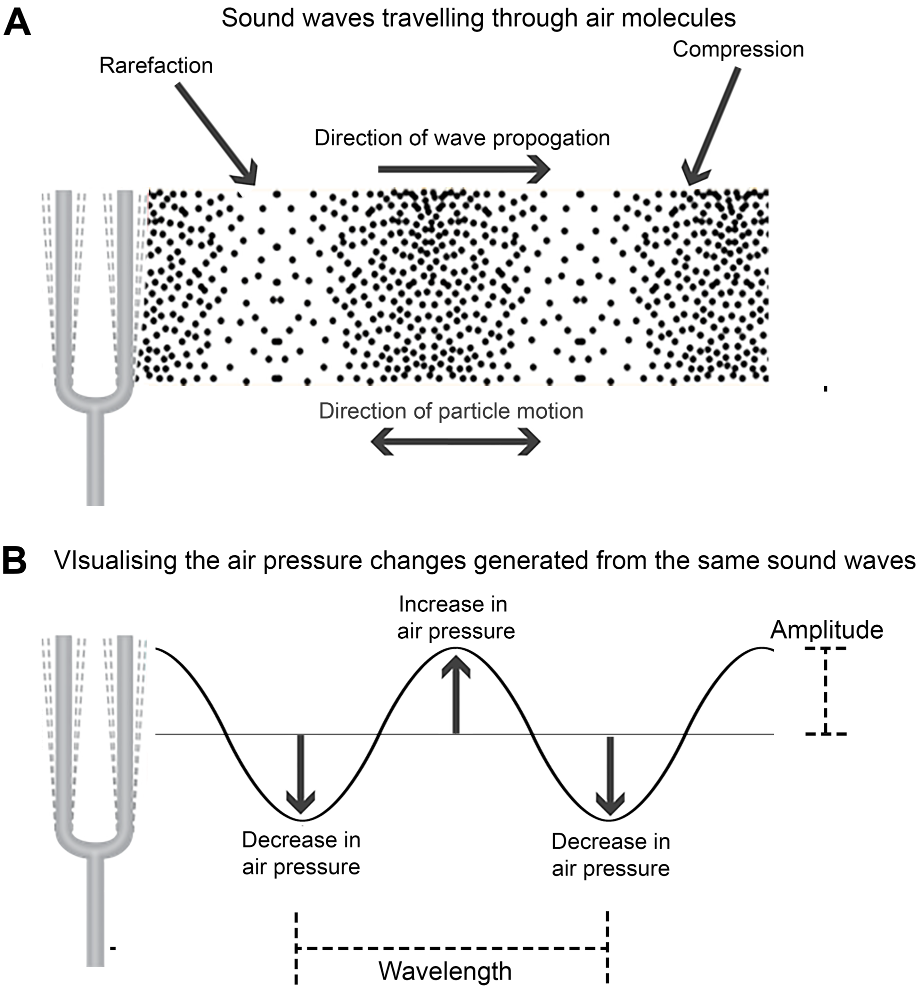
A
Sound waves travelling through air molecules
Rarefaction
Compression
Direction of wave propogation
Direction of particle motion
B
VIsualising the air pressure changes generated from the same sound waves
Increase in air pressure
Amplitude
Decrease in air pressure
Decrease in air pressure
Wavelength

Figure 2.3. **A.** Schematic depicting the movement of a sound wave through air molecules as a tuning fork vibrates. The energy of these sound waves propagates outwards whiles the individual particles oscillate back and forth. **B.** Visualising the air pressure change of a single horizontal slice of area from the sound depicted in panel A. Here the y axis represents air pressure as measured by deviance from an atmospheric baseline (e.g., in Pa). The x axis represents the respective distance (e.g., in mm) from the tuning fork.

2.3.2 The psychophysical measurement of loudness

Humans can pick up an astonishing range of auditory intensities. The weakest sounds we can detect are 1,000,000,000,000 (one trillion) times less intense than the most intense sounds we can receive without damaging our ears. Psychoacoustic research has had to grapple with the problem of measuring loudness not only because of the huge spectrum of intensities that we can receive, but also because *there is no straightforward relationship between physical intensity and perceived intensity.* The measurement of subjective loudness is complicated by at least six factors: 1) behavioral judgments of the intensity of an auditory signal are not directly proportional to the physical intensity of the signal (e.g., Fechner, 1860/1966), 2) it is difficult to reliably quantify perceived loudness in a numerical manner through simple introspection (e.g., Stevens, 1955), 3) semantic definitions of loudness are subjective and can hold different meanings across individuals and cultures (e.g., Florentine, Namba, & Kuwano, 1986), 4) reported loudness is affected not only by sound pressure (i.e., intensity) but also by other physical properties of a soundwave such as frequency (e.g., Suzuki & Takeshima, 2004), 5) environmental context effects can alter loudness ratings (e.g., Berliner & Durlach, 1973; M. Epstein, 2007), and finally - and of particular relevance to my thesis - 6) it is not clear if loudness is solely a function of auditory signals that reach our ear or if it also involves integrating information about causal source intensity (e.g., Norman-Haignere & McDermott, 2018, September 16). As a result of these limitations, there has been great variability in loudness scales that have been reported in the literature. While some loudness metrics are commonly used (e.g., the sone scale), there is no single loudness metric that is universally accepted in the psychoacoustic literature. Each scale that is used to capture loudness is subject to the limitations of the methods used to derive it. Below I will provide a brief summary of the key paradigms that have been used to measure subjective loudness in the field of psychophysics.

A pioneer in loudness measurement: One of the earliest scientific accounts of loudness came from Johann Krüger in 1743 (cited in Marks & Florentine, 2011). Krüger suggested that the perception of intensity was *directly* proportional to the intensity of a stimulus itself. A little over one hundred years later, Gustav Fechner contested Krüger's theory noting that when you double the amount of people in a choir, the average acoustic intensity would be approximately doubled yet the resulting sound was subjectively experienced as being much less than twice as loud (Fechner, 1860/1966). He suggested that the perception of intensity was instead proportional to the logarithm of the physical intensity of the stimulus. Conceptualising loudness through this logarithmic linking function was a seminal leap forward in linking intensity to experience. To this day the most commonly used scale for quantifying a sound's physical intensity, the Decibel (dB) scale, closely resembles Fechner's formula. Fechner (1860/1966) sought to prove his theory by adapting Weber's law that quantified intensity units with the smallest distinguishable differences between stimulus intensities, known as just-noticeable-differences (JNDs). Fechner argued that discrimination thresholds could generate a scale that linked perceived intensity to stimulus intensity. Embedded in this assumption was the notion that each JND unit had the same psychological magnitude. For example, a sound of 4 JND units above the reference would be twice as loud as a sound of 2 JND units above the reference. Using this assumption, Fechner indirectly inferred subjective loudness through JND units. A problem with this approach is there is evidence to contradict Fechner's assumption that JND's provide constant units of loudness when employed across different sound frequencies and in different environmental conditions (Ozimek & Zwislocki, 1996; Riesz, 1933). Although Fechner's theory is of theoretical interest, it has been described as a difficult and impractical method of measurement (Florentine, 2011). Following this landmark account, there have been a plethora of more practical paradigms developed for measuring perceived loudness. As discussed below, the prominent paradigms that have been used to quantify the subjective loudness of an auditory stimulus have been: magnitude estimation, magnitude production, judging ratios, additivity, and magnitude matching.

An estimation-based approach to loudness - the sone scale: The most popular approaches for generating loudness scales have come from overt judgments of loudness ratios between two sounds. These procedures were instrumental in developing the sone scale of loudness, a widely used scale for quantifying the subjective loudness of auditory stimuli. The sone scale was primarily assembled by Stevens and was compiled with the use of both magnitude estimation and magnitude production procedures (Stevens, 1955, 1956). The *magnitude estimation* procedure involved asking participants to compare the loudness of comparison sounds against a reference sound, which was always considered to have a unit of measurement equal to 1. Participants were asked to assign a loudness value to the comparison sound that captured proportionally how much louder or softer the comparison sound felt (e.g., if the comparison sound felt twice as loud as the reference sound, they would assign it with a value of 2). *Magnitude production* procedures involve presenting a reference tone and asking participants to adjust a comparison tone to be proportionally louder by a certain amount (e.g., twice as loud). Stevens's sone scale suggested that perceived loudness was proportional to a power function of the physical intensity of a tone $Sones = K * I^{\beta}$. Here loudness is quantified in sones (2 sones is perceived as being twice as loud as 1 sone), 'k' represents a constant that depends on the participants and units used, 'I' represents intensity (dB SPL) and 'β' represents the power function that transforms intensity into units proportional to loudness. Estimates of this power function β have varied between 0.3 and 0.6 (Ellermeier & Faulhammer, 2000; Moore, 2012; Richardson & Ross, 1930; Stevens, 1975; Zimmer, 2005). Stevens' work also played a role in highlighting how different methods of measurement (e.g., magnitude estimation vs. magnitude production) led to slightly different power functions. These differences drew attention to the dependency of loudness scales on the specific methods of measurement. Nonetheless, the sone unit has served as one of the standard measures of subjective loudness for a little over 70 years.

Loudness matching, a common validity check: *Loudness matching* involves measuring the level in which a comparison sound is judged to be equal volume with a reference sound. The popular unit of measurement in this paradigm is the phon. A phon is the intensity level (in dB) at which a comparison sound appears to be of equal loudness to a 1000 Hz tone. Equal loudness matching has traditionally been used as the benchmark method for verifying the internal consistency of loudness scales. For example, this technique has been used to verify whether the units of loudness in a given scale remain consistent when applied across different auditory stimulus parameters (e.g., across frequencies) (H. Fletcher & Munson, 1933; Marks & Florentine, 2011). A key claim established through loudness matching is that loudness partially depends on frequency as well as sound level: for example, a 100 Hz comparison tone required greater intensity (dB) to be judged as having equal loudness to a 1000 Hz reference tone at 30 dB (H. Fletcher & Munson, 1933). Loudness matching techniques work on the assumption that listeners can identify one dimension of an auditory stimulus - namely, loudness - whilst discounting other dimensions of the stimulus, such as pitch. An important issue that needs to be accounted for in loudness matching paradigms is the potential for the second of two successive sounds to be perceived as louder or softer than the first, depending on the interstimulus (ISI) interval (Hellström, 1979; Stevens, 1955). There is evidence to suggest that longer ISIs can lead to the second sound being perceived as louder, whilst shorter ISI can lead to the second sound being perceived as softer (Hellström, 1979). The loudness matching paradigm has also been suggested to promote a measurement bias known as the regression effect (Stevens & Greenbaum, 1966). Here, listeners tend to overestimate the comparison stimulus at low levels and underestimate it at high levels (Scharf, 1961; Zwicker, Flottorp, & Stevens, 1957).

Beyond this, the marked levels and mechanical characteristics of the mechanism that is used to adjust the volume of the comparison stimulus may also bias results (Guilford, 1954; Stevens & Poulton, 1956). While loudness matching procedures have typically involved using magnitude production procedures similar to those described above, more contemporary approaches have used forced choice adaptive staircase procedures, detailed below.

A comparative approach to loudness matching: A common contemporary approach, based on loudness matching, which has been widely used to quantify subjective loudness percepts is the two-interval forced choice (2IFC) method. This approach requires participants to make a comparative judgement on each trial, such as: *'which of the two intervals contained the louder sound?'*. In this method, the intensity of the comparison stimulus varies across trials, whilst the intensity of the reference stimulus remains fixed. An early example of a comparative 2IFC procedure was called the methods of constants. Here intensity levels of the comparison stimuli were preset, and presented in a random order. Because there is no selective 'honing in' on key areas in the methods of constants (i.e., the intensities at which the comparison and reference stimuli are perceived as being most similar), there are equal amounts of trials in the areas where the comparison intensity is least informative (e.g., where the comparison intensity is much higher than the reference intensity) compared to where it is most informative. Because of this, the methods of constants is inefficient and requires a very large number of trials. In a contemporary context, technology has facilitated the use of adaptive staircases in which the intensity of the stimulus presented on any given trial varies as a function of the participant's previous responses. These approaches often offer superior precision and reliability in sensory measurement, and produce sensitive and efficient estimates relative to non-adaptive testing methods (Leek, 2001). There are multiple types of algorithms that can determine how the comparison stimulus adapts to the participant's response. These range from basic 'up-down' formulas (e.g., if the participant says the comparison was louder than the reference, then the intensity of the comparison tone is reduced on the next trial) (Jesteadt, 1980; Levitt, 1971), to more complex procedures based on maximum-likelihood estimates (e.g., J. L. Hall, 1981; Takeshima et al., 2001). There are also Bayesian methods which determine stimulus levels by updating the posterior probability after each trial, and using this posterior to determine the stimulus level on the subsequent trial such that it maximises the expected information that can be gained on that trial (Kontsevich & Tyler, 1999). The common index of loudness equality in 2IFC methods is generally the point of subjective equality (PSE), which is the point at which there is a 50% probability that the participant chooses the comparison sound as being louder than a reference sound. Although relatively sensitive and efficient, measurement biases have also been found within these 2IFC methods. For example, regression type biases can be an issue: listeners tend to overestimate the comparison stimulus at low levels and underestimate it at high levels in adaptive staircase studies (Florentine, Buus, & Poulsen, 1996). It has also been noted that biases may occur as a result of the order in which the reference stimulus is presented. For instance, one previous study found that the presentation order of a fixed reference stimulus (i.e., whether the reference varied randomly across blocks of trials, randomly within blocks of trials, increased across blocks of trials or decreased across blocks of trials), affected the loudness judgements in the task (Silva & Florentine, 2006). Overall, while there may be biases in adaptive procedures, these can mostly be compensated for by careful consideration of the experimental design and staircase parameters.

Loudness summation - an additivity-based approach to loudness: A different approach was utilised by H. Fletcher and Munson (1933) who suggested that loudness was the result of the additive sum of auditory receptor activity. This led to the development of the additivity theory of loudness. A key assumption here is that a sound delivered monaurally should be half as loud as the same sound delivered binaurally, because half as many auditory receptors would be activated. Based on this assumption H. Fletcher and Munson (1933) used loudness matching to construct a loudness scale by comparing the perceived loudness of sounds presented binaurally vs. monaurally. There was initially some evidence to support this theory (H. Fletcher & Munson, 1933, 1937; Hellman & Zwislocki, 1963; Marks, 1978). However, more recent evidence obtained with a variety of other experimental paradigms suggests that perceived loudness is less than doubled when comparing binaurally vs. monaurally presented sounds (Gigerenzer & Strube, 1983; Scharf & Fishken, 1970; Zwicker & Zwicker, 1991). Specifically, these studies found that binaurally-presented sounds were perceived as being between 1.2 and 1.8 times as loud as monaurally presented sounds. Scharf and Fishken (1970) used magnitude estimation and found that the ratio in question was not 2, but between 1.4 and 1.7. Culling and Edmonds (2007) and Keen (1972) also reported a loudness ratio of 1.4 between monaural and binaural presentations in loudness matching tasks. In short, there is strong evidence that there is not perfect binaural loudness summation. In light of this, it has been suggested that contralateral binaural inhibition may occur, in which input from one ear inhibits input from the other ear (Gigerenzer & Strube, 1983). Contemporary loudness models have taken contralateral binaural inhibition into account. Moore and Glasberg (2007) included inhibitory ear interactions to predict that sounds presented binaurally would be approximately 1.5 times as loud as the same sound presented monaurally. This new model, which incorporates contralateral binaural inhibition, has been found to generate more accurate predictions than the original model which did not incorporate this modification (Moore et al., 1997).

As discussed in the previous section, JND, loudness scaling and loudness matching approaches have been the most influential paradigms used in psychoacoustics for measuring subjective loudness. Across all of these paradigms it has been established that when using basic auditory stimuli (such as pure tones), intensity is the driving force in determining loudness (Fechner, 1860/1966; H. Fletcher & Munson, 1933; Florentine, 2011; Stevens & Poulton, 1956). However, intensity is not the only factor that influences perceived loudness; frequency and other factors such as spectral content, duration, adaptation and the presence of background sounds can also affect perceived loudness to some degree (Florentine, 2011; Suzuki & Takeshima, 2004). These findings have been highlighted by the loudness matching procedure, which is the gold standard approach for comparing the relative loudness of stimuli across different parameters (Florentine, 2011). For example, equal loudness matching established the ISO standard equal-loudness contours (ISO 226, 2003), in which sinusoid sounds at different frequencies were equated in loudness to 1000 Hz tones (phons) (Suzuki & Takeshima, 2004). Scaling procedures led by Stevens also have provided useful 'psychological' scales of loudness that attempt to capture estimates of the subjective intensity of a sound. While multiple attempts have been made to establish a universal model of loudness, the different scales each have their relative strengths and weaknesses, which are determined by the methods used to construct them (as described above). Nonetheless, while it is impossible to directly capture the experience of subjective loudness, these paradigms have provided us with useful information about the loudness equality and/or the rank ordered loudness of different sounds.

2.3.3 The physiological correlates of auditory intensity

The conversion of sound pressure changes in the air into a pattern of neural changes in the brain are the result of multiple mechanical transformations. Soundwaves generate movement that transfer through parts of the outer, middle and inner ear, until the energy is transmitted through the liquid of the cochlea (Pickles, 2013). The basilar membrane of the cochlea is the primary structure responsible for the conversion of sound energy into neural signals. The top of the basilar membrane is covered in hair cells called the organ of corti. The energy carried through the cochlea fluid displaces the inner hair cells of the organ of corti. This movement is mechanically transduced into neural signals in the auditory nerve (Wright, Davis, Bredberg, Ülehlová, & Spencer, 1987). When tiny hairs (stereocilium) on the tip of each inner hair cell flex, action potentials are initiated (Wright et al., 1987). These action potentials translate down towards the auditory nerve, with each inner hair cell being connected to between 10-30 auditory nerve fibers (Pickles & Corey, 1992). The neural impulse travels along the auditory nerve, through the brain stem, midbrain, and ultimately, to the auditory cortex. While we lack a detailed understanding of how sound intensity is parsed by the auditory cortex (Moore, 2012), many studies have demonstrated that the activity level of the auditory cortex is dependent on sound intensity. To this end, some studies have utilised functional magnetic imaging (FMRI), which measures the metabolic response generated from a change of oxygenation in the blood triggered by neural activity (Fox & Raichle, 1986).

2.3.4 FMRI studies of sound intensity

FMRI studies have consistently found that increases in auditory intensity are associated with increases in the volume of activated area in the auditory cortex (Brechmann, Baumgart, & Scheich, 2002; D. A. Hall et al., 2001; Hart, Hall, & Palmer, 2003; Hart, Palmer, & Hall, 2002; Jäncke, Shah, Posse, Grosse-Ryuken, & Müller-Gärtner, 1998; Lasota et al., 2003; Mulert et al., 2005; Röhl & Uppenkamp, 2012). Increases in auditory intensity have also been found to increase the magnitude of BOLD signal in the auditory cortex (Brechmann et al., 2002; D. A. Hall et al., 2001; Hart et al., 2003, 2002; Langers, van Dijk, Schoenmaker, & Backes, 2007; Mohr, King, Freeman, Briggs, & Leonard, 1999; Sigalovsky & Melcher, 2006).

With regards to the anatomical regions most sensitive to changes in auditory intensity: the temporal cortex has commonly been found to be most responsive to changes in sound level, including the transverse temporal gryrus (GTT, or Heschl's gyri, HT, which includes the primary auditory cortex) (D. A. Hall et al., 2001; Hart et al., 2003, 2002; Jäncke et al., 1998; Langers et al., 2007; Mohr et al., 1999; Reiterer, Erb, Grodd, & Wildgruber, 2008; Röhl & Uppenkamp, 2012; Strainer et al., 1997; Thaerig et al., 2008; Woods et al., 2010), the superior temporal gyrus (D. A. Hall et al., 2001; Jäncke et al., 1998; Reiterer et al., 2008; Thaerig et al., 2008) and the planum temporale and the planum polare (S. M. A. Ernst, Verhey, & Uppenkamp, 2008;

Langers et al., 2007; Reiterer et al., 2008). Additionally, the brain stem and midbrain have also been shown to increase their activity in response to increasing sound intensities. Röhl and Uppenkamp (2012) found activation in all levels of the auditory pathway with increasing sound intensity, including the inferior colliculi (IC) and medial geniculate bodies (MGB). Sigalovsky and Melcher (2006) found an increased activation with increasing level in the cochlear nucleus, superior olive, IC, MGB and auditory cortical areas. Whether the auditory cortex follows an amplitopic structure (i.e., whether it is spatially organised such that different cortical neurons are responsive to different auditory intensities), has also been investigated. Results here have been highly equivocal, and thus it is fair to say that the field is yet to resolve whether the auditory cortex follows an amplitopic arrangement (Bilecen, Seifritz, Scheffler, Henning, & Schulte, 2002; Hart et al., 2003; Lockwood et al., 1999; Pantev, Hoke, Lehnertz, & Lütkenhöner, 1989; Sigalovsky & Melcher, 2006).

A handful of studies have also compared how behavioral judgements of loudness track against neural activity as measured with FMRI (D. A. Hall et al., 2001; Langers et al., 2007; Röhl, Kollmeier, & Uppenkamp, 2011; Röhl & Uppenkamp, 2012). D. A. Hall et al. (2001), used FMRI to measure responses to a simple 300-Hz tone at six different sound levels (66 - 91 dB SPL). They compared the extent and magnitude of BOLD activity in the superior temporal gyrus. Loudness was estimated using a model created by Moore and Glasberg (1996), which used auditory input to predict 1) the excitation pattern in the auditory nerve, and in turn, 2) subjective ratings of loudness. When collapsed across stimuli class, the extent and magnitude of the BOLD signal in the superior temporal gyrus had a significant positive correlation with loudness (in phons) but not with intensity (in dB SPL). In another study, Röhl and Uppenkamp (2012) used a broadband pink noise stimulus presented at different levels (20-100dB) to compare the associations between loudness, sound intensity and neural activation of the auditory cortex, inferior colliculi, and medial geniculate bodies (measured with FMRI). Loudness estimates were recorded using a categorical loudness scale that consisted of 11 options that sat between the anchors ‘inaudible’ and ‘too loud’ (Heller 1985). These results revealed that perceived loudness was correlated with BOLD activity in the auditory cortex, but not the inferior colliculi or medial geniculate bodies. This finding provides support for the idea that subjective loudness is represented in the auditory cortex.

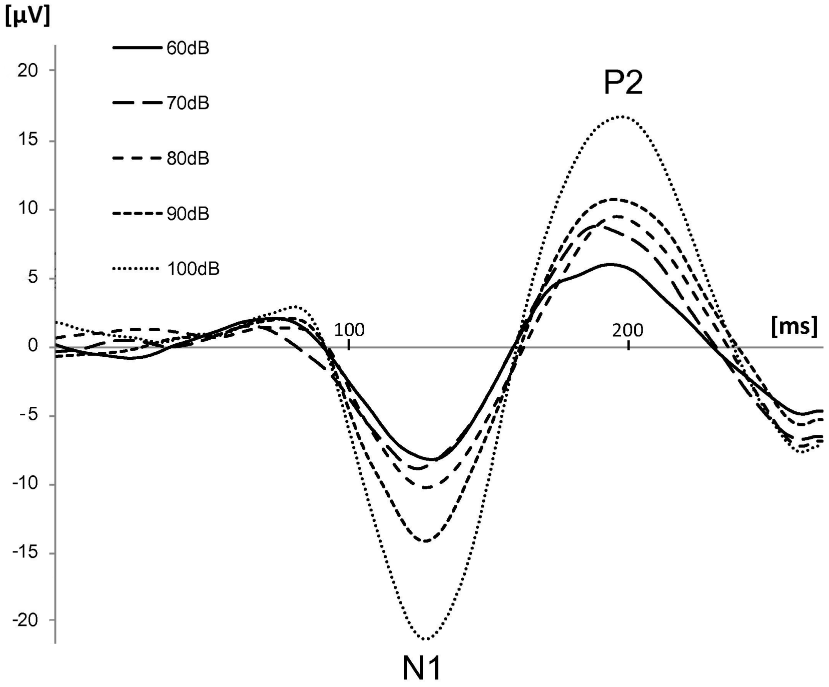

Figure 2.4. The amplitude of the auditory N1 component increases with sound intensity in a positive monotonic fashion. This is an example reproduced from (Hagenmuller et al., 2016). It depicts the auditory evoked potential of a single participant recorded at electrode Cz, which is the electrode at the top of the scalp.

2.3.5 EEG/MEG studies of sound intensity

An alternative tool for measuring the neurophysiological response is elecroencephalography (EEG), and the related technique magnoencephalography (MEG). EEG and MEG are also non-invasive brain imaging techniques which have a temporal resolution that is far superior to FMRI. Both EEG and MEG can measure changes in brain activity with millisecond precision. EEG measures the changes in electrical activity of the brain as recorded at the scalp, while MEG measures the changes in magnetic fields produced by these electrical currents.

Simple sounds, such as pure tones, elicit a characteristic waveform in the EEG known as the auditory-evoked potential. The auditory-evoked potential consists of a number of distinctive peaks and troughs, known as components. By far the most commonly investigated component in previous studies of auditory intensity is the N1 component. The N1 component (illustrated in Figure 2.4) is the first major negative peak in the auditory-evoked potential. The N1 peak occurs approximately 75-125 ms after the onset of an auditory stimulus. The N1 has a fronto-central topography, meaning that it is typically maximal at the vertex of the scalp, around electrode Cz and FCz. The key reason why the N1 component has been investigated in the majority of previous studies of auditory intensity is that the amplitude of the N1 has been shown to be dependent on the physical intensity of sounds (as shown in Figure 2.4) (Brocke, Beauducel, John, Debener, & Heilemann, 2000; Dierks et al., 1999; Hegerl, Gallinat, & Mrowinski, 1994; Lütkenhöner & Klein, 2007; Mulert et al., 2005; Näätänen & Picton, 1987; Reite, Zimmerman, Edrich, & Zimmerman, 1982; Soeta & Nakagawa, 2009; Vasama, Mäkelä, Tissari, & Hämäläinen, 1995). That is, high intensity sounds elicit larger N1 amplitudes than low-intensity sounds. Source localization has revealed that the N1 component of the auditory-evoked potential is generated in the supratemporal plane of the auditory cortex (Godey, Schwartz, De Graaf, Chauvel, & Liegeois-Chauvel, 2001; Näätänen & Picton, 1987; Zouridakis, Simos, & Papanicolaou, 1998). Further evidence that the auditory cortex is the source of the N1 comes from Mulert et al. (2005) who took simultaneous EEG and FMRI measurements. They found a high correlation when comparing the extent of activation of the auditory cortex as measured from FMRI against an EEG-based estimate of the extent of activation of the auditory cortex obtained using a source localization method. Mulert et al. (2005) concluded that there was a close relationship between the sound level dependence of FMRI signal and the amplitude of the N1 component of the auditory-evoked potential.

2.4 The effect of sound source distance and reverberation on perceived loudness

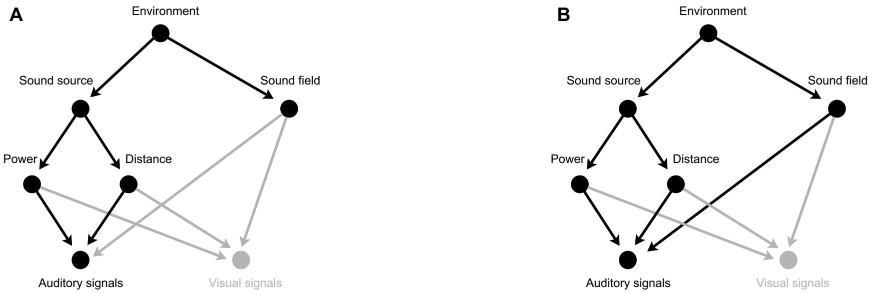

Figure 2.5. Influence graph for an example scenario. The black arrows are the causal influences that are focused on in this section. A) is a representation of an anechoic environment B) is a representation of a reverberant environment.

In the previous section, I summarised research that investigated how aural intensity can affect (a) loudness judgements, and (b) neurophysiological activity. An important common feature of these studies was that they all only considered aural intensity by manipulating the power of the sound source. These studies, did not consider the effect of manipulating the *distance to the sound source.* The distance of a sound source is an important consideration in studies of this nature, as an auditory signal that arrives at a sensory receptor is not only influenced by the power of its source, but also by the *distance between the receiver and the source.* This is because a soundwave will attenuate in intensity as it travels outwards. Specifically, within an optimal free-field environment in which source power is held constant, there will be a 6 dB reduction in sound pressure level every time source distance is doubled (Coleman, 1963). Within reverberant sound fields it is possible for less of a reduction in intensity with distance. Zahorik (2002a) found that within an auditorium which had a reverberant sound field, when intensity was computed based on the entire waveform (based on both the direct and reverberant energy), there was only a 4 dB reduction in intensity with the doubling of distance. Reverberant sound fields often occur indoors where room walls, ceilings and objects reflect sound energy (Mershon & King, 1975). However, outdoor environments that contain reflective objects (such as trees in a forest) can also produce reverberation (D. G. Richards & Wiley, 1980).

In the studies described in section 2.3, loudness was measured simply as the apparent intensity of the auditory signal. However, because of the potential discrepancy between physical intensity at-the-ear and the intensity of a sound at its source, a participant can either be directed to make a loudness judgement based on the signal at-the-ear (i.e., a *proximal* loudness judgement) or based on the estimated intensity of the source (i.e., a *distal* loudness judgement of source power). When investigating our capacity to determine distal source intensity, it has been found that if auditory intensity is the *only* cue available, a listener's judgements as to whether a sound source is 'nearer' or 'louder' are interchangeable (Gamble, 1909). If we are in a dark anechoic room, and we hear a beeping tone increase in intensity, this change could be because (a) the speaker has increased its volume or (b) the speaker has moved closer. Thus, in anechoic conditions, the auditory system cannot disambiguate whether changes at-the-ear have been caused by a manipulation of source power or source distance (Zahorik, Brungart, & Bronkhorst, 2005).

2.4.1 Loudness constancy

If one can judge that the power of a sound source is stable, despite variations in source distance altering the intensity of the signal reaching the ear, then it is said that *loudness constancy* has been demonstrated. It has been suggested that if we display loudness constancy, it may be adaptive in communication (Pollack, 1952; Warren, 1981). For example, we regularly experience others speaking at different distances and as a consequence the speech sounds may vary in intensity at-the-ear. When representing the speech of others, the ability to 'discount' these contextual intensity variations to form a stable representation of the input's content may facilitate speech comprehension. Whilst loudness constancy has not been demonstrated to tones presented in anechoic conditions, auditory signals typically also contain other cues, such as reverberation (Kolarik, Moore, Zahorik, Cirstea, & Pardhan, 2016).

In reverberant sound fields, two studies of particular note have demonstrated that participants tend to display loudness constancy (Altmann et al., 2013; Zahorik & Wightman, 2001). Zahorik and Wightman (2001) required participants to estimate the (a) sound source power, and (b) distance of a sound source, simulated within a reverberant hall. The signal at-the-ear varied in intensity due to both the sound source being simulated at different distances, and with differing amounts of power. Participants were asked to assign the noise burst of each trial with any numerical number to estimate the sound source's power (i.e., estimating the total amount of auditory energy produced by the source at its location). The experiment used a free-modulus estimation procedure (Stevens, 1975) in which no boundaries were placed on the estimation range, and participants could assign any number to the first noise. Despite the intensity of the signal at-the-ear varying due to the sound source being presented at different distances, it was found that listeners were able to relatively identify source power (i.e., participants could relatively identify the different volume settings of the sound source). The ratio of direct-to-reverberant sound energy has previously been found to vary as a function of distance, decreasing as sound source distance increases (Bronkhorst & Houtgast, 1999; Mershon & King, 1975; Zahorik, 2002a). One potential hypothesis explaining the findings of Zahorik and Wightman (2001) is that participants were using this ratio as a distance cue to account for intensity changes at-the-ear caused by the variation of source distance. However, Zahorik and Wightman (2001) also found that participants were systematically biased to overestimate the distance of the sound source at close distances (<1m) and underestimate the distance of the sound source at far distances (>1m). This distance estimation bias when using reverberant cues has been found throughout the literature (Anderson & Zahorik, 2014). Consequently, Zahorik and Wightman (2001) presented an alternate hypothesis: that energy in the reverberant tail facilitated loudness constancy. The energy in the reverberant component of the sound wave (i.e., the total energy arriving at-the-ear after reflecting off surfaces in the space) remained approximately constant with changing source distance. Thus, participants may use the total amount of reverberant energy to determine source power. This theory was supported by a subsequent study by Altmann et al. (2013) who again used free-modulus estimation to acquire estimates of source power from participants. They found that loudness constancy was only present in the strong reverberant condition (i.e., the sound field associated with a slower reduction in reverberant energy) but not the weak reverberant condition. (i.e., the sound field associated with a faster reduction in reverberant energy). This result was observed despite source distance estimates not differing significantly across conditions. These results again indicated a dissociation between source distance and source power estimates. An important point to note here is that, both the studies of Zahorik and Wightman (2001) and Altmann et al. (2013) required the estimation of source power (i.e., the total amount of sound energy emitted from the sound source). It thus remains unclear as to whether reverberant cues also influence apparent loudness estimates of the proximal signal, as estimated at-the-ear.

2.5 The role of visual signals in audition

Up until now we have only considered how the physical properties of the auditory signal can influence its perceived loudness. Yet this may be ignoring an important part of the picture. There are many reasons to believe that visual information may also influence auditory perception. *Often we do not only hear auditory input, but we also see the movement that generates the sound.* In this way, vision can provide complementary information that aids in the disambiguation of auditory input. Vision has been found to influence auditory percepts when visual signals reliably predict the location, the identity and (to a lesser extent) the timing of auditory signals (Alais, Newell, & Mamassian, 2010). Thus, it is somewhat surprising that the influence of visual information on loudness perception and its neural consequences have not been more heavily investigated. *Addressing this issue is the primary aim of the current thesis.* In the following section, I will provide a summary of the conditions under which vision has been shown to interact with audition.

2.5.1 Audiovisual interactions - psychophysics

When audiovisual sensory streams share a certain amount of spatiotemporal synchronicity, they are often bound together as a unitary event (Calvert, Spence, & Stein, 2004). This process is known as audiovisual 'binding'. In the process of audiovisual binding, vision has regularly been found to dominate and 'capture' the apparent location of auditory signals. This is potentially because vision has higher spatial precision compared to audition (Alais et al., 2010). That is, whilst our unimodal estimation of visual distance is relatively reliable and precise, estimation of auditory distance is highly variable, less accurate, and often systematically biased to be overestimated at close distances and underestimated at far distances (Loomis, Klatzky, Philbeck, & Golledge, 1998; Zahorik, 2002a; Zahorik et al., 2005). Anderson and Zahorik (2014) examined localisation precision by asking participants to make distance estimates as to the location of a loudspeaker in a reverberant hall, based on both visual and auditory distance cues. They found that when participants' were provided with either visual-only or audiovisual cues, estimates of the distance of the loudspeaker were more accurate and less variable than when using auditory cues alone. With regards to evidence that visual cues dominate over auditory cues when attempting to locate a sound source, an early example has been provided by the 'proximity-image effect' (Gardner, 1968). In this paradigm, five speakers were placed at five different distances from the participant. These speakers were placed directly in the line of sight of the participant, such that only the closest speaker was visible. Speech sounds were then relayed from the furthermost speaker. Despite this, the closest (and only visible dummy) speaker was consistently perceived as being the actual sound source. Following this, Mershon, Desaulniers, Amerson, and Kiefer (1980) demonstrated that when the only visible (dummy) sound source was placed farther than an actual-yet-occluded sound source, participants' perceived sounds as coming from the dummy speaker. In a similar way, when we attempt to locate a sound source not at a particular depth, but rather on the horizontal or vertical plane, visual information can again influence the perceived location of the auditory signal. This effect underlies the ventriloquist illusion, an illusion in which it is possible to make one's voice appear as if it were coming from a visible prop at a different location (Thurlow & Jack, 1973). When testing the ventriloquist effect, Thurlow and Jack (1973) found that visual cues, such as a video of a speaker moving their mouth, could capture the perceived location of an auditory speech signal that was coming from a disparate angle on both the vertical and horizontal planes.

When visual information provides predictive information about the content of spoken auditory signals it can also influence how the signals are perceived. For example, in the McGurk effect (McGurk & MacDonald, 1976) the perceived content of an auditory signal can be altered with accompanying visual information. In the McGurk effect, an auditory phoneme (such as 'ba') will be perceived differently when it is coupled with a video of someone uttering a different phoneme (such as 'ga'). Specifically, participants often report hearing a novel, 'fused' phoneme (such as 'da'). Additionally, Saldaña and Rosenblum (1993) suggested that the McGurk effect could be extended to non-speech stimuli. They compared the perceived sound of a cello string being plucked vs. the sound of a cello string being bowed. The results showed that discrepant visual information influenced the reported identity of the sounds (i.e., whether they were perceived as being plucked or bowed), although the effect was markedly smaller compared to when speech stimuli were used. Conversely, the presentation of congruent visual information can aid the comprehension of artifact-contaminated sounds. Studies have shown that speech comprehension is higher when videos of the speaker are present, both in impoverished and clear hearing conditions (Erber, 1975; Reisberg, Mclean, & Goldfield, 1987; Sumby & Pollack, 1954). It has been suggested that vision can aid speech comprehension by (a) complimenting the more ambiguous components of auditory signals with additional disambiguating cues, and (b) exploiting corroborative information between the less ambiguous components of auditory signals with correlated cues (Campbell, 2007).

In contrast to the previous studies in which vision was found to influence audition, in audiovisual tasks in which the perceptual judgement depends on temporal information, audition has been shown to influence vision. Whilst vision has a superior spatial resolution than audition, audition has the superior temporal resolution (Burr, Silva, Cicchini, Banks, & Morrone, 2009; Goodfellow, 1934; Irwin, Hinchcliff, & Kemp, 1981; Viemeister & Plack, 1993). In the flash illusion, in which a single flash of light is accompanied with two (or more) beeps that are spaced 57 ms apart, participants often report that they experienced two flashes (Shams, Kamitani, & Shimojo, 2000). Fendrich and Corballis (2001) conducted an experiment where participants were required to judge the position of a rotating clock-hand at the time in which the flash occurred. The flash was offset with a click that occurred 100, 50 or 0 ms before or after the flash. It was found that participants judgements of the clock hand position were shifted towards the position of the hand at the time of the click. Fendrich and Corballis (2001) conducted a second experiment, which was identical except for the fact that they flipped the roles of the click and the flash. A similar result was found: when participants judged the position of the clock hand, the flash introduced a similar (though smaller) attractive capture effect (Fendrich & Corballis, 2001). These biases were dubbed as temporal ventriloquism. In a two-interval temporal order judgement task Morein-Zamir, Soto-Faraco, and Kingstone (2003) asked participants to judge which of the two LED lights (top or bottom) turned on first. They found that when sounds were played just prior to the first flash and just subsequent to the second flash, participants' accuracy was improved. This suggested that the flashes were easier to discriminate with flanking beeps, as they were captured by the auditory stimuli which were temporally further apart. Furthermore, when auditory clicks and visual flashes are presented as fluttering and flickering in synchrony (e.g., at a rate of ten times per second), increasing or decreasing the rate of the clicking sound has been found to respectively increase or decrease the perceived flicker rate of the light. This capture of visual temporal perception has been termed 'auditory driving'. Interestingly, the reverse is not true: changing the flicker rate of a light does not influence the perceived flutter rate of the auditory click (Gebhard & Mowbray, 1959; Myers, Cotton, & Hilp, 1981; Shipley, 1964).

Overall, these studies indicate that multisensory integration typically occurs in such a way that more weight is given to the more reliable sensory modality (Fisher, 1968; Freides, 1975; Howard & Templeton, 1966; O'Connor & Hermelin, 1972; Welch & Warren, 1980). A hallmark study probed this effect by artificially degrading the reliability of visual signals (Alais & Burr, 2004). In this study, participants were asked to judge the location of visual-only, audio-only and audiovisual stimuli. These stimuli were, a) 15ms visual light 'blobs' (projected onto a translucent perspex screen), b) audio 'clicks' (delivered by two external speakers at the edge of the screen; the apparent position of the click was modulated by interaural time differences), and c) an audiovisual combination of both stimuli (in which participants were requested to imagine these inter-modal cues as a unitary event, i.e., a ball hitting the perspex screen). Localisation judgements were gauged using a two-interval forced choice task in which participants had to decide the interval in which the stimulus was perceived as being more towards the left. In the audiovisual condition, one interval contained visual and auditory stimuli that were in spatial conflict (due to being displaced leftwards or rightwards from each other). The resolution of the visual light 'blobs' were also systematically impoverished by altering 'blob' size. For each unimodal condition (i.e., the audio-only and visual-only conditions), precision scores were calculated using the inverse of the variance of location judgements. Using these unimodal precision scores, it was possible to apply a maximum likelihood estimation (MLE) model to predict how a statistically optimal observer would weigh the combination of audio and visual streams, so that variance in judgements would be minimised. It was found that as the quality of the visual stimuli decreased, so too did their dominance in determining source location. Remarkably, participants produced audiovisual judgements that were similar to those which were predicted by the MLE model, suggesting that the average participant weights the relative contributions of audio and visual information in a statistically optimal manner, depending on their relative reliabilities. Optimal integration based on the relative reliability of visual vs. auditory information has also been argued to play a role in speech perception (Shams, Ma, & Beierholm, 2005), and to account for the sound induced flash illusion (Ma, Zhou, Ross, Foxe, & Parra, 2009). Going beyond the audiovisual domain, there have been many accounts that provide further evidence that when integrating multiple sensory cues, participants weigh the input from each sensory mode by its relative reliability in a statistically optimal way (M. O. Ernst & Banks, 2002; M. O. Ernst & Bülthoff, 2004; Knill & Richards, 1996; Knill & Saunders, 2003; Körding et al., 2007; Körding & Wolpert, 2004; Roach, Heron, & McGraw, 2006; Rowland, Stanford, & Stein, 2007; Y. Sato, Toyoizumi, & Aihara, 2007; Shams et al., 2005)

2.5.2 Audiovisual interactions - neurophysiology

In addition to the aforementioned evidence from psychophysics studies, there is also substantial neurophysiological evidence of multisensory integration effects between auditory and visual stimuli. Early FMRI demonstrations of cross modal interactions in the primary sensory cortices found that visual lip reading activated the auditory cortex in the absence of auditory input (Calvert et al., 1997; Pekkola et al., 2005). Direct projections have been identified between areas of the auditory cortex, and the primary visual cortex in primates (Falchier, Clavagnier, Barone, & Kennedy, 2002; Rockland & Ojima, 2003). Primate studies have also shown superior temporal sulcus as having bidirectional connections with the visual cortex, and to contain multisensory neurons responsible for coding both the sight and sound of facial movements, hand actions, and body movements (Barraclough, Xiao, Baker, Oram, & Perrett, 2005; Kaas & Collins, 2004). Similarly, hemodynamic studies have shown that areas such as the superior temporal gyrus in the auditory cortex is a multisensory area in humans (Callan et al., 2004; Calvert, Campbell, & Brammer, 2000).

An increasingly popular way to examine audiovisual interactions has been via the use of EEG and MEG. This is because these two measurement tools have high temporal resolution and are effective at capturing interaction effects in real time. The additive model assumes a multisensory interaction has occurred when the neurophysiological response to bimodal stimulation (i.e., audio and visual) differs from the sum of the activation from each unimodal response (Barth, Goldberg, Brett, & Di, 1995). There is a large body of ERP evidence to suggest multisensory interactions can occur as early as 100ms post stimulus (Alais et al., 2010). Videos containing anticipatory motion temporally predicting the onset of an auditory signal have been shown to have a sub-additive effect, reducing the amplitude of the auditory N1 component elicited by the sound (Vroomen & Stekelenburg, 2010). This attenuation effect has been found when visual cues are used to signal the temporal onset of a variety of sounds, including ecological speech stimuli (Van Wassenhove, Grant, & Poeppel, 2005), non-speech stimuli - such as hand claps (Stekelenburg & Vroomen, 2007) and abstract shapes colliding (Vroomen & Stekelenburg, 2010). There is also evidence that the N1-amplitude is reduced when visual signals predict the onset of a sound at a specific spatial location (i.e., visual prediction of an auditory stimulus at a central location and centrally located auditory stimuli), as compared to when the sound was presented at an unpredicted location (i.e., visual prediction of an auditory stimulus at a central location and laterally located auditory stimuli) (Stekelenburg & Vroomen, 2012).

These studies provide converging evidence to support the notion that when the brain has visual information that accurately and reliably predicts forthcoming auditory input, there is a multisensory interaction which results in the reduction of the amplitude of the N1-component of the auditory-evoked potential (Hughes, Desantis, & Waszak, 2013). Why do these N1 amplitude reductions occur, and what is the functional interpretation of this effect? One theory is that having predictive information about a forthcoming sound reduces signal uncertainty, which consequently reduces the computational demands on the auditory processing regions of the brain (Besle, Fort, Delpuech, & Giard, 2004). This account is broadly consistent with predictive coding theories of perception (Clark, 2013; Friston, 2005b, 2009). Predictive coding accounts of perception argue that the brain processes the sensory inputs it receives in conjunction with top-down predictions about the inputs it *expects* to receive. According to this account, the differences between (top down) predictions and (bottom up) sensory input reflect prediction errors, which requires additional processing and is associated with increased neural activity (Friston, 2009). Neural activity is therefore relatively suppressed when (bottom up) sensory input matches the (top down) prediction. When stimuli increase in their predictability across other multisensory paradigms (e.g., experiments that compare temporally predictable self-generated vs temporally unpredictable externally generated sounds), it is been found that there is also reduction in the auditory N1 amplitude to the more predictable stimuli (i.e., self-generated stimuli) (Hughes et al., 2013). Given that the amplitude of the auditory N1 is intensity dependent (as discussed previously; see Mulert et al., 2005), it has been suggested that the phenomenon of N1-amplitude reduction reflects more predictable sounds being processed as though they were physically softer than the less predictable sounds (Hughes et al., 2013). While there is now a substantial body of research reporting sensory attenuation in the context of self-generated actions predicting auditory input (e.g., Blakemore, Wolpert, & Frith, 2000), the question of whether external visual information that predicts auditory input also causes sensory attenuation has received less attention. One relevant study investigated the effect of either pushing a button to generate a sound or *observing* a button being pushed to generate a sound, on the perceived loudness of that sound. It was demonstrated there was a reduction in the perceived intensity of sounds in both the push and observation conditions, relative to a condition in which there were no visual cues as to the temporal onset of the sound (A. Sato, 2008). This result provides evidence that visual cues about an upcoming auditory event can reduce the perceived intensity of the auditory event.

2.6 Visual signals, auditory processing and loudness

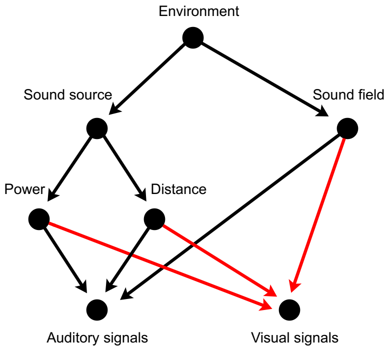

Figure 2.6. Influence graph for an example scenario. The red arrows are the key relationships investigated within this section *and* more broadly are the focus of this thesis.

In the previous section, I have outlined the impact that visual information can have on auditory perception. I have shown how the behavioral interaction of audio and visual information may be integrated in a manner that weighs the relative reliability of each sensory modality. This weighting process is thought to minimize the uncertainty around the representation of the 'what', 'where' and 'when' of the sensory event. In regards to the current thesis: if our auditory system aims to generate optimal percepts of external events, vision is well situated to influence the perception of auditory intensity. We often see a visual event tied to the causation of a sound at a particular level. For example, imagine you *witness* an event happen across the street, either a giant explosion of a gas main, or someone lighting their cigarette with a cigarette-lighter. Similarly, imagine you are in a soccer match and you either *see* someone with a wide open mouth desperately shouting for the ball while they are either standing a metre in front of you or on the far side of the football field. In all of these examples, visual signals carry information regarding the anticipated intensity of the auditory signal. Utilising visual information may increase the precision in which an auditory event can be represented. As depicted in Figure 2.6, visual signals can carry causal information about sound source power, source distance, and the sound field in which the source exists. The role of this visual information in representing the intensity of a sound has rarely been considered. In the section below, I summarise the existing evidence for this proposed association.

2.6.1 Visual cues have the capacity to provide information about the distance of the sound source

Before engaging with evidence that relates visual distance cues to the perceived loudness of sounds, I will first establish that visual signals can provide information about the distance of objects (or sound sources). While the external world exists in three-dimensional (3D) space, we receive all visual signals as two dimensional (2D) projections on our retina. Despite this, the human perceptual system is able to effectively transform these 2D images into 3D representations (Howard & Rogers, 2002). With the aim of understanding this process, vision researchers have examined the accuracy of depth perception and the mechanisms that support it. Vision is considered to be the most precise sensory modality for spatial localisation (Alais et al., 2010) and unsurprisingly, distance estimates informed by vision have been found to be accurate and reliable (Anderson & Zahorik, 2014; Da Silva, 1985).

In depth-measurement paradigms, estimates of distance may be measured based on the egocentric distance of an object, which is the distance from the object to the self, or the exocentric distance of objects, which is the relative distance between two objects (Foley, 1980). Multiple studies have suggested that the perceived egocentric distance of a target (D') varies as a power function (n) of its physical distance (D), multiplied by a constant (C) in the following formula D'=CD^n (Da Silva, 1985). Across a range of environments, it was found that the perceived distance of a target could be predicted on average by raising the physical distance of the target to an exponent that was close 1 - an almost linear linking function (Da Silva, 1982; E. J. Gibson & Bergman, 1954; E. J. Gibson, Bergman, & Purdy, 1955). This has been demonstrated in open fields (e.g., for a review see Da Silva, 1985) and laboratories (Collins, 1976; M. Cook, 1978). Levin and Haber (1993) provided evidence that this almost linear function may even hold at further distances. They required participants to verbally estimate distances of stakes placed up to 40 ft away on a flat grassy area. Here subjects made accurate distance estimates that were linear (with an exponent of 1). Likewise subjects were equally accurate when judging exocentric distances as they were for egocentric distances. In contrast, there is a small amount of evidence that as the distance between the self and an object grows larger, people tend to underestimate egocentric distance (Gilinsky, 1951; Loomis, Da Silva, Fujita, & Fukusima, 1992). For instance, in a natural setting Loomis et al. (1992) found that when asked to match the depth on the ground plane of an interval, with a length provided on the frontal plane, participants estimated the equivalent interval as longer on the ground plane relative to the frontal plane. A caveat to this finding is that the mean exponent linking physical distance to perceived distance will slightly vary as a function of the psychophysical task (Baird, 1970; S. P. Rogers & Gogel, 1975), instructions (Carlson, 1977) and distances ranges (Da Silva, 1985).

Similar depth exponents have also been attained using photographs. Anderson and Zahorik (2014) measured participants egocentric depth estimates of a speaker based on either auditory, visual or audiovisual signals. An HDTV monitor was used to present photographs of a speaker at different distances in a hall. Participants were allowed to estimate distance in units of meters or feet and the presentation distances of the speaker ranged from 0.3 to 9.8 m. In this experiment participants' visually based estimates of distance was linked to actual physical distance with the exponent 0.92. This exponent was very close to the exponent that was found estimating distance in real world environments (Da Silva, 1985). It is however noteworthy that in other instances distance may be underestimated when displayed though virtual head mounted displays (Thompson et al., 2004) and when viewing photo realistic stimuli on large screen display systems (Klein, Swan, Schmidt, Livingston, & Staadt, 2009; Renner, Velichkovsky, & Helmert, 2013).

It has been established that distance estimates can be relatively accurate, but what visual cues do we use to support such estimates? Visual depth cues are typically categorised as being monocular and binocular. Monocular cues can by extracted with one eye, whilst binocular cues necessarily involve both eyes. Monocular depth cues may

Table 2.2. Key monocular cues extracted from Howard and Rogers (2002)

Terms	Definition
Monocular cues: pictorial	
Linear perspective	The lines of parallel objects that appear to converge towards a vanishing point as they approach the horizon.
Foreshortening	The apparent compression of lengthwise surfaces as they increase in egocentric distance.
Texture gradient	The texture of surfaces appear to become increasingly fine as they increase in egocentric distance; this is a result of the pattern on the surface varying in both the orientation and size as a function of distance.
Relative size	When viewing two objects of constant physical size, the object that is located at a closer egocentric distance will generate an retinal image that is larger.
Relative height	When two objects are located on a flat ground plane, the object that is further away will occur at a higher position in the visual field.
Occlusion	When one object appears to partially obscure or overlap over the outline of another object.
Monocular cues: motion	
Motion parallax	Objects moving at a constant speed will move across the visual field more quickly when they are close by relative to when they are further away.

also be further specified as being either pictorial or motion based cues. Pictorial cues are the static elements within an image that generate a sense of depth, whilst any moving components are motion cues. Definitions of the primary monocular cues can be seen in Table 2.2. It is important to note that these monocular parameters are invariant to scale; they specify the relative layout of the objects in a scene independent of absolute distance (Loomis & Philbeck, 1999). That is, if the relative proportions remain constant, smaller objects that are closer together may present an identical image on the retina when compared to larger objects that are proportionately further away. Binocular cues distinctly require both eyes and hold the potential to provide absolute distance estimates. The primary binocular depth cues are detailed in Table 2.3.

Linear perspective and foreshortening (J. J. Gibson, 2014; Hendrix & Barfield, 1995), texture gradient (J. J. Gibson, 1950; Todd & Akerstrom, 1987), relative size (Surdick, Davis, King, & Hodges, 1997), relative height (Dunn, Gray, & Thompson, 1965; W. Epstein, 1966), occlusion (Chapanis & McCleary, 1953), stereopsis (Julesz, 1986; Mayhew & Frisby, 1981), convergence (W. Richards & Miller, 1969), and motion

Table 2.3. Key definitions extracted from Howard and Rogers (2002)

Terms	Definition
Binocular cues	
Vergence	The simultaneous shifting of either eye in opposite directions to fixate on something at a specific egocentric distance. Convergence is the movement of either eye inward when fixating on a close target and divergence is the movement of either eye outward when fixating on a target further away. Vergence eye movements provide absolute distance information, i.e., they are depth cue that is not invariant to scale.
.	
Stereopsis	Because each eye is located at a different position they receive slightly different visual input (binocular disparity). This difference may be used as depth cue. Objects that are closer than the point of fixation (and the horopter) will have their image displaced in opposite directions on each eye's fovea, whilst objects that are further away than fixation (and the horopter) will have their image displaced in the same direction on each eye's fovea.

parallax (E. J. Gibson, Gibson, Smith, & Flock, 1959; B. Rogers & Graham, 1979) have individually been demonstrated to affect our perception of an object's depth. It is important to note, that perspective cues (i.e., linear perspective, foreshortening and texture gradient) only disambiguate the depth of a specific object effectively when that object is linked to a ground intercept. Otherwise the object may appear to be floating and in this situation its height may be confounded with its relative distance (DeLucia, 1991; Kim, Ellis, Tyler, Hannaford, & Stark, 1987). Whilst it has been well established that each of these monocular and binocular cues can individually affect a sense of an objects depth, what has been less well established is the relative effectiveness of each cue. The studies that have attempted to determine which depth cues are most effective have found that perspective cues (i.e., foreshortening, linear perspective, and texture gradient) that incorporate ground intercepts and stereopsis are the most effective individual cues (e.g., Barfield & Rosenberg, 1995; Kim et al., 1987; Ritter, 1979; Surdick et al., 1997). It has even been found that the presence of drop-lines (a perspective cue that provides a ground intercept) were the single biggest factor increasing the accuracy with which participants may estimate an object's depth (Hendrix & Barfield, 1995).

In some instances stereopsis has demonstrated equivocal effectiveness to perspective cues in facilitating the discrimination of an objects depth (Kim et al., 1987; Surdick et al., 1997) and in other instances the addition of stereo cues increased the reliability and precision of depth estimates (Allison, Gillam, & Vecellio, 2009; McCann, Hayhoe, & Geisler, 2018). Although stereopsis has been demonstrated to be an effective cue, it relies on retinal disparity which decreases as viewing distance increases (Davis & Hodges, 1995; Ritter, 1979). Because of this it is thought that at near distances perspective cues are relatively less useful than stereo cues and at further distances stereo cues are relatively less useful than perspective cues (Howard & Rogers, 2002). Surdick et al. (1997) found that when participants were provided with stereo cues and viewing distance increased from 1m to 2m, the ability to discriminate depth was significantly reduced. However, binocular cues may still aid distance discrimination at further distances (Allison et al., 2009; McCann et al., 2018). Using scenes depicting the University of Texas, McCann et al. (2018) reported that while perspective and binocular cues both produced effective distance discrimination, the addition of binocular information improved distance discrimination up to 15m.

Although the investigation of individual depth cues have proved useful, in the natural world it is rare for a visual scene to only convey a single depth cue in isolation. Künnapas (1968) found that as more depth cues are added to a stimulus display, the sense of 3 dimensional depth increased, and generally this also increased the consistency and accuracy with distance judgements. When we view multiple depth cues and they are signaling consistent information, there is evidence that we will fuse this information together to form a unitary depth percept. Such a fusion process appears to produce increasingly reliable and accurate distance estimates (Bruno & Cutting, 1988; Landy, Maloney, Johnston, & Young, 1995). It has also been suggested that in this cue combination process, the influence of each cue (disparity, texture, motion parallax, etc.) is weighted by its reliability and combined in a statistically optimal way (Hillis, Watt, Landy, & Banks, 2004; Kersten et al., 2004; Knill & Saunders, 2003; Landy et al., 1995; Svarverud, Gilson, & Glennerster, 2010; Yuille & Kersten, 2006).

Lastly there are multiple factors related to artificial viewing conditions that may affect distance estimates. Firstly, the field of view can influence judgements (Creem-Regehr, Willemsen, Gooch, & Thompson, 2005). Wu, Ooi, and He (2004) collected distance estimates of objects between 3-8m and found that restricting the field of view to less than 30Âř impaired distance judgements. It has also been found that the accuracy of absolute distance judgements are impaired if there is a disruption of visibility on the ground plane because of a restricted field of view (Sinai, Ooi, & He, 1998; Wu, He, & Ooi, 2007; Wu et al., 2004). Finally, the vantage point may also influence distance perception. There are two prominent theories on how our visual system handles the differing views of 2D objects depicting 3D scenes; experiences that may occur when viewing objects such as a computer monitor or a painting from different locations. The 'vantage-point compensation' hypothesis puts forward that our perceptual system is able to compensate for a shift in vantages point, to view a scene from the incorrect vantage point with any geometrical distortions transformed out (Hagen, 1974; Pirenne, 1970; Vishwanath, Girshick, & Banks, 2005; Yang & Kubovy, 1999). Alternatively, the perspective-transformation hypothesis puts forward that different vantage points will be perceived differently in accordance with the spatial changes from the varying vantage points (Todorović, 2008, 2009).

2.6.1 The potential role of visual cues regarding the (a) distance of a sound source and (b) the sound field of an environment, on the perceived loudness of a sound

Whilst it has been established that loudness constancy occurs within reverberant sound fields (as reviewed in section 2.4), it has not yet been determined whether visual depth cues can also facilitate loudness constancy. If we use causal information to estimate the intensity of an auditory signal - either proximally (i.e., at-the-ear) or distally (i.e., at-the-source) - improving our ability to localise the distance of a sound source may improve the ability to discount variations in auditory intensity induced by distance. We know that distance estimates are more reliable and precise when based on visual as opposed to auditory cues (Anderson & Zahorik, 2014; Loomis et al., 1998; Zahorik, 2002a; Zahorik et al., 2005). If our perceptual systems aim to facilitate loudness constancy, then visual information is well positioned to aid the process of discounting the varying intensity of auditory signals at-the-ear due to the delivery of sounds from different distances.

A handful of studies have investigated the influence of visual cues to source distance on perceived loudness. The findings have been mixed. Mohrmann (1939) utilised a method of adjustment in which participants were required to alter the intensity of a close reference speaker to approximate the intensity of a comparison speaker shown at different distances. Participants were asked to estimate the intensity of the comparison speaker distally at-the-source, and proximally at-the-ear. These judgements were taken with the speaker either in view or in darkness. It was found that loudness constancy (i.e., the perception of stable source loudness across variations in source distance) was highest when estimating distal intensity, and when the speaker was visible. However, a degree of loudness constancy was also demonstrated when participants attended to the proximal intensity at-the-ear, and when the speaker was in darkness. The partial constancy displayed in darkness suggested that reverberant cues may have influenced loudness estimates throughout the experiment. von Fieandt (1951) and Shigenaga (1965) employed a similar method of adjustment, in which participants adjusted the intensity of a close visible speaker to approximate the intensity of a farther visible speaker at its location. Loudness constancy was demonstrated in both experiments; however, only von Fieandt (1951) examined loudness constancy when both visual cues to source distance were provided and when they were not (i.e., loudness estimates performed in the dark). The contribution of vision in von Fieandt's (1951) experiment was difficult to assess as performance was similarly close to complete constancy in both the visual and non-visual conditions. Because auditory distance cues such as reverberation were not controlled for in these two experiments it is hard to disentangle how they may have influenced loudness estimates. Mershon, Desaulniers, Kiefer, Amerson Jr, and Mills (1981) controlled for auditory distance cues by delivering sounds from a hidden source at a fixed location, while a dummy loudspeaker appeared to be delivering the sounds at different distances. Loudness estimates of sound received from the dummy loudspeaker were made through free modulus estimation. Participants were not specially directed to estimate the sound at-the-ear or at-the-source. In the reverberant sound field, participants estimated loudness as increasing as the apparent distance of the source increased. In the anechoic sound field, there was a similar trend of loudness estimates increasing over distance, however at the farther most distance there was a slight reversal of this effect and loudness estimates decreased. It was suggested that this reversal was possibly a result of a failure of the farther most speaker to visually capture the auditory signal due to location incongruities. This is because, in anechoic conditions, sounds can appear closer than in reverberant conditions (e.g., Butler, Levy, & Neff, 1980).

More recently, Altmann et al. (2012) re-examined the effect of visual cues on loudness constancy. In this study, participants heard short bursts of noise delivered via headphones in dark, anechoic conditions. These bursts of noise were paired with the offset of a light source at varying distances. Loudness estimates were taken with a free modulus estimation procedure and no instructions were provided as to whether participants should estimate the proximal or distal loudness. Participants showed no evidence that loudness estimates were influenced by the distance of the light offset. While this result is in contrast to previous findings that demonstrated a visual contribution to loudness constancy, it is unclear if participants were provided with adequate depth cues to accurately perceive the distance of the light source. Neural signals were also measured using MEG. The results revealed that the sounds evoked a larger N1m amplitude when they were presented at the furthest distance, as compared to the closer distances. Surprisingly, this indicated that the neural processing of auditory signals may have been affected by the visual distance information, despite the lack of behavioural loudness constancy. Following this, Berthomieu, Koehl, and Paquier (2019) also investigated the influence of viewing the depth of a sound source, however, they attempted to improve the availability of depth cues through the use of a virtual reality environment. Noise bursts were generated from a virtual sound source (a loudspeaker) to resemble being delivered in (a) a sports hall, (b) a concert hall, and (c) anechoic conditions. The speaker was simulated in a virtual reality room at 5 different distances ranging from 1 to 16 m. In one condition the speaker was visible. In the other condition the view of the speaker was obstructed by a barrier. Loudness and distance estimates were again taken with a free modulus estimation procedure. It was found that participants' judgements were not influenced by visual cues to sound source depth. Whilst there was initial support for visual depth cues influencing *behavioural* loudness constancy, the more recent work of Altmann et al. (2012) & Berthomieu et al. (2019) has brought these findings into question. Further, *if* loudness is affected by inferring a source's distance, it is not clear whether it affects the perception of the proximal signal, the distal signal, or both.

2.6.2 The potential role of visual signals cuing source intensity on loudness

For a sound to be generated, physical objects have to vibrate to create pressure deviations in the surrounding air particles that cascade out as soundwaves. There is a direct relationship between the power of an object's physical vibrations and the intensity of the resultant soundwave (Moore, 2012). For example, the light tap of a gong, will result in the production of a sound at a relatively lower intensity, than a vigorous strike of the same gong. Our visual system often provides us with cues about the relative power of the collisions we see (e.g., a light tap vs. a vigorous strike), and these cues, in turn, provide information about the intensity of the sound we expect to be elicited. Whilst there has been significant interest in how visual information about the *location* and *identity* of a sound source can influence its auditory representation (e.g., Alais et al., 2010), there has been almost no attention paid to the question of whether visual information about the expected intensity of a sound source can influence its perceptual and neural representation.

Only one previous study has directly manipulated visual information regarding the power of a sound source and explored its effect on perceived loudness. Using both speech and non-speech (clapping) stimuli, Rosenblum and Fowler (1991) required people to (a) rate the amount of perceived effort put into the generation of a sound and to (b) rate the loudness of a sound when paired with the same visual cues. Speech or clapping videos depicting an actor putting 1 of 4 possible levels of effort were paired with clap or speech sounds that varied in intensity (between 47dB and 57dB). When auditory stimuli were paired with a video of an actor who was subjectively rated to be putting in more effort, loudness ratings increased compared to when the actor was rated as putting in less effort. It is important to note that it is possible that response bias influenced Rosenblum and Fowler (1991). This is because if a participant cognitively expects a sound to be louder, the task demands may (consciously or unconsciously) influence their task responses. Thus, whilst this investigation provides initial support for the notion that visually provided source power cues can influence perceived loudness, further investigation is needed to determine whether this effect remains after minimizing response bias.

2.7 Research Motivation

In this Chapter, I have attempted to introduce a framework in which the representation of signals received at sensory receptors are related to their causal influences. I have established how the manipulation of sound source intensity, sound source distance and the sound field itself can affect subjective loudness and its neural representations. Following this, I have outlined how visual signals that carry information about the identity, location and timing of auditory events, can influence the perceptual and neural representations of these auditory events. Finally, I have explored how visual signals that carry information about the *intensity* of auditory events also hold the potential to influence auditory processing. This background provides the motivation for four studies that form the research component of this thesis. These studies are united in their the aim of determining whether causal information may influence the perceptual and neural representations of auditory intensity.

To begin the experimental components of this thesis, Chapters 3 and 4 explore the possibility that visual distance cues provide information that may aid the representation of source intensity. To do this, I used a psychophysical 2-interval forced choice (2IFC) method to attain loudness estimates from healthy human participants while they viewed a simulated speaker which relayed sounds at different distances. Given that the proximal auditory signal (i.e., the signal at-our-ear) is a function of the distance between a sound source and ourselves, visually cuing source distance should help to provide a functional representation of source intensity. In Chapter 3, I explore how *apparent* loudness may be modulated by simulating a speaker relaying sounds at different distances, in an anechoic open field. In Chapter 4, I explore how *source* loudness estimates may be modulated by viewing a speaker relaying sounds at different distances, in a reverberant concert hall. In the latter chapter, I attempt to increase the ecological validity of the stimuli by using real-life recordings taken from within this hall.

Chapters 5 and 6 explore the possibility that providing visual information about the expected power of a sound source can influence subjective loudness and the auditory evoked response. In both chapters, intensity expectations were created by means of a video clip which displayed an actor producing either a vigorous (‘strong’), or a non-vigorous (‘weak’) handclap. In Chapter 5, I used a 2-pair forced choice paradigm to estimate whether the perceived loudness of the handclap was altered by concomitant visual information regarding its expected auditory intensity. In Chapter 6, I used a similar design to explore whether visual cues to source intensity affected the neurophysiological response to the sound. The neurophysiological response was operationalized as the N1 amplitude of the auditory-evoked potential, as measured with EEG.

3 Visual Cues to Source Distance & Loudness (anechoic conditions)

Title: Loudness judgements are not necessarily affected by visual cues to sound source distance

Author contributions:
Conceptualisation: SL, TJW, DJM. Stimuli: DJM. Methodology: SL, TJW, DJM. Programming: DJM. Data collection: SL. Data analysis and presentation: SL, DJM. Writing - original draft: SL, DJM, Writing – review and editing: SL, TJW, DJM. Supervision: TJW, DJM.

Preamble:

As outlined in Chapter 2, the intensity of auditory signals at the ear depends on both the power of the sound source and the distance of the source from the listener. I begin the experimental component of this thesis by exploring whether visual cues to the *distance* of a sound source influence loudness estimates. In particular, I hypothesised that visual information would facilitate loudness constancy, which is the ability to perceive stable loudness despite variations in the distance of a sound source causing variations in the intensity of the signal at-the-ear. Previous experiments have identified a visual contribution to loudness constancy, but these studies did not control for the potential influence of reverberant acoustic cues. This is a limitation as it has been shown that reverberation can influence loudness estimates. In this experiment I attempted to isolate the effect of visual cues to source distance on perceived loudness by removing reverberant acoustic cues that can confound loudness estimates. I did this by designing the experiment in anechoic conditions.

3.1 Abstract

One factor that will influence the intensity of an auditory signal is the distance it has travelled from its source. The further a signal travels, the more its intensity will reduce. Loudness constancy requires that our perception of sound intensity, loudness, corresponds to the source power by remaining invariant to the confounding effects of distance. Here, we assessed the evidence for a potential contribution of vision to loudness constancy through the disambiguation of sound source distance. We presented participants with a visual environment, on a computer monitor, which contained a visible loudspeaker that appeared at a variable distance. This was accompanied by the delivery, via headphones, of an anechoic sound of a variable aural intensity. We measured the point of subjective loudness equality for sounds associated with loudspeakers at different visually-depicted distances. We report strong evidence that loudness judgements were closely aligned with the aural intensity, rather than being affected by the apparent distance of the sound source conveyed visually. Similar results were obtained across different sounds and under different presentation conditions. We conclude that the loudness of anechoic sounds are not necessarily affected by visual information about the distance of the sound source.

Introduction

The task of perception is to extract useful information about the environment from the signals available to the sensory receptors. Achieving this task often requires abstraction across the sensory effects of environmental features that are not task-relevant. For example, perceiving the physical size of an object based on visual signals requires consideration of the effect of object distance on the extent of retinal stimulation. The stable perception of environmental features, such as object size, across non task-relevant effects is termed perceptual constancy.

Here, we investigate constancy in the perception of the power of a sound source (i.e., capacity of a sound source to produce acoustic energy — the perception of its 'loudness'). The power of an emitting object is a feature of the environment that can aid in fundamental processes such as recognition and identification (Bizley & Cohen, 2013). However, because sensory signals are affected by other properties of the environment, the power of a sound source cannot be established solely from examination of its direct effects on sensory receptors.

The challenge of loudness constancy can be illustrated by considering the generative process that produces auditory signals at-the-ear. The sound source has two key properties of current interest—its power (i.e., its capacity to produce acoustic energy) and its distance (i.e., its location in the environment, relative to the perceiver). Importantly, both of these properties combine to affect the auditory signals received at-the-ear of the perceiver. This combination means that accurate estimates of the power of the source cannot be obtained solely by evaluating the intensity of the aural signals, because such aural signals could be produced by any pairing of source power and distance (Bronkhorst & Houtgast, 1999; Coleman, 1962). The outcome of this may be that aural intensity experienced alone is not a functionally useful cue; as stated by Worden (1971, p. 22), "intensity level, per se, holds little biological significance". For example, a given aural intensity could be produced by a sound source with high power at a far distance or a sound source with low power at a close distance.

The contribution of source power to auditory signals can be disambiguated if the distance of the sound source is identified. A potential way in which source distance can be estimated is via the prevailing sound field, which interacts with the sound produced by the source to affect the signal at-the-ear. This sound field encompasses aspects of the environment such as the geometry and acoustic properties of surfaces and media, which interact with the emitted sound waves and affect the received auditory signals. Importantly, such interaction can produce cues to the sound source distance in the auditory signals, such as the ratio of direct to reverberant energy (see Zahorik et al. 2005 and Kolarik et al. 2016, for reviews of auditory cues to distance). This estimate of sound source distance, combined with knowledge of how sound source distance and sound source power interact in producing the auditory signals at-the-ear, would allow the sound source power to be disambiguated and become independent of distance. However, Zahorik and Wightman (2001) found that loudness constancy was evident in situations where accurate source distance judgements were not obtained. Zahorik and Wightman (2001) proposed that loudness judgements can instead be based on the energy of the reverberant component of the auditory signals, which can remain largely invariant to sound source distance in reverberant sound fields. Supporting this proposal, Altmann et al. (2013) found that loudness constancy was only present when there were strong, but not weak, reverberant cues—whereas estimates of sound source distance were not reliably affected by changes in room characteristics.

These results suggest that the capacity for loudness constancy requires the prevailing sound field to support the production of appreciable amounts of reverberant energy. However, Zahorik and Wightman (2001) also acknowledged that estimates of sound source distance obtained from visual signals may also affect loudness judgements. This is a plausible suggestion as visual information tends to support more accurate estimates of a source's distance than auditory cues (Anderson & Zahorik, 2014; Kolarik et al., 2016; Loomis et al., 1998). It would seem desirable for the mechanisms of loudness constancy to be capable of incorporating visually-determined distance estimates in such situations—indeed, Calcagno, Abregu, Eguía, and Vergara (2012) have proposed a key role for vision in establishing the distance of sound sources to aid in the interpretation of aural events.

Previous studies have suggested that visual information is capable of affecting loudness judgements. Mohrmann (1939, as described in Brunswik, 1956, p. 70–72) positioned pairs of loudspeakers at different distances and required participants to use the method of adjustment to equate the loudness of sounds from each loudspeaker, with the sounds presented in alternation. Participants demonstrated a high degree of loudness constancy, which tended to further increase when participants were able to see the scene compared to when completing the task in darkness—suggestive of a visual contribution to the mechanisms permitting loudness constancy. Similar results were obtained by Shigenaga (1965), who concluded that "there is a close relationship between the perception of acoustic distance, the constancy of loudness and the visionary cue" (p. 331). While a similar method was used by von Fieandt (1951), the contribution of vision to their reported findings is difficult to assess as performance was similarly close to complete constancy in visual and non-visual conditions.

Each of these studies were conducted under conditions in which there were likely to be concomitant auditory cues that varied as a function of distance (e.g., reverberation). In contrast, Mershon et al. (1981) investigated the effect of visual cues on loudness judgements in an anechoic environment. Sounds were presented from a hidden loudspeaker while a silent loudspeaker was visible at different distances. Due to the visual capture phenomenon, the silent but visible loudspeaker was perceived to be the sound source. A magnitude estimation approach was used in which participants were asked to rate the loudness of the sounds. Mershon et al. (1981) found that, in an anechoic room, such loudness ratings increased when the apparent distance of the sound source changed from 75cm to 225cm, consistent with the operation of a loudness constancy mechanism driven by visual input. However, the loudness judgements decreased to an intermediate value when the apparent distance of the sound source was further increased to 375cm. Hence, while providing evidence for a visual contribution to loudness, the findings of Mershon et al. (1981) are equivocal when concerning the robustness and generality of such an influence.

More recently, Altmann et al. (2012) adopted a similar approach to investigating the influence of visual cues on loudness judgements. Participants heard short bursts of bandpass noise, via earphones, that were played simultaneously with the offset of a light source at varying distances. The resulting loudness ratings showed no evidence of loudness constancy, and were instead driven by the aural sound intensity and unaffected by the distance of the putative sound source. While potentially inconsistent with a visual contribution to loudness, it is unclear whether participants were able to accurately perceive the distance of the light source due to the limited availability of visual distance cues. Intriguingly, Altmann et al. (2012) also performed magnetoencephalography recordings which suggested that the neural processing of auditory signals may have been affected by the visual distance information, despite this lack of behavioral loudness constancy.

Here, we aimed to assess the evidence for a visual contribution to *apparent* loudness judgements. We constructed the simulated environment shown in Figure 3.1, consisting of a single visible sound source (a loudspeaker) on a grassed open-field. By rendering the loudspeaker at one of three different positions and including multiple visual cues to scene depth, we were able to manipulate the perceived distance of the apparent sound source. Furthermore, the use of an open-field environment was designed to convey the expectation of little reverberant energy (D. G. Richards & Wiley, 1980) to the listener. Accordingly, we delivered sounds with a direct component only, simulating an anechoic environment in which the sound source power and distance are entirely confounded in the auditory signals.

We measured the potential influence of visual cues to sound source distance using a temporal two-interval forced-choice behavioral task, in which a reference stimulus of standard aural intensity and visual distance was presented in temporal proximity to a comparison stimulus of variable aural intensity and different visual distance (nearer or farther). By presenting the comparison with different aural intensities, we identified the aural intensity that was required for the comparison to be perceived as equally loud as the reference. If visual signals to sound source distance can affect loudness judgements, such points of subjective loudness equality (PSEs) would have lower and higher aural intensities for farther and closer sound sources, respectively. Alternatively, if visual

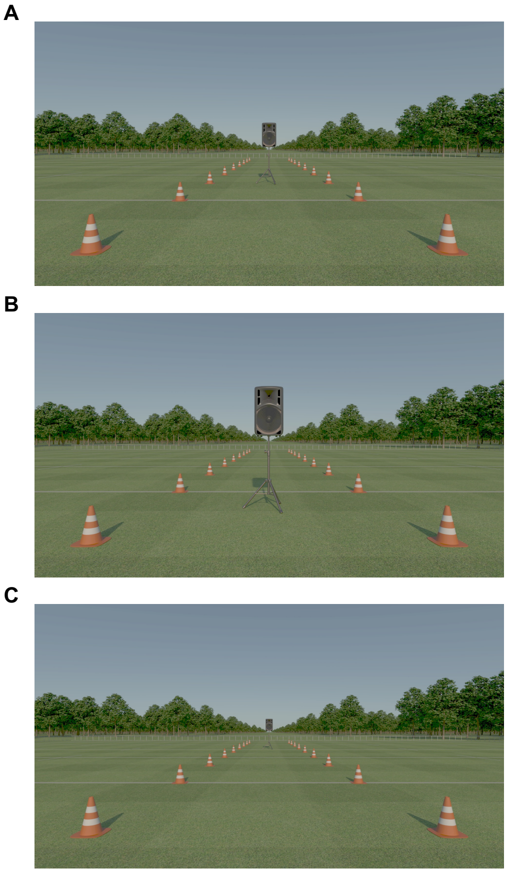

Figure 3.1. Visual stimuli used in Experiments 1 and 2. The loudspeaker is positioned at the reference (**A**), near (**B**), or far (**C**) distance from the observer.

signals to sound source distance are not incorporated in loudness judgements, such PSEs would be equal for farther and closer sound sources.

We performed a series of six experiments, across which we altered the auditory characteristics, the visual presentation, and the relationship between the auditory and visual presentations. In Experiments 1 and 2, participants judged the relative loudness of pure tones delivered with accompanying visual depictions of loudspeakers at different distances. In Experiment 3, this visual environment was altered to include additional visual cues to distance and to provide a visual cue that was synchronous with the auditory presentation. For Experiments 4 and 5, the auditory stimulus was changed to a pink noise burst and a speech utterance (‘ba’), respectively. The final experiment introduced a distance-dependent delay between the visual onset cue and the delivery of the auditory stimulus. To pre-empt our results, we found considerable evidence that the presented visual cues did not influence loudness judgements in these scenarios.

3.2 Experiment 1

3.2.1 Method

Participants

Participants ($N = 15$) with normal or corrected-to-normal vision and no auditory pathologies (self-reported) were recruited from a pool of students enrolled in an introductory psychology course at UNSW Sydney. Participants received course credit for their involvement and gave informed and written consent in accordance with the experiment protocols approved by the Human Research Ethics Advisory Panel in the School of Psychology, UNSW Sydney (#2683). All participants were naïve to the purposes of the experiment.

One participant was not considered due to a computer malfunction which prevented completion of the data collection session. The following analyses were conducted on the remaining 14 participants (3 males, 11 females; ages ranged from 18 to 25 with a median of 18.5).

Apparatus

Auditory stimuli were presented via an 'AudioFile' device (Cambridge Research Systems, Kent, UK) and over-ear headphones (Sennheiser, Wedemark, Germany; model HD 201). The relationship between tone amplitude and the sound intensity level produced by the headphones was determined using an artificial ear, microphone, and analyser (Brüel & Kjæer, Nærum, Denmark; models 4152, 4144, and 2250, respectively). All subsequently reported sound levels are in units of dB SPL as determined by this calibration method.

Visual stimuli were presented on a Display++ LCD monitor (Cambridge Research Systems, Kent, UK) with a spatial resolution of 1920 × 1080 pixels, temporal resolution of 120Hz, and mean luminance of 60 cd/m^2. The relationship between the video signal and monitor luminance was linear. Participants viewed the monitor in a darkened room from a distance of 52cm, via a chin rest, for a visual angular subtense of 76.6° × 43.1°.

The experiment was controlled using PsychoPy 1.83.04 (Peirce, 2007, 2008) and Python 2.7.11.

Stimuli

Auditory stimuli were pure tones with a frequency of 250 Hz and duration of 200ms, with a 5ms Hanning window applied at the start and end of the waveform. Waveforms had a sampling rate of 44.1 kHz and were identical in the left and right channels. A library of such tones was created that ranged from 50 to 80 dB in steps of 0.5 dB.

Visual stimuli were depictions of an outdoor environment in which the observer was positioned on a field with a loudspeaker visible in the scene. As shown in Figure 3.1, the loudspeaker was positioned at a distance of 15m (middle / reference), 7.5m (near), or 30m (far). A depiction was also produced in which the loudspeaker was not present in the scene. Images were rendered using Mitsuba (0.5.0; `http://www.mitsuba-renderer.org`).

Design and Procedure

The experiment had a two-way within-subjects design, with factors of reference level (66, 68, 70 dB) and comparison distance (near, far). The procedure for a given participant was conducted in a single session lasting approximately one hour. The session consisted of a series of six runs, where each run assessed two combinations of reference levels and comparison distances. The combinations were arranged such that they had different reference level and comparison distance. The ordering of runs was randomised for each participant, and there was a self-paced break of at least 30 seconds between each run.

Each run consisted of a series of trials, where each trial consisted of a temporal two-interval forced-choice task. Each interval began with a 750ms preparatory period in which the scene was presented without a visible loudspeaker. The rendering with the loudspeaker at the required position was then presented for 1s, with the opacity increasing linearly to complete visibility over the first 75ms. The auditory stimulus was then delivered while the rendering remained visible for the following 700ms, with the opacity decreasing linearly to complete transparency over the last 75ms. This procedure was then repeated for the second interval. On each trial, one of the intervals contained the loudspeaker at the reference distance and the tone at the reference level while the other interval contained the loudspeaker at the comparison distance and the tone at the comparison level. The interval containing the reference was randomised on each trial. Following presentation of the two intervals, the scene was presented (with no visible loudspeaker) with a written prompt “Which interval contained the louder sound? Press the left arrow key for the first interval. Press the right arrow key for the second interval”. The next trial commenced subsequent to the participant’s button press, with a minimum inter-trial interval of 3 seconds.

The level of the comparison tone on each trial was determined using a Psi adaptive staircase procedure (Kontsevich & Tyler, 1999). Each run contained two separate staircases, one for each different combination of reference level and comparison distance. Each consisted of 30 trials, and the staircase order was randomised within each run. As part of the staircase procedure, participant responses were modelled via a logistic-based psychometric function, after Kingdom and Prins (2010):

$$\psi\left(x;\alpha,\beta,\gamma,\lambda\right) = \gamma + (1-\gamma-\lambda)\ \left(\frac{1}{1+e^{-\beta(x-\alpha)}}\right) \tag{3.1}$$

This psychometric function describes the probability of selecting the comparison interval as containing the louder sound for a given comparison level (x), where γ and λ are the ‘guess’ and ‘lapse’ rates (both fixed at 0.05), α is the point of subjective equality (PSE), and β is the slope. For the Psi procedure, the candidate comparison intensities were between 50 and 80 dB in 0.5 dB increments. This distribution was also used for the point of subjective equality (α), while the slope (β) was given by 50 logarithmically-spaced values between 0.1 and 10.0.

Before commencing the experiment, the participant’s dominant eye was determined using the ‘card test’ (described by Ehrenstein, Arnold-Schulz-Gahmen, & Jaschinski, 2005). This was used to adjust the location of an occluder attached to the chinrest such that the participant viewed the monitor through their dominant eye only. This monocular viewing was designed to remove the influence of stereopsis cues to the true depth structure of the testing booth and promote immersion in the depicted scene.

Participants were then introduced to the task via a set of computer-based instructions. They then completed a practice run, which was identical to a given random run in the experiment. Following completion of the practice run, the experimenter visually evaluated the resulting psychometric functions to determine if it appeared that participants understood the task requirements. This judgement was based on the concordance between the observed response probabilities and the assumed psychometric function—no consideration was given to the relevance to the experiment hypotheses. An additional practice run was completed if necessary.

Analysis

The experiment procedure produced 360 data points per participant, where each data point specified the intensity of the comparison sound and the corresponding participant judgement regarding its loudness relative to the reference sound. With six conditions, this corresponded to 60 data points per condition (2 staircases × 30 trials per staircase) per participant.

Our analysis goal was to compare the evidence for a visual contribution to loudness judgements against a null hypothesis in which loudness judgements are unaffected by concurrent visual signals. To begin, we define an index of vision's influence on loudness as the change in the point of subjective loudness equality with a halving of sound source distance. This index (L_v) is in units of dB. Under the null hypothesis, the loudness judgements are unaffected by the depiction of the sound source, to give an L_v of zero. Under the alternative hypothesis, the depiction of the sound source affects loudness judgements such that a depiction of a nearer sound source requires a higher sound intensity at-the-ear to be perceived as equally loud as a farther sound source. For example, a point sound source in a free field with a given source level that produces 70 dB at-the-ear from the reference distance (15m) would produce 76 dB at-the-ear when at the near distance (7.5m) with the same source level, for an L_v of 6 dB.

We used a hierarchical Bayesian approach for our data analysis and hypothesis evaluation. We begin by assuming that L_V for each participant is drawn from a normal distribution with a particular mean (L_V^μ) and standard deviation (L_V^σ). The L_V^μ parameter is critical for our hypothesis comparison, and requires the setting of a prior that reflects our belief of the plausibility of obtaining particular values under the alternative hypothesis. As discussed above, the maximum plausible value is around 6 dB—however, we consider this to be relatively unlikely as the environment does not depict a point sound source or a free field, and because humans do not typically achieve complete constancy in experiment scenarios. Hence, we set the prior for L_V^μ as a normal distribution with a mean of 3 dB and standard deviation of 1.5 (see Figure 3.3 for a depiction). Setting the standard deviation to half of the mean follows the recommendation of Dienes (2014), and expresses our belief that L_V^μ, under the alternative hypothesis, would be somewhere between 0 dB and 6 dB, with intermediate values being more plausible. We set the prior for L_V^σ to be vague, as a uniform distribution with lower and upper bounds of 0.01 and 3, respectively.

With such L_V estimates, each participant's PSEs for each of the six within-subjects conditions (two visual distances, three reference levels) can then be determined by the addition (PSE for the near conditions) and subtraction (PSE for the far conditions) of their L_V and the reference levels. We then assume that each participant's psychometric functions have a common slope across the within-subjects conditions, and that such slopes have a prior distribution that is uniform with lower and upper bounds of log(0.005) and log(10), respectively. Finally, we model each trial in the observed data as a Bernoulli distribution, with the probability parameter given by a logistic psychometric function in Equation 3.1. The alpha parameter of this logistic function is given by the estimated PSE for this participant, comparison distance, and reference level, the beta parameter is given by the estimated slope for this participant, and the catch and lapse rates are fixed at 0.05.

The analysis model was implemented in PyMC3 (Salvatier, Wiecki, & Fonnesbeck, 2016). Markov chain Monte-Carlo (MCMC) sampling was performed using a No-U-turn sampler (Hoffman & Gelman, 2014) with PyMC3's default initialization. A total of 20000 draws were used for each of 3 independent chains in the sampling process, after discarding the initial draws (2000) used in initializing the sampler, which were then concatenated and thinned by a factor of 5 to produce the posterior distributions. Sampling quality was assessed by visual inspection of sampling traces and autocorrelations, and by consideration of the match between the fitted psychometric functions and each participant's raw data (shown in Figure 3.2 for a representative participant).

Following model estimation, we used the Savage-Dickey method (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010) to compute a Bayes factor to quantify the evidence in the comparison of the null hypothesis (L_V^μ is 0 dB) and the alternative hypothesis (L_V^μ is most likely to be between 0 dB and 6 dB, with intermediate values being more plausible). Specifically, we apply kernel density estimation (Gaussian kernel, automatic bandwidth selection using Scott's method as implemented in `scipy`) to the samples from L_V^μ to obtain the posterior probability at 0 dB. The Bayes factor was then computed as the ratio of this posterior probability and the prior probability evaluated at 0 dB, and is communicated as the $\log_{10}$ of this ratio.

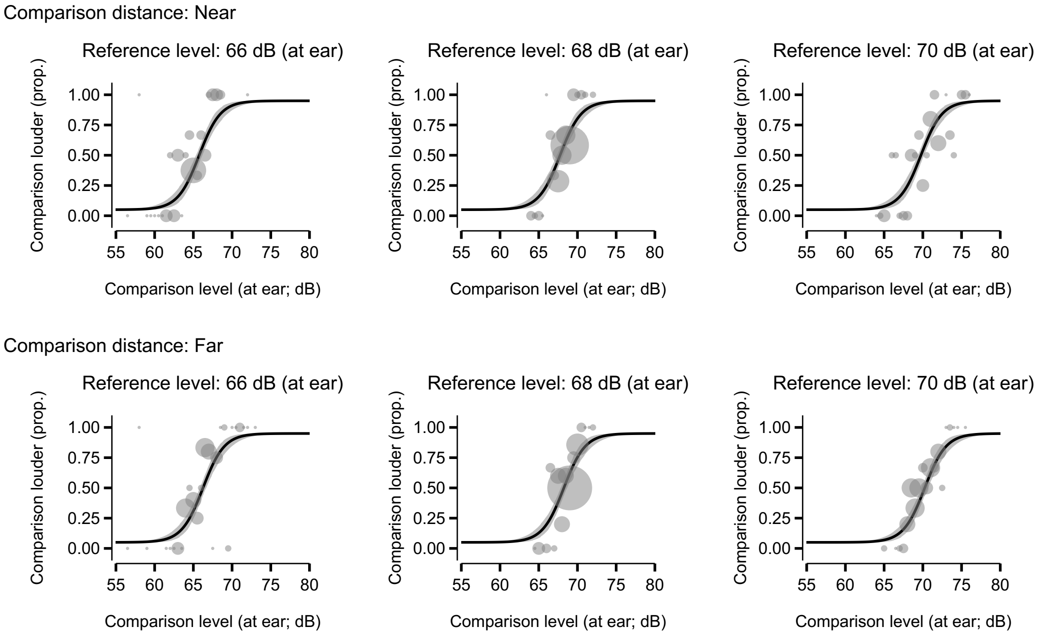

Figure 3.2. Psychometric functions from a representative participant in Experiment 1. Each panel depicts a single condition, with rows varying in the distance of the depicted comparison loudspeaker (near, far) and columns varying in the intensity at-the-ear of the reference sound (66, 68, 70 dB). Within each panel, points show the proportion of trials in which the participant identified the comparison sound as louder for a given comparison sound level. The point sizes are proportional to the number of trials at that particular comparison sound level. The solid line represents the mean of the posterior psychometric function distribution, and the surrounding grey area represents its 95% HPD interval.

3.2.2 Results and Discussion

We measured the sound level that was required for a tone emitted by a visible loudspeaker to be perceived as equally loud as a reference tone sound level emitted from a visible loudspeaker that was at a closer or farther distance. If distance estimates obtained from vision can affect loudness judgements, participants would be affected by the apparent distance of the sound emitter such that closer and farther loudspeakers would require higher and lower sound levels, respectively, to be perceived as equally loud as the reference. We constructed a hierarchical Bayesian model to compare the evidence for this hypothesis against a null hypothesis in which visually-depicted distance has no effect on loudness judgements.

We find that the posterior distribution for the mean visual influence on loudness index was close to zero, with a mean of -0.32 dB (95% HPD $[-0.73, 0.08]$). As shown in Figure 3.3, the posterior density increased its mass at zero between the prior for the alternative hypothesis and the posterior. Quantification as a Bayes factor indicated moderately strong evidence for the null hypothesis against our particular alternative hypothesis ($\log_{10} BF_{1,0} = -1.14$).

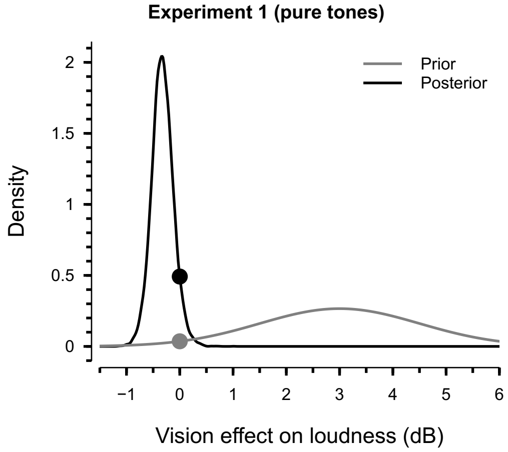

Figure 3.3. Evaluation of the vision effect on loudness in Experiment 1. The grey line shows the prior density for the mean vision effect on loudness (L_V^μ) parameter under the alternative hypothesis. The black line shows the posterior density for this parameter given the data obtained in Experiment 1. The filled circles mark the density of the prior (grey circle) and posterior (black circle) at a vision effect on loudness of 0 dB.

The results of this experiment do not support the prediction that visual cues to distance would affect judgements of loudness in a way that would be expected from the operation of a loudness constancy mechanism. Instead, the results of this experiment provide evidence for participant judgements of loudness being unaffected by the visual signals.

Given this lack of support for the loudness constancy hypothesis in this experiment, we next conducted an experiment to probe the generality of these results. We made two primary changes in Experiment 2. First, we removed the intermediate level (68 dB) of the reference level factor and replaced it with a new level of the comparison distance factor in which the comparison stimulus was presented at the same distance as the reference. The rationale for this change was to encourage participants to register the differential positioning of the loudspeakers in the near and far conditions via exposure to situations in which the position of the loudspeaker does not change. Second, we changed the task instructions in an attempt to avoid biasing participants towards making their judgements based on the aural level. This involved alterations designed to promote consideration of each loudspeaker presentation as a separate object with potentially different capacities to produce sound.

3.3 Experiment 2

3.3.1 Method

Participants

An additional set of unique participants ($N = 18$) was recruited as per the procedures for Experiment 1.

Apparatus

The apparatus were as per Experiment 1.

Stimuli

The stimuli were as per Experiment 1.

Design and Procedure

The design and procedure were as per Experiment 1, with the exception that the reference levels were changed to 66 and 70 dB (the 68 dB reference was removed) and the comparison distances were changed to near, far, and reference (where ‘reference’ is an addition in which the comparison distance is the same as the reference distance).

Analysis

The analysis was as per Experiment 1. Four participants were excluded due to an inability to obtain reasonable parameter estimates from their data (see Supplementary Figure 2). The following analyses were conducted on the remaining 14 participants (4 males, 10 females; ages ranged from 18 to 23 with a median of 19).

3.3.2 Results and Discussion

Participants appeared to again be unaffected by the apparent distance from the simulated sound source. The posterior distribution for the mean visual influence on loudness index was close to zero, with a mean of -0.07 dB (95% HPD $[-0.26, 0.10]$). As shown in Figure 3.4, the posterior density increased its mass at zero between the prior for the alternative hypothesis and the posterior. Quantification as a Bayes factor indicated very strong evidence for the null hypothesis against our particular alternative hypothesis ($\log_{10} BF_{1,0} = -1.92$).

The results of Experiment 2 are consistent with those from Experiment 1—there is no indication that participants were demonstrating loudness constancy in their judgements. Together, Experiments 1 and 2 both provided little support for the notion that visual cues to distance necessarily affect loudness judgements so as to support loudness constancy.

In Experiment 3, we sought to alter the visual stimuli to provide a more compelling impression of the environment and of the position of its visible sound sources. In particular, we were concerned that the ‘magical’ appearance and disappearance of the loudspeakers in Experiments 1 and 2 may have constituted an ecological incoherence that may have prevented participants from integrating the visual cues to distance into their loudness judgements. Hence, we developed a modified environment in which the loudspeakers at each of the three distances were present simultaneously, as shown in

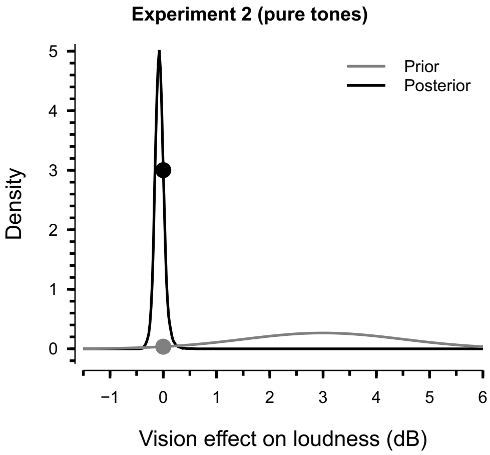

Figure 3.4. Evaluation of the vision effect on loudness in Experiment 2. The grey line shows the prior density for the mean vision effect on loudness (L_V^μ) parameter under the alternative hypothesis. The black line shows the posterior density for this parameter given the data obtained in Experiment 2. The filled circles mark the density of the prior (grey circle) and posterior (black circle) at a vision effect on loudness of 0 dB.

Figure 3.5 (which is similar to the arrangement suggested by Eisler, 1981). To prevent an expectation of binaural auditory cues, we rotated the position of the camera such that the active loudspeaker was always positioned directly in front of the observer. We anticipated that the addition of such simultaneous relative size cues and rich motion parallax cues would increase the sense of immersion and of the presence of three objects at different distances.

Furthermore, we introduced a visual cue to loudspeaker activation that was synchronous with the onset of the auditory stimulus. A light source at the top of each loudspeaker emitted red light while the associated sound was being played. This both served as an indication that the frontally-positioned loudspeaker was the source of the emitted sound and provided an audio/visual synchrony cue to aid in the perceptual binding of the sound to the loudspeaker.

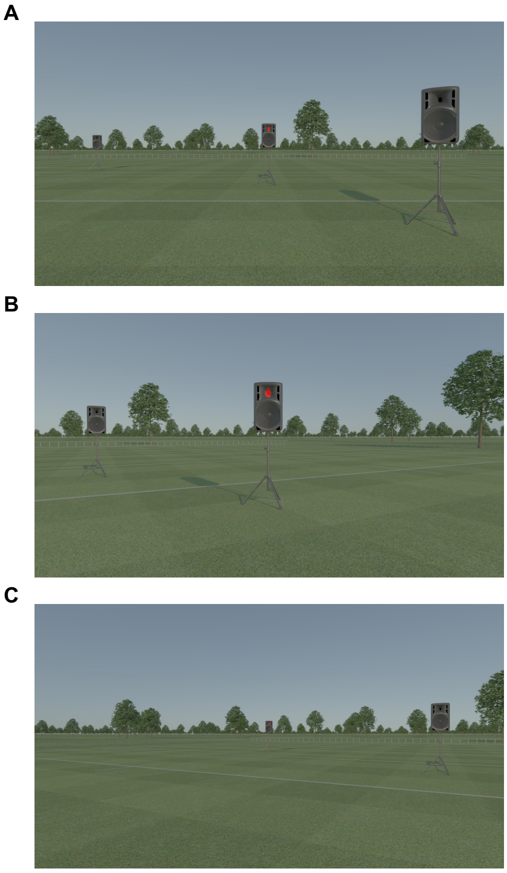

Figure 3.5. Sample frames from the visual stimuli used in Experiments 3-6. The camera is centered on the loudspeaker at the reference (**A**), near (**B**), or far (**C**) distance from the observer. The renderings depict the display during presentation of the sound, which includes the presence of a red light on the central loudspeaker that is absent when the sound is not playing.

3.4 Experiment 3

3.4.1 Method

Participants

An additional set of unique participants ($N = 16$) was recruited as per the procedures for the previous experiments.

Apparatus

The apparatus were as per the previous experiments.

Stimuli

There were three primary changes to the visual stimuli in this experiment. First, the near and far speakers were moved horizontally to permit all three speakers (near, far, reference) to be visible simultaneously. The near and far speakers were moved such that the camera would need to rotate 30° to the right and left, respectively, to focus on the speaker. Second, a light source was added to the upper section of each speaker. Additional images were rendered depicting a focus on each speaker with this light active, which emitted a red glow. Third, a set of 121 images was rendered in which the angle and distance of the camera target changed linearly between the near and far loudspeaker positions. Examples of the resulting stimuli are shown in Figure 3.5.

The auditory stimuli were as per the previous experiments.

Design and Procedure

The presentation sequence on a given trial was changed from Experiment 2 to accommodate the dynamic transition between the camera's focus on different loudspeakers. Each interval began with a 500ms static period in which the camera was focused on the speaker from the previous interval (for the first of the two intervals, this was the second interval in the previous trial). There was then a period in which the presented image changed on each frame to display the transition to the focus loudspeaker for the interval. The duration of this period was either 500ms if the transition involved the reference distance or 1000ms if it was between the near and far speakers. There was then a 750ms static period in which the camera was focused on the relevant loudspeaker, followed by the synchronous presentation of the scene with the light on the relevant loudspeaker active and the onset of the sound. The image with the active light was shown for 125ms, before the light was turned off and the static image displayed for an additional 500ms. This was then repeated for the second interval.

Analysis

Two participants were excluded due to an inability to obtain reasonable parameter estimates from their data. The following analyses were conducted on the remaining 14 participants (4 males, 9 females; ages ranged from 18 to 23 with a median of 19; demographics unknown for 1 participant).

3.4.2 Results and Discussion

Consistent with the results of Experiments 1 and 2, participants once again appeared to be unaffected by the apparent distance from the simulated sound source. The posterior distribution for the mean visual influence on loudness index was close to zero, with a mean of −0.16 dB (95% HPD $[-0.57, 0.26]$). As shown in Figure 3.6, the posterior density increased its mass at zero between the prior for the alternative hypothesis and the posterior. Quantification as a Bayes factor indicated very strong evidence for the null hypothesis against our particular alternative hypothesis ($\log_{10} BF_{1,0} = -1.59$).

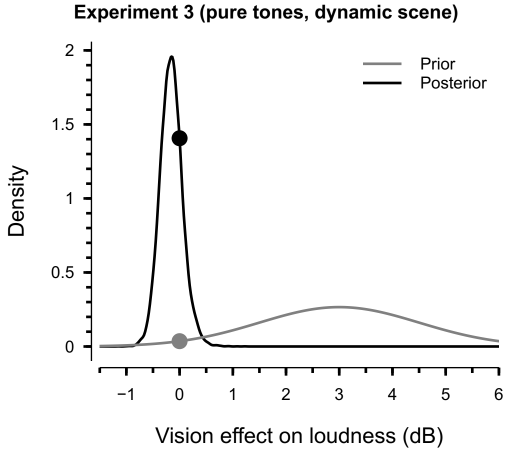

Figure 3.6. Evaluation of the vision effect on loudness in Experiment 3. The grey line shows the prior density for the mean vision effect on loudness (L_V^μ) parameter under the alternative hypothesis. The black line shows the posterior density for this parameter given the data obtained in Experiment 3. The filled circles mark the density of the prior (grey circle) and posterior (black circle) at a vision effect on loudness of 0 dB.

The alterations to the visual stimuli in this experiment had little apparent effect, with loudness judgements again being consistent with the aural level and seemingly unaffected by the apparent distance of the sound source as conveyed by visual cues.

Taken together, Experiments 1–3 provide considerable evidence that loudness constancy is not necessarily achieved when cues to distance are provided solely through the visual modality.

However, Experiments 1–3 each only considered one form of auditory stimulation—250 Hz pure tones. While an appealingly simple stimulus, pure tones may not be representative of the conditions in which loudness constancy is typically expressed. Indeed, Shigenaga (1965) reported a complex variation of the degree of loudness constancy with pure tone frequency and Mohrmann (1939, as described in Brunswik, 1956, p. 70–72) found that tones were among the forms of auditory stimulus that were least able to elicit high degrees of loudness constancy.

In the next experiment, we sought to extend the form of auditory stimulation beyond pure tones. We retained the loudspeaker depiction used in this experiment, but changed the auditory exemplars to pink noise bursts. Noise was used as the auditory stimulus in Mershon et al. (1981), who reported equivocal degrees of loudness constancy, and by Altmann et al. (2012), who reported a lack of behavioural loudness constancy. Mohrmann Brunswik (1939, as described in 1956, p. 70–72) also had a condition where the stimulus was auditory noise and reported substantial degrees of loudness constancy, however such judgements were very similar regardless of the presence of visual cues.

3.5 Experiment 4

3.5.1 Method

Participants

An additional set of unique participants ($N = 15$) was recruited as per the procedures for the previous experiments.

Apparatus

The apparatus were as per the previous experiments. A calibration was performed in which the relationship between pink noise waveform root-mean-square (RMS) and the level produced by the headphones was determined and used to generate stimuli in dB units.

Stimuli

Auditory stimuli were pink noise bursts that were 200ms in duration. A library of noise samples was created that ranged from 50 to 80 dB in steps of 0.5 dB. Each noise sample was created by first generating a set of phases from the frequency domain representation of a sequence drawn from a uniform distribution. These phases were then combined with amplitudes that were proportional to the inverse of the frequency between 20 Hz and 20 kHz and then converted into the time domain. The resulting distribution was then z-scored before being multiplied by the desired RMS. A 5ms Hanning window was applied to the start and end of resulting waveform, and the same waveform was entered into the left and right stereo channels.

The visual stimuli were as per Experiment 3.

Design and Procedure

The design and procedure were as per Experiment 3.

Analysis

Two participants were excluded due to an inability to obtain reasonable parameter estimates from their data. The following analyses were conducted on the remaining 13 participants (6 males, 7 females; ages ranged from 18 to 21 with a median of 19).

3.5.2 Results and Discussion

Changing the sound from a pure tone to a pink noise burst appeared to have little influence on the results, with participants once again appearing to be unaffected by the apparent distance to the simulated sound source. The posterior distribution for the mean visual influence on loudness index was close to zero, with a mean of 0.08 dB (95% HPD $[-0.09, 0.26]$). As shown in Figure 3.7, the posterior density increased its mass at zero between the prior for the alternative hypothesis and the posterior. Quantification as a Bayes factor indicated very strong evidence for the null hypothesis against our particular alternative hypothesis ($\log_{10} BF_{1,0} = -1.90$).

The aim of this experiment was to expand the consideration of the role of visual cues to sound source distance in loudness judgements beyond pure tone auditory

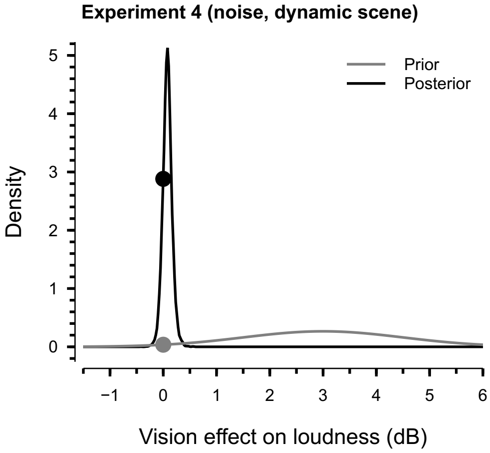

Figure 3.7. Evaluation of the vision effect on loudness in Experiment 4. The grey line shows the prior density for the mean vision effect on loudness (L_V^μ) parameter under the alternative hypothesis. The black line shows the posterior density for this parameter given the data obtained in Experiment 4. The filled circles mark the density of the prior (grey circle) and posterior (black circle) at a vision effect on loudness of 0 dB.

stimuli. Hence, we replaced the presentation of pure tones with the presentation of pink noise bursts. However, the results of this experiment are very similar to those of the previous experiments using pure tones—participant judgements of loudness was seemingly unaffected by visual cues to the distance of the sound source and were instead closely aligned with the aural level of the noise.

In the next experiment, we expanded the range of auditory stimulation by evaluating loudness judgements to temporally-structured waveforms in which an adult male utters the syllable ‘ba’. Mohrmann (1939, as described in Brunswik, 1956, p. 70–72) obtained the highest degrees of loudness constancy when speech comprised the auditory stimulation, and vocalisations produced the highest loudness constancy in the report by von Fieandt (1951). Furthermore, Rosenblum and Fowler (1991) showed that the loudness of speech can be affected by the perceived effort of the speaker as conveyed through vision.

3.6 Experiment 5

3.6.1 Method

Participants

An additional set of unique participants ($N = 23$) was recruited as per the procedures for the previous experiments.

Apparatus

The apparatus were as per the previous experiments.

Stimuli

Auditory stimuli were formed from a single recording of an adult male uttering the syllable ‘ba’. The waveform was approximately 275ms in duration, with an 11ms Hanning window applied to the beginning and end of the waveform. Intensity was manipulated by z-scoring the resulting waveform before multiplication by the desired RMS, and a library of samples with intensities from 50 to 80 dB in steps of 0.5 dB was formed.

The visual stimuli were as per Experiments 3 and 4.

Design and Procedure

The design and procedure were as per Experiments 3 and 4.

Analysis

Two participants were excluded due to an inability to obtain reasonable parameter estimates from their data, and an additional three participants were excluded due to not completing a full session. The following analyses were conducted on the remaining 18 participants (10 males, 8 females; ages ranged from 18 to 34 with a median of 19).

3.6.2 Results

Changing the sound from a pink noise burst to an utterance ('ba') appeared to have little influence on the results, with participants once again appearing to be unaffected by the apparent distance to the simulated sound source. The posterior distribution for the mean visual influence on loudness index was close to zero, with a mean of -0.10 dB (95% HPD $[-0.20, 0.00]$). As shown in Figure 3.8, the posterior density increased its mass at zero between the prior for the alternative hypothesis and the posterior. Quantification as a Bayes factor indicated strong evidence for the null hypothesis against our particular alternative hypothesis ($\log_{10} BF_{1,0} = -1.41$).

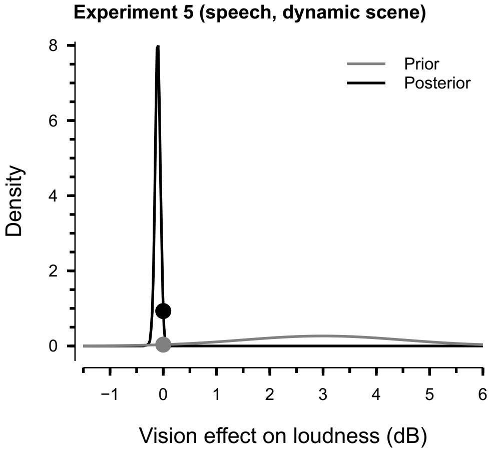

Figure 3.8. Evaluation of the vision effect on loudness in Experiment 5. The grey line shows the prior density for the mean vision effect on loudness (L_V^{μ}) parameter under the alternative hypothesis. The black line shows the posterior density for this parameter given the data obtained in Experiment 5. The filled circles mark the density of the prior (grey circle) and posterior (black circle) at a vision effect on loudness of 0 dB.

The aim of this experiment was to evaluate the potential role of visual cues to distance in loudness judgements for sounds mimicking human speech. As previous reports have highlighted that the loudness of speech is particularly susceptible to being affected by concomitant cues to the distance of the sound emitter (Mohrmann, 1939, as described in Brunswik, 1956, p.70–72; Pollack 1952; von Fieandt, 1951), we expected that this would provide an opportune stimulus for demonstrating the capacity for visual cues to inform loudness judgements. However, consistent with the results of Experiments 1–4, we find that the loudness judgements were closely aligned with the aural level of the utterance and were only slightly affected by the perceived distance of the sound emitter.

In Experiments 3–5, we added a visual indicator that flashed on the currently active loudspeaker coincident with the onset of the auditory stimulus. The purpose of this indicator was to both emphasise which of the three loudspeakers was currently producing the sound and to promote the binding of the auditory and visual signals through temporal synchrony. However, this approach does not reflect the ecological constraint that sound travels more slowly than light—and there is evidence to suggest that the perceptual system incorporates such knowledge (Jaekl, Seidlitz, Harris, & Tadin, 2015; Sugita & Suzuki, 2003). Hence, the inclusion of a distance-dependent audiovisual delay may be more effective in promoting binding than audiovisual simultaneity—and may also provide a new multisensory cue to sound source distance. We evaluated this possibility in Experiment 6.

3.7 Experiment 6

3.7.1 Method

Participants

An additional set of unique participants ($N = 19$) was recruited as per the procedures for the previous experiments.

Apparatus

The apparatus were as per the previous experiments.

Stimuli

The auditory stimuli were as per Experiment 5, and the visual stimuli was as per Experiments 3–5. However, the onset of the auditory stimulus was delayed relative to the onset of the 'active light' on the focused loudspeaker. The delays were approximately 22, 43, and 87 milliseconds for the near, reference, and far distances, respectively.

Design and Procedure

The design and procedure were as per Experiments 3–5.

Analysis

No participants were required to be excluded, and the following analyses were conducted on the complete set of 19 participants. No demographic information was collected for this experiment.

3.7.2 Results

The addition of a distance-dependent asynchrony between the auditory and visual indicator (light flash on the active speaker) appeared to have little influence on the results, with participants once again appearing to be unaffected by the apparent distance to the simulated sound source. The posterior distribution for the mean visual influence on loudness index was close to zero, with a mean of -0.01 dB (95% HPD $[-0.11, 0.09]$). As shown in Figure 3.9, the posterior density increased its mass at zero between the prior for the alternative hypothesis and the posterior. Quantification as a Bayes factor indicated decisive evidence for the null hypothesis against our particular alternative hypothesis ($\log_{10} BF_{1,0} = -2.36$).

The aim of this experiment was to evaluate the influence of distance-dependent audiovisual asynchrony on loudness judgements. As previous reports have suggested that the perceptual system is aware of the likely differences in the timing of auditory and visual stimulation from sound sources at a distance (Jaekl et al., 2015; Sugita & Suzuki, 2003), we were concerned that the audiovisual synchrony in Experiments 3–5 may have reduced the perceived binding of the auditory and visual stimulation or confounded the interpretation of the apparent distance of the sound source. However, we obtained results that were consistent with those from Experiments 3–5; that is, there was little indication that the loudness judgements were affected by the apparent distance of the sound source. Although the capacity of the perceptual system to utilise audiovisual delays is controversial (Arnold, Johnston, & Nishida, 2005; Lewald & Guski, 2004), this result from Experiment 6 argues against the absence of meaningful audiovisual asynchrony as being the critical determinant in the apparent lack of a visual contribution to loudness judgements in this series of experiments.

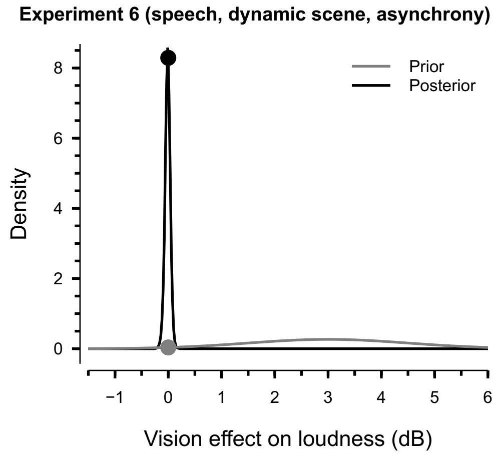

Figure 3.9. Evaluation of the vision effect on loudness in Experiment 6. The grey line shows the prior density for the mean vision effect on loudness (L_V^μ) parameter under the alternative hypothesis. The black line shows the posterior density for this parameter given the data obtained in Experiment 6. The filled circles mark the density of the prior (grey circle) and posterior (black circle) at a vision effect on loudness of 0 dB.

3.8 General Discussion

The aim of this study was to assess the evidence for a visual contribution to apparent loudness through the provision of source depth cues. By using a simulated outdoor scene with a visible loudspeaker as the sound source, we measured how human participants' judgements of relative loudness are affected by the depiction of sound source distance. In a series of six experiments, we find strong evidence that visual information simulating the distance of a sound source was not incorporated into the loudness judgements—which were instead consistent with the intensity of the auditory signals received at-the-ear. This result was obtained for pure tones, pink noise bursts, and speech utterances and across variations of the visual and audiovisual environment.

This apparent lack of a visual contribution to loudness is consistent with the behavioral findings reported by Altmann et al. (2012). Such an outcome is also potentially consistent with the anechoic experiment reported by Mershon et al. (1981) which, as described in the Introduction of this Chapter, was unclear in the strength of evidence for a visual contribution to loudness. Importantly, it is also in agreement with the proposal by Zahorik and Wightman (2001) that the key determinant of loudness is reverberant energy. However, this absence of a visual influence is inconsistent with studies that have reported that judgements become closer to constancy when visual information is available (Mershon, 1981; Mohrmann, 1939, as described in Brunswik, 1956, p. 70–72; Shigenaga, 1965). The key discriminating feature of such studies, relative to those reporting a lack of a visual contribution to loudness (such as the current study), appears to be that they were each conducted in reverberant rather than anechoic environments. This distinction implies that the presence of reverberation is required for vision to influence loudness. Another possibility is that by providing rich 3D depth cues the presence and plausibility of speaker depth was greater in these studies, and that this was a factor in demonstrating a visual influence on loudness. We cannot comprehensively reject the idea that sitting in a testing cubicle, in front of a 2D computer monitor does in some capacity (either cognitively or perceptually) undermine the immersion needed for the simulation of the speaker at a distance to be perceptually experienced as a speaker that is actually distant. Withstanding this, we note that there is evidence to suggest that 2D pictorial depth cues may provide comparable depth estimates to real environments (Anderson & Zahorik, 2014; Plumert, Kearney, Cremer, & Recker, 2005; Surdick et al., 1997). We also note that other experiments have employed stereoscopic depth cues (Berthomieu et al., 2019) and physical distance cues (Altmann et al., 2012) and observed comparable results.

To demonstrate that the visual system is indeed considering the loudspeakers to be positioned at different distances, we have constructed the variant of the Ponzo illusion shown in Fig. 3.10. In this depiction, each of the loudspeakers is of identical height in the image and is positioned at the vertical locations corresponding to the 'near', 'reference', and 'far' distances in Experiments 1 and 2. Perceptually, the three loudspeakers appear to be of different physical sizes (despite being the same physical size in the image) - an 'illusion' that supports our contention that their apparent distance is being registered as being at different locations by the visual system.

Why might the presence of a reverberant environment be a necessary precondition for vision to inform loudness judgements? We suggest two possible explanations. First, the perceptual system may have an expectation about the likely reverberation that would be elicited when a sound source produces acoustic energy in the environment (Traer & McDermott, 2016). The absence of such reverberation may produce a perceptual conflict that is resolved in favour of independence, rather than incorporation, of the auditory and visual information—that is, vision may not contribute to the interpretation of auditory signals because vision and audition are considered to be relating to different environmental sources. When reverberation is present, this may be sufficient to localise the sound source as being external to the

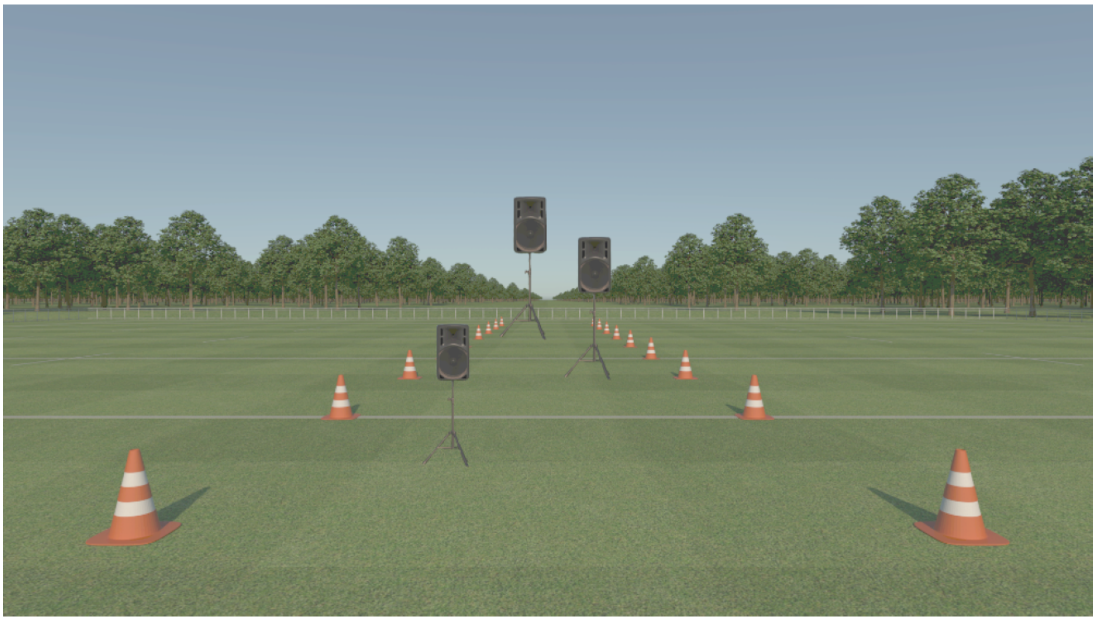

Figure 3.10. Demonstration that the visual system considers the loudspeakers to be positioned at different distances. Each loudspeaker is of identical height in the image, but appears to be of differing physical size.

perceiver and allow the system to incorporate information provided by vision. Second, reverberation may be necessary because the relevant information provided by vision pertains to reverberation. Thus far, we have focused on the potential for vision to identify the distance of the sound source. However, as mentioned in the Introduction of this Chapter, vision also has the potential ability to obtain information about the prevailing sound field. For example, vision can allow the identification of the reflective surfaces in the environment—which can provide estimates of parameters relating to reverberation such as room size (Calcagno et al., 2012). This information could then be incorporated into the interpretation of auditory signals, such as by supporting the discrimination of the direct and reverberant components, thereby affecting loudness in situations with appreciable reverberant energy.

The apparent necessity of reverberation for loudness constancy has the intriguing consequence that perception would be unable to sustain constancy in anechoic or weakly reverberant environments. Indeed, a dependence of loudness on the strength of reverberation was reported by Altmann et al. (2013). The ecological foundations, and potential behavioral consequences, of this apparent failure of perceptual constancy is an interesting avenue for future research.

Finally, it is relevant to consider whether the task requirements in the current experiments may have obscured any loudness constancy mechanisms. Specifically, participants were instructed to judge the relative loudness of sounds emitted by objects in the environment; if this was interpreted as being required to make judgements based on the intensity of the signal at-the-ear, we may not have captured a normal perceptual experience in which loudness constancy is evident. However, the presence of a perceptual constancy mechanism is often such that judgements are pulled in the direction of constancy even when attempting to consider sensory cues in isolation. Indeed, Mohrmann (1939, as described in Brunswik, 1956, p. 70–72) reported considerable loudness constancy with the availability of visual cues even when participants were explicitly instructed to adopt an attitude towards aural intensity. Hence, we suggest that task instructions are unlikely to completely explain the lack of loudness constancy that we observed in this study.

3.9 Conclusion

Despite providing information about the distance of a sound source and its environment, vision does not necessarily affect loudness judgements. Under anechoic conditions, when simulating the visual distance of a sound source on a 2D monitor with multiple rich monocular cues, loudness appears to be determined by the intensity of aural signals at-the-ear irrespective of concurrent visual information. This is consistent with the necessity of reverberation for loudness constancy, and the potential role of vision in such reverberant environments is an important direction for future research.

4 Visual Cues to Source Distance & Loudness (reverberant conditions)

Title: Source loudness estimates are not improved with visual cues to sound source distance in a reverberant environment

Author contributions:
Conceptualisation: SL, TJW, DJM. Stimuli: PWA, PZ. Methodology: SL, TJW, DJM. Programming: DJM. Data collection: SL. Data analysis and presentation: SL, DJM. Writing - original draft: SL. Writing – review and editing: SL, TJW, DJM. Supervision: TJW, DJM.

Preamble:

In the previous chapter we found that, in anechoic conditions, visual cues to the distance of the sound source did not influence loudness estimates. It may be that visual capture of the auditory signals did not occur because a certain amount of reverberation was needed to localise an auditory cue at an external location. Alternatively, it may be that visual capture did occur, but that visual cues only influence loudness when they relate information both about the source's distance, and the properties of the sound-field that give rise to the reverberant component of auditory input. In the following study we addressed these possibilities by examining whether visual cues to the distance of the sound source disambiguated loudness estimates of the source in reverberant conditions. In addition, it has been found that the degree of loudness constancy is highest when participants are directed to estimate the distal power of a signal (i.e., at-its-source) as opposed to estimating the proximal intensity of a signal (i.e., at-the-ear). Thus, to maximise the possibility of eliciting loudness constancy we required participants to make loudness judgements based on the distal signal of the sound source.

4.1 Abstract

Abstract

An auditory signal reaching the ear may be modulated both by the capacity of the sound source to produce acoustic energy (i.e., sound source power) and by the distance of the source from the listener. Loudness constancy requires that our perception of sound source power will remain invariant to the confounding effects of distance. Here, we used a reverberant environment to assess the evidence for the potential contribution of both auditory and visual cues to loudness constancy when taking *distal* source power estimates. We presented participants with a visual environment, on a computer monitor, which contained photographs of a loudspeaker at a particular distance in a concert hall, taken from the participant's perspective. This was accompanied by the delivery, via headphones, of a virtual sound source based on binaural room impulse response (BRIR) measurements at particular distances in the same concert hall. We measured the point of subjective equality between sounds relayed from a close reference loudspeaker (1.22m) against comparison loudspeakers that were placed further in the hall (2.44, 3.45, 4.88, 6.9 or 9.75m). Four different conditions were employed in a between subjects design: an audio only condition (n=41), a congruent audiovisual condition (n=40), a condition in which the comparison speakers were visually shifted to be farther than the concomitant virtual sound source (n=40) and a condition in which the comparison speakers were visually shifted to be closer than the concomitant virtual sound source (n=38). We report partial loudness constancy in the audio-only condition, based on reverberant cues. However, we did not find evidence that the inclusion of a congruent visible sound source increased loudness constancy. Finally, we found that in both incongruent conditions, the power of the comparison speaker was perceived to increase relative to the congruent conditions. If we use visual cues to disambiguate a sound source's depth, and if our sense of source depth influences our perception of a sound source's power, then visually shifting the apparent distance of the sound source should result in a shift of the perceived power of the source that is relative to its distance. Thus, the results of the incongruent manipulation indicate that visual cues to source distance do not facilitate accurate source power estimates.

4.2 Introduction

A function of perception is to decode information about the external environment from signals arriving at sensory receptors. However, in many instances the information arriving at our sensory receptors has an ambiguous relationship with its external causes. Consequently, when organising input into a representation the perceptual system has to solve an inference problem; it is possible for the same features of an object to generate different signals at a sensory receptor. Whilst multiple sensory consequences can return from a single object, we often perceive the features of the external object as being invariant (Walsh & Kulikowski, 1998). This capacity is labelled perceptual constancy. A well known kind of perceptual constancy is size constancy. Size constancy posits that while the size of an object subtended on the retina varies with distance, we have the capacity to perceive the physical size of that object as remaining stable (Brunswik, 1944).

A form of perceptual constancy that has received much less attention is loudness constancy. The intensity of a physical signal received at-the-ear may be ambiguous as it holds the possibility of being determined by both sound source power and source distance. This is because as a sound wave travels outwards its intensity attenuates and thus a sound source that is further away will have a lower intensity at-the-ear (Bronkhorst & Houtgast, 1999; Coleman, 1962). If one has the capacity to determine sound source power as stable despite changes in the physical signal at-the-ear, loudness constancy has been demonstrated. In many ways this is analogous to size constancy in vision, where intensity at-the-ear represents retinal size and sound source power represents physical object size (Zahorik & Wightman, 2001). In anechoic environments, when the intensity of auditory signal is the only variable manipulated, loudness constancy is not demonstrated as judgements of louder and closer can be interchangeable (Zahorik et al., 2005). However, often auditory signals do not only have intensity cues but also other cues such as frequency and reverberation (Kolarik et al., 2016; Zahorik et al., 2005). The direct to reverberant ratio of sound energy has been found to relay relative source distance information (Bronkhorst & Houtgast, 1999; Mershon & King, 1975; Zahorik, 2002a). Investigating whether this information may in turn influence source loudness estimates, Zahorik and Wightman (2001) found that within a simulated reverberant environment participants do have the capacity to display loudness constancy. Here, without visual cues, participants were able to estimate invariant source loudness despite the amplitude of the signal at-the-ear varying due to modulation of the sound source's distance. Unexpectedly, this study also found that distance estimates were more biased than source loudness estimates. Based on this, it was suggested that reverberant energy retained in an auditory waveform independently relays information about the distal power of a sound source. In support of this, a subsequent study reported that loudness constancy was only present in a strong but not weak reverberant environment; however, sound source distance estimates were not reliably affected by the changes to the prevailing sound field (Altmann et al., 2013). This finding was consistent with the suggestion of Zahorik and Wightman (2001) of a dissociation between distance and loudness estimates.

Nonetheless, it is still possible that improving the localisation of a sound source may facilitate loudness constancy. A natural hypothesis is if we can determine the egocentric distance of a sound emitter, we may be able to better estimate the distal power of that sound source by discounting the distance-based intensity attenuation that the proximal signal has undergone. Vision is one such mode that offers the capacity to more accurately determine the location of a sound source (Anderson & Zahorik, 2014). The possibility that visual information about source distance is harnessed when determining the power of an auditory source has been scarcely investigated and is yet to be resolved. Mohrmann (1939) positioned a pair of loudspeakers at different distances and required participants to equalise the source volume between a close speaker (0.75m) and a comparison speaker that varied in distance (2.37 or 7.5m). Judgements were taken estimating either the intensity of the sound at-the-ear or at-the-source (i.e., source power). It was found that constancy was higher when estimating the intensity of signals at-the-source over the intensity of input at-the-ear. Participants also demonstrated partial loudness constancy in darkness but demonstrated a higher degree of loudness constancy when the scene was visible. The presence of partial constancy in darkness suggested that unimodal auditory cues may have contributed to the ability to gauge the power of the source in the dark. Yet, the increase in constancy with the presence of a visible sound source suggested that visual cues played a role in loudness estimates on top of other auditory cues present. Following this, von Fieandt (1951) & Shigenaga (1965) conducted studies that replicated components of Mohrmann's (1938) method of comparison. In a music studio and on an outdoor rooftop respectively, these studies asked participants to estimate the distal power of visible speakers at different distances. Estimation was obtained through the adjustment of the intensity of the close speaker. They again found that participants were able to display loudness constancy and approximately estimate distal source power.

In a subsequent study Mershon et al. (1981) noted that all previous studies that investigated the effect of visually signaling source distance on loudness constancy, also physically varied the distance at which sounds were delivered. Thus, it was possible that auditory distance cues (such as the direct to indirect reverberant ratio) may have been present, and these cues may have systematically interacted with visual cues to facilitate partial constancy. Their study aimed to control for this confound by delivering sounds from a static location (through a hidden loudspeaker) whilst having a silent but visible dummy loudspeaker move between 3 distances (75, 225 and 375cm). Visual capture led the silent dummy speaker to be perceived as the sound source. Apparent loudness of sounds delivered from the dummy loudspeaker were taken through free modulus estimation. The experiment utilised 3 different presentation conditions: an anechoic room where the hidden sound source was far away (420cm), a reverberant room where the hidden source was far away (420cm) and a reverberant room with a close hidden sound source (60cm). To demonstrate a loudness constancy like effect, ratings of sounds would need to be louder as the apparent distance of the sound source increased. In both reverberant environments participants estimated the signal as increasing in loudness as the apparent distance of the source increased. In the anechoic environment there was a similar trend of loudness estimates increasing over distance, however between 225cm and 375cm there was a reversal of this affect and loudness estimates decreased. It was suggested that this was possibly because of a failure of visual capture with anechoic sounds generating location incongruencies.

More recently, Altmann et al. (2012) examined loudness constancy using short bursts of noise delivered via earphones in anechoic conditions. These bursts of noise were paired with the offset of a light source at varying distances (60, 120, 240cm). Loudness estimates were taken with a free modulus estimation procedure and participants were not specifically directed to attend to either the distal or proximal intensity. Results demonstrated no evidence of loudness constancy. Neural signals were also measured using magnetoecephalography (MEG). The furthest distance was associated with a larger N1m signal than the closer distance suggesting that despite finding no behavioural indication of loudness constancy there may have been a neural integration of the processing of sounds perceived to be farther away. Looking to improve visual depth cues, Berthomieu et al. (2019) examined whether loudness estimates were influenced by source depth in a virtual reality environment. The visual environment was a room with a speaker presented at 5 distances (1, 2, 4, 8, 16m). The speaker was obstructed by a panel wall in a non-visible condition. Noise bursts were simulated from the virtual sound source within 3 different sound fields; a sports hall, a concert hall and anechoic conditions. Apparent loudness and distance estimates were taken with a free modulus estimation procedure. It was found that participants' judgements were not influenced by visual cues to sound source depth. In virtual reality, Berthomieu et al. (2019) provided the most convincing control of auditory depth cues in reverberant and anechoic environments to date, and failed to demonstrate any visual influence on loudness constancy. However, it is noteworthy that the use of barriers to occlude visual cues in this study may have confounded the comparison between visual and non-visual loudness estimates. This is because barriers have been found to affect loudness estimates (Aylor & Marks, 1976).

An initial aim of our study was to replicate the audio-only loudness constancy effect using a different method to magnitude estimation. The only studies that have demonstrated auditory-only loudness constancy have involved delivering auditory signals in reverberant conditions and measuring loudness with a free modulus magnitude estimation procedure (Altmann et al., 2013; Zahorik & Wightman, 2001). Marks and Florentine (2011) noted that there has been substantial variability in loudness estimates depending on the method of measurement. Consequently, examining constancy estimates using a novel paradigm is valuable as it enables us to probe the robustness of previous constancy findings. Whilst there is not a perfect solution to measuring loudness without bias, adaptive 2IFC methods are a good option for recording sensory judgements as they are both sensitive and efficient (Leek, 2001). We used a Bayesian adaptive forced choice method which has been found to have both of these qualities (Kontsevich & Tyler, 1999).

The primary aim of our study was to resolve whether visual cues to source distance may play a role in influencing source loudness estimates. As described above, in experiments requiring the distal estimation of source power, findings have been mixed. Four studies found loudness constancy elicited from visual cues in reverberant conditions (Mershon et al., 1981; Mohrmann, 1939; Shigenaga, 1965; von Fieandt, 1951) and one study failed to find an influence of visual cues on apparent loudness estimates in reverberant conditions (Berthomieu et al., 2019). Altmann et al. (2012), Mershon et al. (1981), Berthomieu et al. (2019) and Chapter 3 of this thesis failed to find loudness constancy elicited from visual cues in anechoic conditions. Thus, all the evidence supporting the influence of visual cues on loudness constancy has been generated within environments in which reverberation was present; this effect has not been observed within anechoic environments. It could be that reverberation is needed to facilitate the binding of audio and visuals signals at different depths. It has been found that participants are sensitive to real world regularities while interpreting reverberant noises (Traer & McDermott, 2016). Further, anechoic sounds tend to be perceived as being closer than reverberant sounds (Butler et al., 1980; Mershon & King, 1975). For this reason, it is possible that reverberation is necessary for audio and visual information to be perceived as sharing a spatiotemporal location and to be integrated as a unitary event. Additionally, it may be that participants are inclined to demonstrate higher constancy under conditions in which they are directed to estimate intensity at-the-source rather than at-the-ear (Mohrmann, 1939). While Berthomieu et al. (2019) failed to find an influence of visual cues on the *apparent* loudness of sounds in reverberant conditions, we look to establish whether *distal* source power estimates are influenced by visual cues in reverberant conditions.

A further point of consideration is that we are cuing the visual distance of the sound source using photographs displayed on a 2D computer monitor. Critically, we believe this is a suitable method as photographs have previously been shown to produce reliable and accurate estimates of source distance (Anderson & Zahorik, 2014). Furthermore, in the present experiment, we have employed the exact stimuli as those used by Anderson and Zahorik (2014). Anderson and Zahorik (2014) measured participants egocentric depth estimates of a loudspeaker in a hall depicted at different distances in photographs. The photographs were viewed on a HDMTV. Sounds were synthesised with binaural room impulse responses (BRIR) of the speaker at each respective distance and delivered via headphones. Participants were to estimate the egocentric distance of the speaker in units of either meters or feet. The presentation distance of the speaker ranged from 0.3 to 9.8 m. In this experiment participants' perceived estimate of distance of the loudspeaker was linked to the actual physical distance of the loudspeaker by the exponent 0.92. Not only did this demonstrate accurate estimates of source distance but further this exponent was very similar to the exponent derived from studies in which participants were required to estimate target distances in real world settings. In a review of distance perception in open fields, Da Silva (1985) found that the mean exponent linking perceived distance with physical distance to be 0.99.

We hypothesise that in the audio-only condition, reverberant cues present in the auditory signal will facilitate the estimation of source loudness such that partial loudness constancy will be demonstrated. Secondly, we hypothesise that the visibility of a sound source's location will improve participants' ability to accurately estimate source loudness. To further disentangle this relationship we systematically shifted the visual distance of the speaker to be closer or farther than the distance of the auditory cue. If the visual distance of a sound source is accounted for when estimating source loudness, we hypothesise that the visual-nearer condition should shift estimates to be softer compared to the congruent audiovisual condition; conversely, the visual-farther condition should shift estimates of the sound source to be louder than the congruent audiovisual condition.

4.3 Methods

Participants

A total of 159 participants were recruited from a pool of students enrolled in an introductory psychology course at UNSW Sydney. Participants received course credit for their involvement and gave informed and written consent in accordance with the experiment protocols approved by the Human Research Ethics Advisory Panel in the School of Psychology, UNSW Sydney (#2683). All participants were naïve to the purposes of the experiment.

Apparatus

Auditory stimuli were presented via one of two identical 'AudioFile' devices (Cambridge Research Systems, Kent, UK) and over-ear headphones (Beyerdynamic, Heilbronn, Germany; model DT990 Pro). The sound level produced by the headphones was determined using an artificial ear, microphone, and analyser (Brüel & Kjæer, Nærum, Denmark; models 4152, 4144, and 2250, respectively).

Visual stimuli were presented on a Display++ LCD monitor (Cambridge Research Systems, Kent, UK) with a spatial resolution of 1920 × 1080 pixels, temporal resolution of 120Hz, and mean luminance of 60 cd/m^2. The relationship between the video signal and monitor luminance was linear. Participants viewed the monitor in one of two darkened rooms from a distance of 54cm, via a chin rest, for a visual angular subtense of 73.7° × 41.5°. The experiment was controlled using PsychoPy (Peirce, 2007, 2008).

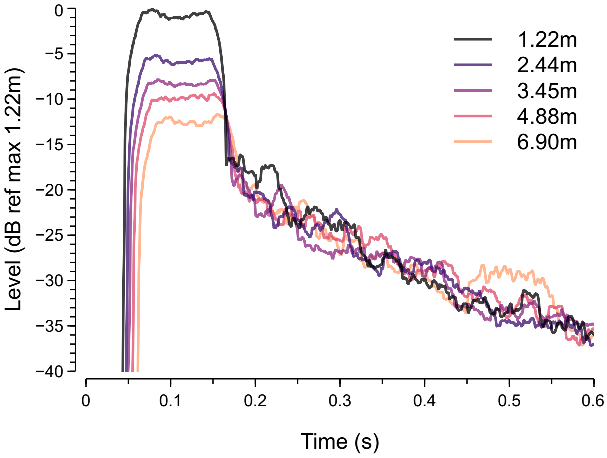

Figure 4.1. The average difference in proximal sound level over time. The difference in level is relative to the peak of the signal received from the reference speaker. The colour of each line represents the distance of the speaker.

Stimuli

The auditory and visual stimuli were obtained from the study reported in Anderson and Zahorik (2014). Briefly, Anderson and Zahorik (2014) positioned a sound source at a range of distances within a large concert hall and collected binaural room impulse responses (BRIRs) and photographs of the scene. Samples of white noise (100ms duration) were convolved with the BRIRs to produce waveforms with properties consistent with being produced from the sound source position within the concert hall. The level over time of the unaltered waveforms is shown in Figure 4.1. The photographs of the sound source (i.e., the loudspeaker) at each distance are shown in Figure 4.2. For further details of the BRIR and photograph acquisition and environment characteristics, see Anderson and Zahorik (2014).

For use in the current study, we first resampled the post-convolved waveform for each distance to 44.1kHz (from 48kHz). We then created a library of sounds by multiplying the waveform for each distance so as to modulate its level by up to ±15 dB (in 0.25 dB increments). This was used to simulate the increase or decrease in the level of the sound source. The visual depictions of each sound source distance were then cropped and resized to show the field of view that was consistent with the computer monitor at the viewing distance.

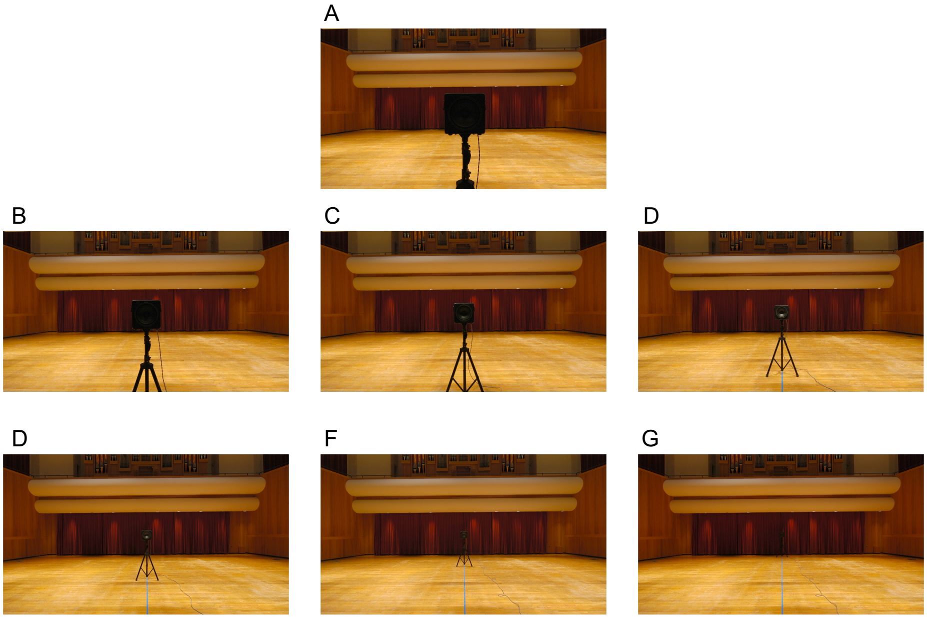

Figure 4.2. The Visual stimuli in this experiment consisted of photographs of a loudspeaker appearing at different distances from the observer in Comstock Hall. The reference speaker appeared at **(A)** 1.22m, while the comparison speakers appeared at **(B)** 1.72m, **(C)** 2.44m, **(D)** 3.45m, **(E)** 4.88m,**(F)** 6.90m and **(G)** 9.75m.

Design and Procedure

The study consisted of four between-subjects conditions, where each condition had a within-subjects manipulation of comparison source distance (four levels). The levels of the comparison source distance factor were always 2.44m, 3.45m, 4.88m, and 6.90m in the auditory modality. These comparison source distances corresponded to 1, 1.5, 2, and 2.5 doublings of the reference source distance (1.22m). As shown in Table 4.1, the distances of the comparison source varied in the visual modality for the visual-closer and visual-farther conditions. The first two sets performed the task either *with* (audiovisual) or *without* (audio-only) a visual depiction of the sound source and the acoustic environment. Two follow up sets of participants performed a task similar to the previous audiovisual condition in all respects, except that the visually depicted sound source was systematically nearer (visual-nearer) or farther (visual-farther) than the concurrent auditory signal.

Table 4.1. Condition properties

Condition	Modality	Units	Reference	Comparison			
All conditions	Auditory	Metres	1.22	2.44	3.45	4.88	6.9
		Doublings	0	1	1.5	2	2.5
Audio-only	Visual	Metres	-	-	-	-	-
		Doublings	-	-	-	-	-
Audiovisual	Visual	Metres	1.22	2.44	3.45	4.88	6.9
		Doublings	0	1	1.5	2	2.5
Visual-nearer	Visual	Metres	1.22	1.72	2.44	3.45	4.88
		Doublings	0	0.5	1	1.5	2
Visual-farther	Visual	Metres	1.22	3.45	4.88	6.9	9.75
		Doublings	0	1.5	2	2.5	3

The procedure for a given participant was conducted in a single session lasting approximately one hour. The session consisted of a series of four runs, where each run assessed two levels of the comparison source distance factor. The combinations were arranged such that the levels contained one of each of the two distances closest to the reference and the two distances farthest from the reference. The ordering of runs was randomised for each participant, and there was a self-paced break of at least 30 seconds between each run and halfway through each run.

Each run consisted of a series of trials, where each trial consisted of a temporal two-interval forced-choice task. Each interval began with a 750ms preparatory period in which the screen was uniformly mid-grey. In the conditions with visual presentations, the image with the loudspeaker at the appropriate position was then presented for 700ms, with the image opacity increasing linearly to complete visibility over the first 200ms. The auditory stimulus was then delivered while the image remained visible for the following 2200ms, with the opacity decreasing linearly to complete transparency over the last 200ms. This procedure was then repeated for the second interval.

On each trial, one of the intervals contained the sound source at the reference distance (1.22m) and at the reference level (74dB LAFmax) while the other interval contained the sound source at the comparison distance and at the comparison level. The interval containing the reference was randomised on each trial. Following the presentation of the two intervals, a written prompt appeared "Which sound was produced by a loudspeaker with a higher volume setting? Press the left arrow key for the first sound. Press the right arrow key for the second sound". The next trial commenced subsequent to the participant's button press, with a minimum inter-trial interval of 3s. The experiment session for a given participant produced 240 data points (60 per comparison distance), where each data point consisted of the level of the comparison sound (in dB, referenced to the unaltered sound from the comparison distance) and a response indicator (whether the participant considered the comparison sound to be louder than the reference sound).

The level change of the comparison sound on each trial was determined using a Psi adaptive staircase procedure (Kontsevich & Tyler, 1999). Each run contained two separate staircases, one for each different comparison distance. Each staircase consisted of 30 trials, and the staircase order was randomised within each run. As part of the staircase procedure, participant responses were modelled via a logistic-based psychometric function that described the probability of selecting the comparison interval as containing the louder sound for a given comparison offset. The function had free parameters for the point of subjective equality (PSE; α) and the slope of the psychometric function (β) and fixed parameters for the lower (0.05) and upper (0.95) asymptotes. The PSE is the amount of artificial adjustment in level of the comparison sound source that was required for it to be perceived equally often as being louder and softer than the reference sound source.

Before commencing the session, the participant's dominant eye was determined using the 'card test' (described by Ehrenstein et al., 2005). This was used to adjust the location of an occluder attached to the chinrest such that the participant viewed the monitor through their dominant eye only. This monocular viewing was designed to remove the influence of binocular cues to the true depth structure of the testing booth and promote immersion in the depicted scene. Participants were then introduced to the task via a set of computer-based instructions before commencing the experiment.

Exclusions

Participants and their data were evaluated against a set of criteria that determined whether they were excluded from subsequent analyses. Participants were excluded due to equipment failure (1 participant), non-compliance with instructions (1 participant), and self-reported vision impairment (3 participants). Based on the raw trial responses, we excluded participants who had a bias towards responding the first or second interval (less than a 45% or more than a 55% probability of responding the second interval, across the experiment; 10 participants). We also fitted a model (see below for details on the general statistical modelling framework) containing parameters for the PSE and slope for each condition and repeat, separately for each participant, to assess the fundamental resemblance of the data to a broad family of psychometric functions. We excluded participants where: the mean uncertainty in the PSE estimates (parameterised as the width of the 95% highest posterior density intervals) was greater than 15 dB (18 participants); any of the differences in the mean PSE estimates between the two repeats of each condition were greater than 12.5 dB (14 participants) ; the geometric mean of the slopes across conditions was greater than 12.5 (17 participants); the maximum slope across all conditions was greater than 30 (24 participants); and problems were indicated in the model estimation process (6 participants). Most participants that were excluded did not satisfy more than one of the criteria. After such exclusions, there were 116 participants in total; 34 participants in the audio-only condition, 26 participants in the congruent audio-visual condition, 30 participants in the visual-closer condition, and 26 participants in the visual-farther condition.

Statistical approach

The experiment session for a given participant produced 240 data points (60 per comparison distance), where each data point consisted of the level of the comparison sound (in dB, referenced to the unaltered sound from the comparison distance) and a response indicator (whether the participant considered the comparison sound to be louder than the reference sound). We are interested in the PSE parameter extracted for each comparison speaker at each distance. The PSE parameter captures the artificial level change to the comparison sound that is required to be perceived as equally loud as the reference sound. To extract PSEs for each participant, the trial-wise responses were broken into runs and modelled as Bernoulli events in which the probability of indicating that the comparison sound was perceived to be louder than the reference sound was given by a cumulative normal psychometric function. This function had PSE, spread, and lapse rate parameters. The PSE & spread of the psychometric functions produced separate estimates for each participant, speaker distance and repeat. The PSE for a given participant, speaker distance and repeat was assumed to be drawn from a normal distribution with a mu of 0 and SD of 10. The spread for a given participant, speaker distance and repeat was assumed to be drawn from a parent log normal distribution, with a mu of $\log 7.5$ and SD of 1. The lapse rate parameter sets the upper and lower asymptotes of the psychometric function, and was given a fixed value of 5%. A PSE of 0 indicates that participants have displayed perfect constancy and accurately estimated the power of the source at its location. In this case, the comparison speaker would not require any artificial adjustment to be perceived as having equal amounts of power as the reference. Conversely, if participants display no constancy and base their loudness estimates on the direct component of the waveform reaching their ear, the comparison speaker will require a positive adjustment at PSE. In this case, the comparison speaker needs to artificially increase its power as its distance increases. This is because the waveform attenuates in intensity over distance travelled. Specifically, as seen in Figure 4.1, the comparison speaker would be required to increase by approximately 6 dB at 2.44m, 8 dB at 3.45m, 10 dB at 4.88 m and 13 dB at 6.9m.

The first aim of this study was to identify whether the PSEs for each comparison distance in the audio-only condition reflected perfect constancy. To estimate the level of constancy expressed we extract the mean PSE and the 95% confidence interval around that mean for each comparison distance. A one-way repeated measures analysis of variance (ANOVA) was run on the audio-only condition to assess whether there was a change in the degree of constancy across comparison speaker distance. When a violation to the assumption of sphericity occurred, a Greenhouse-Geisser correction was applied. Statistical significance was assessed against a Type 1 error rate of 0.05. The second and primary aim of this study was to estimate the effects of varying sound source distance on loudness under different conditions of audiovisual presentation. To investigate the effect of viewing a speaker distances on mean PSE estimates we conducted a (4) x 4 mixed ANOVA. This analysis had a within-subject factor of comparison speaker distance for the auditory signal and a between-subject factor of visual presentation condition. Again, when a violation to the assumption of sphericity occurred across within-subject conditions, a Greenhouse-Geisser correction was applied. Statistical significance was assessed against a Type 1 error rate of 0.05.

4.4 Results

Figure 4.3 shows the mean change of the comparison at PSE (dB) at each auditory distance (m). Each visual-presentation condition is split into a different coloured symbol. PSEs that deviate from a comparison change value of 0 reflect a deviation from perfect loudness constancy. PSEs based only on the intensity of the direct waveform reaching the ear will require a comparison change of approximately 6 dB at 2.44m, 8 dB at 3.45m, 10 dB at 4.88m and 13 dB at 6.9m.

First, we identified whether the PSEs at each comparison distance for the audio-only condition reflected perfect constancy. The closest comparison at 2.44m required a mean adjustment of 1.846 dB, 95% CI $[1.071, 2.621]$, the comparison at 3.45m required a mean adjustment of 2.955 dB, 95% CI $[1.946, 3.963]$, the comparison at 4.88m required a mean adjustment of 3.721 dB, 95% CI $[2.731, 4.712]$ and finally, the farthermost comparison at 6.9m required a mean adjustment of 3.793 dB, 95% CI $[2.374, 5.212]$. Comparison sounds at PSE required between 4-8 dB less intensity than the direct component of the reference waveform. However, these comparison sound estimates required a further 2-4 dB reduction in intensity to reflect perfect source power estimation. This indicated that participants displayed partial loudness constancy. Following this, a one-way repeated measures analysis of variance (ANOVA) was run on the audio-only condition. This was to assess whether there was a change in the degree of constancy across comparison speaker distances (2.44, 3.45, 4.66, 6.9m). There was a significant main effect of distance ($F_{1,33} = 85.549$, $p < 0.001$, $\eta_p^2 = 0.211$). Paired t-test comparisons revealed that constancy decreased between 2.44m and 3.45m ($MD = -1.109$, $SE = 0.409$, $p = 0.009$), trended towards a decrease between 3.45m and 4.88m ($MD = -0.767$, $SE = 0.388$, $p = 0.057$) and did not significantly change between 4.88m and 6.9m ($MD - 0.72$, $SE = 0.509$, $p = 0.889$). This indicated that loudness constancy decreased up until the furthermost comparison distance.

Second, we identified whether the PSE for each comparison distance differed based on the visually cued distance of the sound source. To investigate this we employed a (4) x 4 mixed ANOVA, with a within-subject factor of auditory distance (2.44, 3.45, 4.66, 6.9m) and a between-subject factor of visual-presentation condition (audio only, congruent audiovisual, visual-closer, visual-farther). We find a significant main effect of distance ($F_{3,336} = 6.301$, $p = 0.02$, $\eta_p^2 = 0.053$) and of visual-presentation condition ($F_{3,112} = 4.084$, $p = 0.009$, $\eta_p^2 = 0.099$). We also find a significant interaction effect between distance and visual-presentation condition ($F_{9,336} = 2.175$, $p = 0.041$, $\eta_p^2 = 0.055$). This indicated that the effect of auditory source distance on loudness judgements differed based on the visual-presentation of source distance. Simple effects were used to further examine the main-effect of visual-presentation condition. To assess the influence of congruent visual information on source loudness estimates, the congruent audiovisual condition was compared against the audio-only condition. We did not find evidence that the difference between the auditory-only and congruent audiovisual condition ($MD = -0.582$, $SE = 0.878$) was statistically significant ($F_{1,58} = 0.439$, $p = 0.510$, $\eta_p^2 = 0.008$). To assess whether biasing visual source distance information influenced loudness estimates, we compared both the visual-closer and the visual-farther conditions against the congruent audiovisual condition. The difference between the visual-closer and congruent audiovisual condition ($MD = 0.633$, $SE = 0.955$) was not statistically significant ($F_{1,54} = 0.439$, $p = 0.511$, $\eta_p^2 = 0.008$). However, the difference between the visual-farther and congruent audiovisual condition ($MD = -2.357$, $SE = 1.074$) was statistically significant ($F_{1,50} = 4.815$, $p = 0.033$, $\eta_p^2 = 0.088$). This indicated that the comparison speakers in the visual further condition needed less volume to reach PSE with the reference than the congruent audiovisual condition.

Following this, we unpacked the visual-presentation condition and auditory distance interaction using pairwise comparisons. Pairwise interactions were completed by computing difference scores between two distance combinations and then conducting independent sample t-tests on these differences. First, we examined the audio-only and congruent audiovisual condition. We do not find significant pairwise interactions at any distance combination. Next, we examined any interactions between the congruent audiovisual, visual-farther and visual-closer conditions. We find a significant interaction

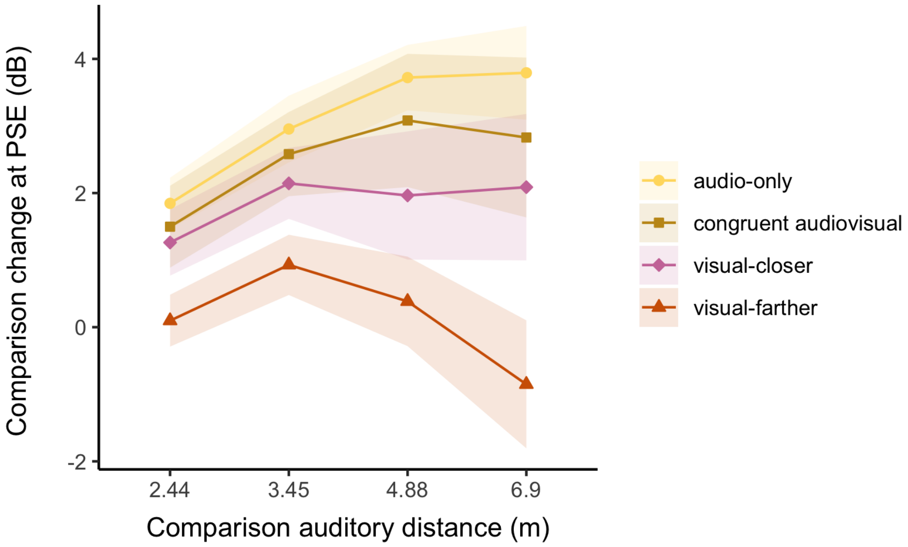

Figure 4.3. Mean PSEs. Each coloured symbol represents a visual-presentation condition with shading indicating standard error. The x-axis captures the distance of the comparison speaker when simulating the auditory signal. The y-axis represents the amount a comparison speaker had to be artificially adjusted, at its location, to be perceived as loud as the closer reference speaker.

between the visual-farther and congruent audiovisual condition at 2.44 and 6.9m $(MD = -2.279, SE = 1.057, F_{1,50} = -4.652, p = 0.036)$. In the visual-farther condition, the comparison speaker at the furthermost distance had a lower PSE relative to the closest comparison, whilst in the congruent audiovisual condition, the comparison speaker at the furthermost distance had a higher PSE relative to the closest comparison. We do not find significant pairwise interactions at any other distance combination across these three conditions.

4.5 Discussion

The present study examined judgements of source loudness when both auditory and visual cues to a sound sources distance were manipulated. Using stimuli acquired from a concert hall we simulated a reverberant scene with a loudspeaker delivering sounds at different distances. First, we looked to quantify the degree of constancy displayed in auditory-only conditions. Second, we looked to assess whether loudness constancy was increased with the inclusion of ecologically congruent visual information. Third, we looked to explore whether visibility plays *any* role when making source loudness judgements by systematically biasing the visual cues to depict the comparison sound source as being closer or farther than the paired auditory signal.

First, we examined the effect of auditory-only cues on source loudness estimates. To date there have been two studies that have specifically gauged loudness constancy in reverberant conditions using only the auditory signal (Altmann et al., 2013; Zahorik & Wightman, 2001). Zahorik and Wightman (2001) and Altmann et al. (2013) used magnitude estimation procedure requiring participants to estimate source power. Zahorik and Wightman (2001) found near perfect constancy and Altmann et al. (2013) found near perfect constancy within 'strong' reverberant conditions but a failure of constancy in 'weak' reverberant conditions. In the present study we elected to use a Psi adaptive staircase 2IFC procedure. We found that participants could partially account for the distance-related intensity attenuation of the farther comparison source. The PSE occurred when the comparison sound source produced sounds that were 4-8dB less than the reference sound source. However, the comparison sound source required a further 2-4dB reduction to reflect an accurate estimation of its source's power (i.e., to reflect the point at which the speaker required no artificial adjustment at its location). Thus, when completing the auditory-only condition, participants underestimated comparison source power and demonstrated only partial loudness constancy. This finding is in contrast to Zahorik and Wightman (2001) and Altmann et al. (2013) who found near perfect constancy. It could be that the difference in our finding was driven by changing the method of measurement from magnitude estimation to 2IFC. Another possibility is that the difference is driven by the amount of reverberation within each sound field. Zahorik and Wightman (2001) simulated the BRIR of a virtual sound source in a hall in which the time it took for the signal to decay by 60 dB (T_{60}) was approximately 0.7 seconds. Altmann et al. (2013) used a reverberation chamber in which the weak reverberation condition had a T_{60} of approximately 0.14 seconds and the strong reverberation condition had a T_{60} of approximately 1.03 seconds. The BRIRs from the concert hall we employed had a reverberation time T_{60} of 1.9s (Anderson & Zahorik, 2014). Comparing these studies we can see, there was a failure of constancy in the sound field with the weakest reverberation (T_{60} = 0.14s), there was near perfect constancy in sound fields with medium amounts of reverberation (T_{60} =1.03, 0.7s) and there was partial constancy in the sound field with the most reverberation (T_{60} = 1.9s). It is possible that there is an optimal level of reverberant signal that facilitates accurate source loudness perception. Future studies will need to disentangle whether these constancy differences are the result of the type of reverberation present in a sound field.

Next, we examined the effect of visual distance cues on loudness estimates. We predicted that source power estimates would improve with the inclusion of congruent visual cues. The rationale was that by visually increasing the accuracy of source distance representations, one could then more accurately account for variation in a physical signal at the ear due to source distance. Our simple effect comparisons did not demonstrate significant evidence for a hypothesis that the congruent visibility of a source's distance altered the degree of loudness constancy relative to the audio-only condition. It is possible that visual depth cues did not produce a decisive improvement in the capacity of participants to estimate source loudness because reverberation was already facilitating a ceiling level of loudness constancy. In this case, a ceiling level of constancy established by auditory cues would mask any influence of visual cues on loudness.

To further dissect whether visual cues play a role in source loudness estimation, we shifted visual cues of the sound source to be either systematically closer or systematically farther than the concomitant auditory signal. Simple effect comparisons revealed that shifting the sound source to be farther increased the degree of constancy relative to the congruent audiovisual condition. When the source distance of the comparison speaker appeared farther, the comparison speaker needed a lower adjustment value than the congruent audiovisual condition (i.e., the comparison sound source was perceived to be louder in the visual-farther condition). Surprisingly, simple effect comparisons did not demonstrate evidence that shifting the sound source to be closer altered the degree of constancy relative to the congruent audiovisual condition. This effect is difficult to interpret as if participants were using visual estimates to facilitate loudness constancy, the visual-nearer condition should have returned comparison speaker sounds that required higher adjustment values (i.e., where the sound source would be judged as being relatively softer) compared to the congruent condition. That is to say, if visual source depth was a cue that was harnessed to appropriately inform source power estimates, the shifting of the speakers in opposing directions should in turn shift source power estimates in opposing directions. In fact, the visual-closer condition had means PSEs that were in the direction of lower adjustment values (i.e., the sound source was judged as being relatively louder) than both the congruent audiovisual and audio-only conditions, although this difference was non-significant. One potential explanation is that a predictive coding mechanism is driving both incongruent conditions to have comparison sound sources that appear louder. A predictive coding account posits that our brain is constantly making top down predictions of bottom up sensory input (Friston, 2009, 2012). Any difference between prediction and input generates a prediction error, and this error manifests as an increase in the strength of neural activation. This prediction error process was theorised to aid sensory systems in verifying and correcting representations of the external environment. The degree of neural activation from prediction errors has been suggested to also influence the perceived strength of a signal (Hughes et al., 2013). Therefore, it is a possibility that in both incongruent conditions, participants are generating prediction errors between the expected auditory waveform based on the visual distance cue, and the received auditory distance cue. The prediction errors may in turn amplify the perceived loudness of the comparison signal.

Our findings demonstrate that we do not appropriately use source distance cues to disambiguate source power. It has been suggested that in multisensory integration we weight sensory cues by their relative reliability (Alais et al., 2010). One hypothesis explaining our findings may be that visual distance cues do not provide relatively reliable estimates of source power, and therefore are not weighted to have a meaningful influence. Transforming proximal signals into a representation of distal sound source power using source distance cues may be dependent on multiple parameters. For example, in anechoic conditions, every time distance is doubled a signal tends to decrease by 6dB (Coleman, 1963), however in reverberant sound fields it can be attenuated by less (Zahorik, 2002a). Further, whether a sound source projects its signal uniformly or in one specific direction can also affect attenuation (Kolarik et al., 2016), as can environmental factors such as wind (Traunmüller & Eriksson, 2000). Alternatively, visual source power cues may be a more precise indicator of distal intensity. It has been found that visual cues to source power can influence loudness estimates (Chapter 5; Rosenblum & Fowler, 1991), and that visual cues to source power can also influence the brain's initial auditory evoked response in accordance with expectations (Chapter 6). These studies did not manipulate source distance and consequently did not disentangled whether source power information influences the representation of the apparent intensity of auditory signals, independent of distance. Future research will need to resolve this question.

A mechanistic theory is that, loudness estimates are influenced by vision through the organisation of input with a top down, anticipatory neural template (van Laarhoven, Stekelenburg, & Vroomen, 2017). In the EEG and MEG literature it has been suggested that to record the early integration of visual information in the auditory areas, visual information needs a predictive 'head start' (>100ms) when cuing a sound (Aoyama, Endo, Honda, & Takeda, 2006; Senkowski, Saint-Amour, Kelly, & Foxe, 2007;

Vroomen & Stekelenburg, 2010). In this account, the neural mechanism driving an auditory interaction requires a precise predictive template to be formed prior to the reception of the sound. For this template to be predictive it requires both temporal precision (van Laarhoven et al., 2017) and possibly also intensity precision as established by cuing both source distance and sound source power. In the present study, it was impossible to form a precise predictive template in anticipation of a sound because there were not predictive temporal cues, nor were there source power cues. To begin exploring this account, future studies may increase the precision with which source power is visually predicted by providing temporal onset and source power cues in the presence of a source distance manipulation.

A potential limitation of the present study is a failure of our visual stimuli to cue sound source depth. However, it is noteworthy that using the exact same stimuli, Anderson and Zahorik (2014) demonstrated that visual cues increased the accuracy and reduced the variance of participants auditory depth estimates. Nonetheless while these stimuli have supported accurate and reliable estimates of sound source depth, it could be that such stimuli only support the cognitive estimation of source depth and fail to support the perceptual integration of the objects at each physical distance. While we did not employ binocular depth cues, we note that there are striking similarities between our results and Berthomieu et al. (2019) who provided binocular cues to depth through the use of virtual reality. Although Berthomieu et al. (2019) required participants to judge 'apparent loudness' and not source volume, they also did not find evidence of a significant influence of visual cues on loudness estimates in either anechoic or reverberant environments.

In conclusion, this paper supports the notion that we use reverberant energy provided in an auditory signal as a cue to estimate source loudness. The sound field of this concert hall facilitated partial loudness constancy. We did not find evidence that the congruent visibility of a sound source's distance had any pronounced effect on source power estimates. We also found that systematically biasing the visual distance of a sound source shifted source loudness estimates to increase regardless of whether the visual manipulation made the source appear closer or father than its auditory signal. This final manipulation provides further support for the notion that we do not appropriately employ visual signals to account for distance-based variation in sound source intensity.

5 Visual Cues to Source Power & Loudness (behavioural experiment)

Seeing hands and hearing claps: seeing the power of a sound source modulates perceived loudness

Author contributions:
Conceptualisation: SL, DJM, TJW. Stimuli: SL. Methodology: SL, DJM, TJW. Programming: DJM. Data collection: SL. Data analysis and presentation: SL, DJM, TJW. Writing - original draft: SL. Writing – review and editing: SL, TJW, DJM. Supervision: TJW, DJM.

Preamble:

The intensity of an auditory signal at-the-ear depends on both the power of the sound source and the distance of the source from the listener. In the previous two Chapters we presented evidence suggesting that visual cues as to the distance of a sound source do not seem to influence the perceived loudness of the auditory signal. In the following Chapter we shifted to explore whether visual cues as to the *power* of a sound source could influence the perceived loudness of the auditory signal.

5.1 Abstract

An auditory event is often accompanied by characteristic visual information. For example, the sound level produced by a vigorous handclap relates to the speed of the hands as they move toward collision. Here, we tested the hypothesis that visual information about the power of a sound source is capable of altering the perceived loudness of auditory stimuli. To do this we utilised a psychophysics task to measure loudness judgements of audiovisual handclaps that minimized demand characteristics and, in turn, response bias. Specifically, we used a two-pair forced-choice task, with each pair consisting of a probe and an anchor. The probe was always a non-visible clap and was the same sound level across both pairs in a trial. The sound of one of the anchors was presented with a video depicting a handclap; the sound of the other anchor was not presented with a visible handclap. The visible handclap depicted the production of either a 'strong' clap or a 'weak' clap. The task was to judge in which pair the probe clap was more similar in loudness to its anchor clap. The sound level of the visible clap anchor when depicting a 'strong' clap was always 75dB, while the sound level of the visible clap anchor when depicting a 'weak' clap was always 65dB. The sound level of the anchor paired with the non-visible clap was always ±10dB the visible anchor. From trial to trial, the probe varied in sound intensity between the ranges provided by the two anchors. A Psi adaptive staircase determined the sound level of the probe. The key prediction was that the point of subjective loudness equality (PSE; the point that is perceptually equidistant in loudness from the two anchors) would be relatively increased when the video depicted a 'strong' effort handclap, and relatively decreased when the video depicted a 'weak' effort handclap. We found that the 'strong' visible clap had a PSE that was relatively increased compared to the 'weak' visible clap. This suggests that loudness percepts are constructed through the combination of visual expectations and auditory signals.

5.2 Introduction

Often we do not only experience sounds in isolation, but instead we also *see* the relationship between an object's movement and its auditory consequences. For example, when someone is whispering they will narrowly open their lips, while a shout will involve a wide open mouth. Similarly, a soft clap will involve a slow, weak movement, whilst a loud clap will involve a fast, powerful movement. There is an association between the force at which two objects collide and the intensity of the resulting signal: collisions between fast-moving objects will tend to cause more intense sounds than collisions between slow-moving objects. If the function of the perceptual system is to produce useful models of the external world, visual cues may provide information about the intensity of forthcoming sounds that are synthesised into the perceptual experience of auditory intensity.

In other domains of psychoacoustic research it has been established that vision plays a useful role in representing auditory input. Vision has been found to aid in the identification of the *content* of auditory stimuli (e.g., Campbell, 2007; Erber, 1975; Reisberg et al., 1987; Sumby & Pollack, 1954). The 'McGurk effect' has demonstrated that the perceived content of auditory syllables can be altered when viewing someone's lips producing a different utterance. When watching lips generating a certain phoneme (such as /fa/) and hearing a different phoneme (such as /ba/), the brain will resolve the similar but conflicting stimuli by fusing the information from the auditory and visual streams to generate a new percept (such as /va/) (McGurk & MacDonald, 1976). It has been suggested that the association between seen facial movements and heard speech drives this effect (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Thomas & Jordan, 2004; Yehia, Kuratate, & Vatikiotis-Bateson, 2002). Somewhat analogous to the McGurk effect, the present study will examine whether the brain generates an auditory percept through the fusing of associated information from the auditory and visual streams. The current study differs from previous McGurk-like studies in that it will focus on the impact of merged sensory stream information on signal *intensity* as opposed to signal *identity.*

Within the auditory domain, studies have provided preliminary (but contested) evidence that our auditory system uses causal cues when representing the loudness of auditory input. It has been found that increasing the vocal effort of speech stimuli can increase loudness estimates, even when the sound intensity is fixed (Allen, 1971; Brandt, Ruder, & Shipp Jr, 1969; Lehiste & Peterson, 1959; Mendel, Sussman, Merson, Naeser, & Minifie, 1969). Likewise, it has been found that when the vocal effort of speech stimuli remain fixed, but are delivered at increasing sound intensities, the loudness exponent (i.e., the rate at which loudness increases as sound intensity increases) is reduced as compared to when the vocal effort increases with sound intensity (Brandt et al., 1969). A potential explanation for these findings is that loudness percepts were biased towards recovering the distal features of a speech event (namely, the power of the sound source) based on the amount of vocal effort present (Fowler & Rosenblum, 1991).

However, it has also been proposed that the spectral properties of auditory signals differ as a function of vocal effort, and these basic stimulus differences are a confound in previous studies. This notion is supported by studies that modeled the physiological excitation at the auditory nerve generated by speech produced with differing vocal efforts. These models were able to account for the pattern of loudness judgements based on the different spectral properties of each vocal effort (Brandt, 1972; Glave & Rietveld, 1975). Across these paradigms, the observed inter-dependency between causal cues and spectral properties makes it difficult to disentangle whether higher order inferences or low order acoustic features are the driving cause of this vocal-effort effect.

While there have been several studies in which vision has captured and influenced audition (Alais et al., 2010), to my knowledge there has only been one behavioural study that has investigated the consequences of visually cuing source power on loudness judgments. Using both speech and non-speech stimuli (hand claps), Rosenblum and Fowler (1991) required people to (a) rate the amount of perceived effort put into the generation of the sound, and (b) rate the loudness of the sound when paired with the same visual stimuli. This visual manipulation allowed for the comparison of physically identical auditory signals with different effort cues. Loudness measurements were recorded with a variant of the magnitude estimation procedure (Marks, 1979). When auditory stimuli were paired with a video of a sound emitter that was perceived to be putting in more effort, perceived loudness ratings also increased.

There have been three other studies that have incidentally harnessed visual stimuli that convey auditory intensity information and measured loudness. M. Epstein and Florentine (2009, 2012) measured the binaural loudness summation of speech with and without visual cues. In this study, when changing from monaural to binaural listening conditions, the additive increase in loudness was significantly less when visual cues were present. One hypothesis that may explain these results is that the videos of the speaker (which all depicted the production of sounds at a fixed distance with ‘moderate vocal effort’), contributed to expectations that auditory input would arrive at a stable and ‘moderate’ intensity. If visual expectations influence the processing of auditory intensity, they may be integrated into loudness percepts such that there was a degree of loudness constancy between binaural and monaural listening conditions. Conversely, because source intensity cues were not directly manipulated, it is possible that factors unrelated to intensity expectations confounded this effect. For example, visual information that predicts the temporal onset of auditory input (i.e., a mouth moving before the onset of a speech sound) has been suggested to cause sensory attenuation and this may have influenced such a result (Besle, Fort, Delpuech, & Giard, 2004; Hughes et al., 2013). A study conducted by Aylor and Marks (1976) required subjects to judge the relative loudness of narrowband noise transmitted through different barriers (row of hemlock trees, slat fence, acoustic tile barrier or no barrier). This study had two conditions, one in which participants were blindfolded, and one in which participants were not blindfolded but there was a barrier that obscured the sound source. In the blindfolded condition, there were no differences between loudness estimates for any of the barriers. In the condition in which participants had no blindfold, loudness ratings were relatively attenuated when the barriers did not completely visually obstruct the sound source (i.e., slat fence, no barrier). Based on this finding, it was suggested that when a sound source was occluded by a barrier, participants expected the barrier to diminish the loudness of the auditory stimulus, which in turn raised their loudness estimates.

The studies described above provide tentative evidence supporting the notion that visual intensity expectations influence the perceived loudness of sounds. However, whether these effects are due to post-perceptual response biases, or a true perceptual effect remains a crucial issue to be resolved. Traditionally psychophysical measurements of loudness are taken from either (1) magnitude-production or loudness-matching tasks, where the requirement is to determine when a comparison stimulus is equally as loud as a reference stimulus (Marks & Florentine, 2011; Mohrmann, 1939; Shigenaga, 1965; von Fieandt, 1951), (2) an adaptive form of loudness matching tasks such as 2-interval forced-choice (2IFC) procedures (Silva & Florentine, 2006; Takeshima et al., 2001), or (3) magnitude estimation tasks such as free modulus estimation, where participants are asked to respond with a number that proportionally estimates the loudness of a sound (Stevens, 1956; Zahorik & Wightman, 2001). In the context of the present study, the awareness that a certain visually perceived action is associated with the production of a louder or softer sound (i.e., a 'strong' hand clap should produce a louder sound than a 'weak' handclap; see our manipulation check 5.3.3) may influence the participant's responses. This is because if the task requires participants to estimate which sound was more or less loud (as required by traditional psychophysical methods), the visual cue may provide the participant with a post perceptual 'answer' as to which sound was more or less loud that influences their response. For example, in a free modulus estimation when participants know a certain visual cue 'should' sound louder, they may cognitively inflate their ratings based on the visual explanation. Likewise, in a standard 2IFC method, participants are required to differentiate between two intervals by choosing which interval was louder; if an observer is perceptually uncertain about which interval was louder, they may base their decision (consciously or unconsciously) on the visual information cuing the louder sound. *If these uncertain decisions are driven by these cognitive assumptions and not the true experience of loudness, then the measured effect would be artificially inflated by response bias.* In an attempt to separate audiovisual integration from post-perceptual decisions, Rosenblum and Fowler's (1991) study included a discrepancy-rating between visual effort cues and auditory signals. They suggested that if participants fail to notice a discrepancy between audio and visual streams, then audiovisual integration has occurred and this would be an indication of a true perceptual effect. However, on trials in which the audiovisual stimuli had no discrepancy, when participants incorrectly judged that the audio was louder or softer than the visual cue, loudness ratings were also biased such that the audio was rated respectively louder or softer than when no discrepancy was reported. This effect was presumably due to the random distribution of sensory noise when processing stimuli. This indicated that using correct and incorrect discrepancy ratings to partition trials for 'true' perceptual integration could bias results by partitioning the effects of sensory noise on loudness. They recognised *'Sorting out whether occurrences of audiovisual integration are perceptual or cognitive is germane to all studies involving McGurk-type presentations... It is now clear, however, that a way is needed to determine whether an observed interaction between discrepancy group and video influence is a true indicator of the perceptual nature of the effects or is simply a by-product of the classification of loudness judgments by discrepancy rating'* (Rosenblum & Fowler, 1991, p.984).

In order to measure the influence of visual cues on loudness whilst addressing the problem of response bias, we developed a paradigm that employed the principles of Patten and Clifford (2015) and Jogan and Stocker (2014). Patten and Clifford (2015) shifted the demands of a 2IFC task measuring the tilt illusion to tease apart response bias from a true perceptual effect. Rather than examining in which interval the grating was shifted rightwards or leftwards (as is done in a traditional tilt illusion procedure), they asked participants to choose the interval in which a central test grating was closer to vertical. Patten and Clifford (2015) found that changing the design so the task demands did not coincide with changes to the surround reduced response bias. The design of the present study was adapted from Patten and Clifford (2015); it involved relaying two pairs of sounds and requiring a judgement of which pair contained sounds that were most *similar* in loudness. The second sound in each pair was either a visible or non-visible anchor; these anchors were set at fixed sound intensities. The first sound in each pair was a non-visible probe; this probe was identical for both pairs and varied in sound intensity across trials. We were interested in the point at which the probe was perceived as equally similar in intensity compared to each of the two anchors; this was

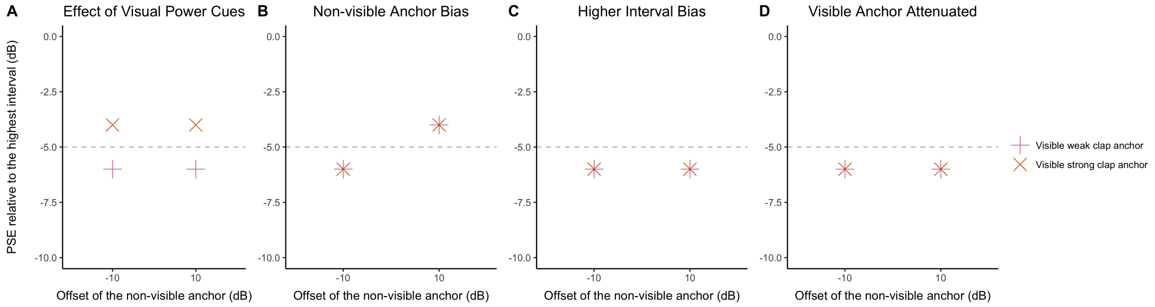

Figure 5.1. Mean PSEs for four potential scenarios. In all four panels, each coloured symbol represents the visual presentation condition, the x-axes capture the offset level of the non-visible anchor, the y-axes capture how much a PSE has reduced relative to the higher dB interval, and the dashed grey lines represent the points that are equidistant between the intensity of either anchor (in dB). These panels depict the direction of effects predicted by the scenario in which: **(A)** a true perceptual effect occurs where the 'strong' visual anchor is perceived as louder than the 'weak' visual anchor; **(B)** participants are biased towards choosing the pair containing the non-visible anchor; **(C)** participants are biased towards choosing the pair with the lower dB interval; **(D)** both visible anchors have their intensity attenuated.

labelled the point of subjective equality (PSE). The PSE was quantified as the point at which there was a 50% chance of choosing either pair as being more similar in intensity. In this task, if the visual anchor biases the clap sound to feel louder (or softer), the PSE of the probe should be respectively shifted up (or down). This is because if an anchor has its intensity amplified, the sound intensity at which the probe is equidistant in intensity between the two anchors also increases.

Critically, this approach no longer requires the estimation of which stimulus had a greater or lesser intensity, but rather which pair of stimuli were most similar in intensity. Looking for *similarity* is a shift away from more traditional 2IFC approaches that look for difference (e.g., traditional tasks may ask "which interval contained a louder/softer sound?"). This alteration shifts the task demands so that information contained in the visual cue cannot be used by the participant when making a decision on an uncertain trial. That is, the association of a cue as being louder or softer can no longer inform a participant's response because the task is not requiring the identification of a louder or softer sound. Our key hypothesis is that the anchor depicting a 'strong' clap will amplify the loudness of its concomitant sound more than the anchor of the 'weak' clap. As a consequence, we predict the 'strong' clap will generate higher PSEs than the 'weak' clap condition, see Figure 5.1A. In this case, the loudness of the visible anchor is a synthesis of both the visual expectation and auditory input.

It is worth noting that there are three additional factors that may influence responses within the current paradigm. While we are not directly interested in these factors, their consideration will aid the interpretation of our data. First, it is possible that participants display a response bias in which there is a preference for choosing either the visible or non-visible anchor; for example, see Figure 5.1B. Second, it is possible that participants display a response bias in which there is a preference for choosing either the higher or lower dB interval; for example, see Figure 5.1C. Finally, it has been suggested that when the onset of a forthcoming auditory stimulus is temporally predicted (i.e., by a video of hands moving towards collision), the experience of that stimulus is attenuated relative to when temporal cues are not provided (Hughes et al., 2013). In this case, the visible anchor would be attenuated relative to the non-visible anchor; for example, see Figure 5.1D. The scenario in which there is a response bias towards the lower dB interval and the scenario in which there is attenuation of both visible anchors cannot be disentangled as they predict the same pattern of results, however, resolving this is not an objective of the present study. Crucially, our experiment was designed such that the effects of these three potential response biases do not confound our key predicted perceptual effect (as shown in 5.1A).

5.3 Methods

5.3.1 Participants

A total of 32 participants were recruited from a pool of university students enrolled in an introductory psychology course at UNSW Sydney. 9 participants were excluded for failing to identify a requisite number of 'catch' trials (see Exclusions). Participants received course credit for their involvement and gave informed and written consent in accordance with the experiment protocols approved by the Human Research Ethics Advisory Panel in the School of Psychology, UNSW Sydney (#2968). All participants were naïve to the purposes of the experiment.

5.3.2 Apparatus

Auditory stimuli were presented via a pair of 'AudioFile' devices (Cambridge Research Systems, Kent, UK) and over-ear headphones (Beyerdynamic, Heilbronn, Germany; model DT990 Pro). The sound level produced by the headphones was determined using an artificial ear, microphone, and analyser (Brüel & Kjæer, Nærum, Denmark; models 4152, 4144, and 2250, respectively). All subsequently reported sound levels are in units of dB SPL as determined by this calibration method.

Visual stimuli were presented on a Display++ LCD monitor (Cambridge Research Systems, Kent, UK) with a spatial resolution of 1920 × 1080 pixels, temporal resolution of 120Hz, and mean luminance of 60 cd/m^2. The relationship between the video signal and monitor luminance was linear. Participants viewed the monitor in one of two darkened rooms from a distance of 54cm, for a visual angular subtense of 73.7° × 41.5°. The experiment was controlled using PsychoPy (Peirce, 2007, 2008).

5.3.3 Stimuli

Auditory claps were produced by convolving an anechoic recording of a clap with a room impulse response. The impulse response was obtained from the "Salford-BBC Spatially-sampled Binaural Room Impulse Responses database" (Satongar, Lam, & Pike, 2014) and characterised a frontally-positioned source in an enclosed room at a distance of 1m. The clap (obtained from `https://freesound.org/people/Anton/sounds/345`) was downsampled to the sampling rate of the impulse response (48kHz) prior to convolution, and the resulting waveform was again downsampled to the sampling rate of the presentation device (44.1kHz). Manipulations of clap level were produced by multiplications of this waveform.

Two potential videos were paired with the auditory stimuli, one in which the actor producing the clap was visible and one in which the actor was not visible. The visible claps were produced by recording videos of the first author (with visible hands, arms, and torso) producing a hand clap with either 'weak' or 'strong' levels of force. These were the same stimuli used in our manipulation check. The recordings were made at a spatial resolution of 1920 × 1080 pixels using a Sony Cybershot RX100 digital camera. Videos were converted to greyscale, resampled to 960 × 540 pixels (giving a viewing angle of 18.5° × 10.4° on the presentation monitor) and temporally cropped such that the sequence had a duration of 54 frames (900ms). The contact of the hands occurred at frame 27 (thus providing 433ms of anticipatory motion). The intensity distributions of the resulting sequences were then normalised by z-scoring across space and time and multiplying by 0.45. Example frames are shown in Figure 5.2. Videos in which the claps were not visible were generated using a static 2-dimensional grey oval to mask the video frame so that no visual motion was visible. At the frame at which the hands

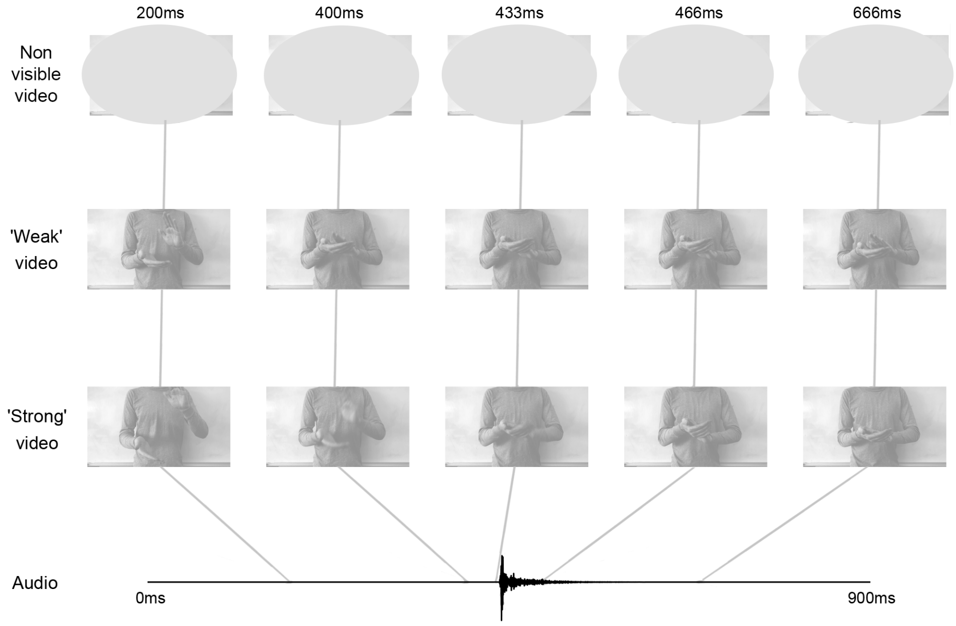

Figure 5.2. The time course of the three videos administered in this experiment.

collided (or equivalent frame for the non-visible clap), a parallel port signal was emitted which simultaneously triggered the onset of the auditory delivery of the clap sound.

Prior to conducting the main experiment we performed a manipulation check on our stimuli. This was to test whether our stimuli were overtly predictive of loudness. To do this we recruited seven naïve participants and asked them to complete three trials. Each trial consisted of the presentation of both a ‘strong’ and a ‘weak’ clap video (with the order counterbalanced across participants) without any accompanying auditory signals (i.e., silent claps). On each trial, participants were asked to judge which of the two videos: a) depicted a ‘louder’ clap; b) depicted a ‘stronger’ clap; and c) depicted a clap in which the hands moved faster. Each of the 7 participants judged the ‘strong’ clap video to be ‘louder’, ‘stronger’, and ‘faster’ than the ‘weak’ clap video.

5.3.4 Design

We used a two-way within-subjects design with factors of visually-conveyed clap effort (weak, strong) and anchor offset (-10dB, +10db). A single trial was comprised of 2 pairs of stimuli, delivering a total of four sounds. Each pair had one ‘probe’ clap sound and

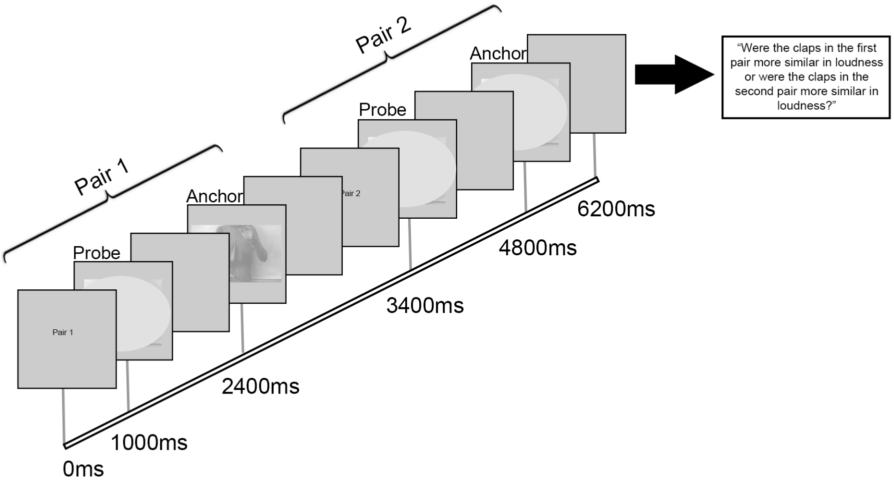
Pair 1
Pair 2
Probe
Anchor
Probe
Anchor
Pair 1
Pair 2
"Were the claps in the first pair more similar in loudness or were the claps in the second pair more similar in loudness?"
0ms
1000ms
2400ms
3400ms
4800ms
6200ms

Figure 5.3. The structure of a single trial in our two pair, four interval design.

one 'anchor' clap sound. The order within a pair was always [probe, anchor]. One of the anchors was paired with the non-visible clap and the other anchor had the visible clap. The order of which pair had the visible clap was randomised. If the visible actor produced a 'weak' clap, the auditory clap sound was always 65dB. If the visible actor produced a 'strong' clap, the auditory clap sound was always 75dB. If the anchor was paired with a non-visible clap the auditory clap sound was ±10dB the sound level of the visible anchor. This meant that the sound levels of the non-visible anchors when compared against the visible 'weak' clap anchor were either 55dB or 75dB. The sound levels of the non-visible anchors when compared against the 'strong' clap anchor were either 65dB or 85dB. The probe also consisted of a non-visible clap and was the same sound level across both pairs. For example, in a given trial of the ['weak' clap anchor, +10dB anchor] condition, the four intervals may consist of a Probe [68dB], visible anchor [65dB], probe [68dB] & non-visible anchor [75dB]. The sound level of the 'probe' shifts around between the dBs of the two anchors, according to a Psi staircase (Kontsevich & Tyler, 1999). This staircase assumed a logistic psychometric function relating the probe dB to the proportion of times it was perceived to be closer in loudness to the anchor with the higher dB level. After the two pairs of stimuli are presented, the task was to say which pair was more similar in loudness. Thus, the task was to determine in which pair the probe was more similar in sound level to its respective 'anchor'. An illustration of an example trial is shown in Figure 5.3.

5.3.5 Procedure

The experimental task was conducted in a single session which lasted approximately 50 minutes. The session consisted of a series of 8 runs, where each run contained 50 trials. The first four trials of each run were designated as practice trials and were discarded. The next 46 trials drew from the set of anchors (in dB): [55, 65], [75, 65], [65, 75], [85, 75], which were presented in a randomized order. Each trial began with 1s of blank screen. Next the onset of the first stimulus [‘probe’] would occur. After this, there was 500ms of blank screen and then the second stimulus [‘anchor’] would occur. Following the initial pair there was 1s of blank screen before the second pair was presented. Following conclusion of these video sequences, a prompt appeared “ *Were the claps in the first pair more similar in loudness or were the claps in the second pair more similar in loudness? Press the left arrow key if the first pair seemed more similar in loudness, or the right arrow key if the second pair seemed more similar in loudness.*” Participants were required to press a key to make their loudness judgement. After pressing the key the next trial began.

Statistical approach

Overall there were four conditions, that were differentiated by the anchors present in each pair: [‘weak’ clap, -10dB], [‘weak’ clap, +10dB], [‘strong’ clap, -10dB], [‘strong’ clap, +10dB]. From each subject we obtained 50 trials for each of the 4 conditions. The key information extracted from each trial was the intensity of the probe (dB) and whether the probe was perceived as being closer in loudness (or not) to the pair that contained the higher intensity anchor. We were interested in the PSE parameter extracted for each of the four conditions. The PSE parameter captures the intensity change to the probe that was required for it to be perceived as equally similar in loudness with each of the two anchors. The intensity change of the probe was relative to the anchor with the higher intensity. For example, in the condition in which we have anchors of 55 dB and 65 dB, a PSE of -5 indicates that the probe felt most similar in loudness with either anchor when it was 5dB below the higher intensity (65 dB) interval. In this case, the PSE reflects the identification of the point that is 5dB from both the high and low anchors. To extract PSEs for each participant, the trial-wise responses were modelled as Bernoulli events, in which the probability of indicating that the probe was more similar in loudness to the higher intensity anchor (dB) was given by a cumulative normal psychometric function. This function included the parameters of

PSE, spread, and lapse rate. The PSE and spread of the psychometric functions produced separate estimates for each participant and condition. The PSE for a given participant and condition was estimated with a prior, which was defined with a normal distribution with a mu of -5 and SD of 2.5. The spread for a given participant and condition was estimated with a prior given by a log normal distribution with a mu of $\log 0.75$ and SD of 1. The lapse rate parameter sets the upper and lower asymptotes of the psychometric function, and was estimated using catch trial errors. It was set with a prior that had a beta distribution with an $\alpha = 3$ and a $\beta = 20$.

The primary aim of this study was to identify whether PSEs differed as a function of viewing the 'strong' or 'weak' clap. To investigate the effect of viewing the clap on mean PSEs, we conducted a (2) x (2) repeated measures ANOVA with factors of visually cued source power and anchor offset. When a violation to the assumption of sphericity occurred, a Greenhouse-Geisser correction was applied. Statistical significance was assessed against a Type 1 error rate of 0.05. It is noteworthy that a visual anchor that is biased to sound 1 dB less intense should result in a $\frac{1}{2}$ dB PSE change. This is because to acquire the point that is equidistant from either anchor we need to halve the difference between both anchors. For example, say the 'weak' clap visual biases its 65dB sound so that it is perceived as being 1dB less loud, the point that is equidistant between its perceived loudness and the 55 dB anchor is now 59.5dB. As a result, to gauge the loudness bias from the PSEs of the 'strong' and 'weak' clap conditions, we must double the mean difference to estimate the bias between either anchor (in dB).

Finally, to test whether participants demonstrated a response bias towards the higher-dB interval or lower-dB interval, we performed a one-sample *t*-test. To do this we pooled the PSEs of all conditions and compared whether our composite variable deviated from the null hypothesis, which is that the PSE is equidistant between the two anchors.

5.3.6 Exclusions

Each run included 4 randomly interspersed 'catch' trials, which allowed the identification of participants who were not reliably attending to the clap events. In these trials both anchors were paired with non-visible videos. The probe sound was always presented at exactly the same sound level as one of the anchors. If participants picked the pair in which the probe was not the same sound level as the anchor, this was considered to be a catch trial error. We excluded participants for exceeding a catch trial error rate of 15% (7 participants). Furthermore, we also introduced ‘catch’ questions, which appeared on approximately 50% of trials. In these trials, after making a loudness judgement, participants were also asked a follow up question "Also, was the video with the visible person in the first pair or the second pair? Press the left arrow key if you think the visible person was in the first pair of sounds, or the right arrow key if you think the visible person was in the second pair of sounds". We excluded participants for exceeding a catch question error rate of 10% (2 participants). After exclusions 23 participants remained. For the remaining participants, these catch trials were not included in the behavioral analysis.

5.4 Results

Figure 5.4 shows on average how many dB the probe had to be reduced relative to the higher intensity anchor at PSE. This is represented for each anchor offset [-10 dB, 10 dB] and visual presentation condition [‘weak’ clap, ‘strong’ clap]. PSEs that deviate from a comparison change value of 5 reflect the deviation (in dB) from the point that is equidistant between either anchor.

To investigate whether the PSEs differed based on the visually cued power of the sound source we employed a (2)x(2) repeated measures ANOVA, with factors of visual presentation condition (‘weak clap’, ‘strong clap’) and anchor offset (-10dB, +10dB). We found a significant effect of visual-presentation condition $(F_{1,22} = 12.463, p = 0.002, \eta_p^2 = 0.362)$. This is our key comparison, and it indicated that when pooling across anchor offset conditions, the central tendency of the PSEs were closer to the highest interval $(MD = 0.84, SE = 0.239)$ when viewing the strong clap $(M = -5.11, SE = 0.21)$ relative to when viewing the weak clap $(M - 4.27, SE = 0.16)$. This effect was consistent with our hypothesis that ‘strong’ clap PSEs would be higher than ‘weak’ clap PSEs.

Following this, we found a significant interaction effect between the anchor offset and visual-presentation conditions $(F_{1,22} = 9.363, p = 0.006, \eta_p^2 = 0.299)$. This indicated that the effect of visual anchor on loudness judgements differed based on the offset anchor present in the other pair. We also found a significant main effect of anchor offset $(F_{1,22} = 11.622, p = 0.003, \eta_p^2 = 0.346)$. This indicated that when pooling across visual presentation conditions, the central tendency of the PSEs were closer to the highest interval $(MD = 0.87, SE = 0.256)$ in the +10dB anchor condition $(M = -5.13, SE = 0.22)$ relative to the -10dB anchor condition

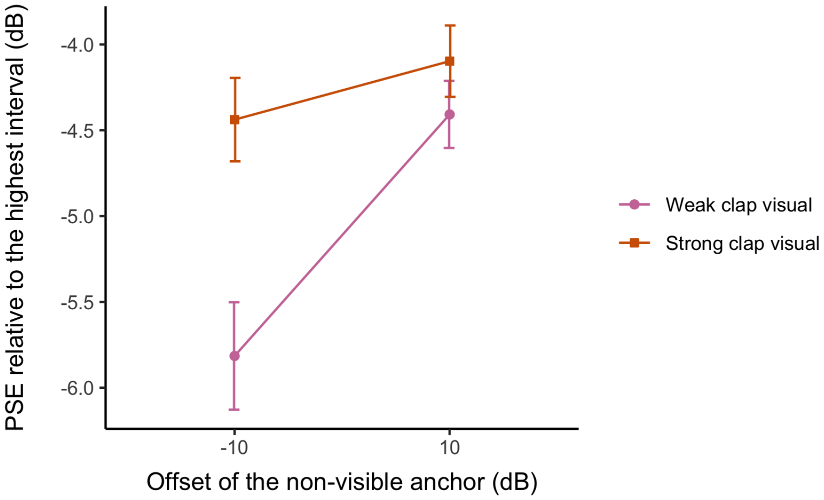

Figure 5.4. Mean PSEs. Each coloured symbol represents the visual-presentation condition of the visible anchor. The x-axis captures dB offset of the non-visible anchor. Error bars indicated 1 unit of standard error. The y-axis represents the number of dB the probe had to be artificially reduced relative to the higher dB interval, to appear most similar in loudness with either anchor.

$(M = -4.25, SE = 0.168)$. We unpacked this interaction using pairwise comparisons. First, we examined the difference between the ‘strong’ and ‘weak’ clap conditions in trials coupled with a -10dB anchor offset. We found a significant difference between the visual presentation conditions when the non-visible anchor was -10dB $(F_{1,22} = 15.17, p = 0.001)$. Here PSEs were closer to the highest interval in trials that contained a ‘strong’ clap video compared to trials that contained a ‘weak’ clap video $(MD = 1.38, SE = 0.35)$. We did not find a significant difference $(F_{1,22} = 1.93, p = 0.179)$ between visual presentation conditions when the offset anchor was $+10$dB $(MD = 0.31, SE = 0.22)$. Overall, these means indicated that the upward shift in PSEs from the -10dB anchor offset to the $+10$dB anchor offset, was less for the ‘strong’ visual condition than it was for the ‘weak’ visual condition.

Lastly, we computed a new PSE variable that was the average of all four conditions. On it we performed a one-sample t-test to gauge whether participants’ overall responses shifted from the point equidistant between either anchor (-5 dB). Overall, we find that PSEs were slightly above the point equidistant between either anchor $(MD = 0.31, SE = 0.15)$, and that this difference was statistically significant $(t_{22} = 2.13, p = 0.045)$.

5.5 Discussion

The key finding of this study was that when a visible anchor contained a video depicting an actor producing a 'strong clap', the PSE of the probe shifted upwards relative to when it contained a 'weak' clap. An upwards shift in PSE indicates that an anchor has increased in perceived intensity; this is because the point which is equidistant from either anchor is now higher. These results are consistent with the hypothesis that the 'strong' clap anchor would shift the perception of auditory input to feel relatively louder, whilst a 'weak' clap anchor would shift the perception of auditory input to feel relatively softer. Specifically, we found that the mean PSE of the 'strong' clap condition was 0.84 dB higher than that of the 'weak' clap condition. In this experiment, we assume a $\frac{1}{2}$ unit shift in loudness at PSE, is the consequence of a 1 unit shift in loudness at an anchor; consequently, we assume the 'strong' anchor was shifted upwards by on average 1.7 dB relative to the 'weak' clap. These findings suggest that visual information cuing the power of a sound source is integrated into the perception of loudness. This effect is somewhat analogous to the McGurk effect, with the key difference being that visual stream information was shown to effect perceived signal *intensity* rather than perceived signal *identity*.

It is also of interest to consider whether response biases or visually-induced sensory attenuation may have influenced our visual presentation effect. If participants displayed a response bias towards either the lower or higher dB interval, we would expect the PSEs to shift respectively below or above the point that is equidistant between the two anchors. Similarly, if participants experienced sensory attenuation to the visible anchor, we would expect the PSEs to shift below the point that is equidistant between the anchors. When combining all conditions together we found that PSEs deviated from the point that was equidistant between the anchors (i.e., -5 dB) by 0.31 dB. This indicated that there may have been a slight response bias towards choosing the pair with the lower dB interval. Further, if participants demonstrated a response bias towards picking the non-visible anchor, we would expect the -10 dB non-visible anchor to generate lower PSEs than the +10 dB non-visible anchor. This is what we found: the -10 dB non-visible anchor generated PSEs that were significantly lower than the +10 dB non-visible anchor. Critically, these response biases cannot account for the fact that the visible 'strong' clap generated significantly higher PSEs than the visible 'weak' clap.

To our knowledge, this is the first study to provide evidence that the effect of visual information on loudness judgements is driven by a perceptual response and not a post-perceptual decision. This finding agrees with the literature that has suggested that perceptual systems serve not to measure the veridical features of sensory input, but instead to functionally estimate events and objects 'out in the world' (e.g., Alais & Burr, 2004; M. O. Ernst & Bülthoff, 2004; Knill & Richards, 1996; Körding et al., 2007; Schutz & Kubovy, 2009; Shams & Beierholm, 2010). Visual cues carrying information about a sound source's power may increase the reliability with which an auditory event can be represented. For example, a visual signal of a 'strong' clap has a high probability of producing a high intensity sound. Exploiting this association by integrating visual information about the intensity of a sound producing event into the auditory percept may enhance the accuracy with which auditory input is transformed into a representation of that event. Whilst this paper represents one of the few examples of visually-based expectations influencing subjective loudness, it is consistent with similar accounts that have demonstrated that visual cues can influence auditory percepts with respect to representing the identity, location and timing of auditory events (Alais & Burr, 2004; Fendrich & Corballis, 2001; McGurk & MacDonald, 1976; Shams et al., 2005; Thurlow & Jack, 1973).

Our results do not disentangle whether the perceptual system aims to represent the intensity of an auditory event proximally at-the-ear or distally at-its-source. Because the distance of the object remained fixed, both proximal intensity and distal source power would increase or decrease together based on the visual power cues we provided (i.e., a strong clap would be of higher intensity both at-the-ear and at-its-source). Without independently varying the distance and power of the sound source, it is impossible to determine which location the perceptual system is estimating. If it represents the proximal intensity of input at-the-ear, we would expect a source with constant power that is visually cued to be closer to be perceived as increasing in loudness. Alternatively, if the perceptual system aims to estimate distal sound-source power, we are likely to discount changes in the physical signal at-the-ear due to variations in the distance of the sound source. Here we would perceive stable loudness independent of location: this capacity has been labelled loudness constancy (Zahorik & Wightman, 2001). In the context of the current experiment, an example of loudness constancy would be the perceived loudness of the 'strong' clap remaining constant regardless of its visually cued distance from the viewer. Previous studies have only modulated visual cues to distance without providing visual cues to source power. Mershon et al. (1981); Mohrmann (1939); Shigenaga (1965), & von Fieandt (1951) all demonstrated loudness constancy. However, in three of these four studies the auditory cue of reverberation may have interacted with visual distance cues and partially contributed to loudness constancy (Mohrmann, 1939; Shigenaga, 1965; von Fieandt, 1951). When controlling for auditory cues more rigorously, recent findings have failed to find that visual distance cues support loudness constancy (Altmann et al., 2012; Berthomieu et al., 2019, Chapter 3, 4). The majority of recent evidence suggests that visual cues to source distance do not contribute to loudness constancy in the absence of source power cues. However, it is unclear whether visual cues to source power are a necessary prerequisite for visual cues to source distance to affect the subjective experience of loudness.

A limitation of the current experiment was that the visible 'weak' clap was always paired with a 65 dB sound and the visible 'strong' clap was always paired with a 75 dB sound. These sound levels were fixed to avoid incongruencies between the visually predicted auditory intensity and the auditory input. Audiovisual incongruencies can generate prediction errors which have been suggested to alter cortical processing and reduce sensory attenuation (e.g., Friston, 2009; Hughes et al., 2013). To minimise the influence of prediction errors, the visual cue that predicted a louder sound was paired with a higher intensity sound than the visual cue that predicted the softer sound. Nonetheless, it would be informative for future studies to pair a variety of source power cues provided visually with *identical* auditory input. This would allow for a direct comparison of the perceptual effects introduced by visually-created intensity expectations. Furthermore, it would be useful to pair source power cues with a wider spectrum of sound levels and presentation contexts. This would help determine whether the effect of visually provided intensity expectations on loudness is an absolute or relative effect (i.e., relative to (a) the sound level the visual signal is paired with, (b) previously experienced audiovisual pairings, and/or (c) the environment the audiovisual signal is received within).

To date, the only other study that has investigated causal sound power cues and loudness also involved the use of stimuli that depicted goal directed actions, namely speech and clapping (Rosenblum & Fowler, 1991). In the present study we utilised clap stimuli, and so we cannot rule out the possibility that this perceptual effect is driven by an action-specific multisensory mechanism. It has been suggested that specific perceptual mechanisms may be allocated to mirroring the actions of others (Grèzes, Costes, & Decety, 1999; Iacoboni et al., 1999; Konorski, 1967; Koski et al., 2002). Because the only studies that have investigated the influence of visual signals on loudness have employed stimuli that depict gestural movements, it is possible these actions are the only types of visual cues that influence loudness. Alternatively, it is possible loudness percepts are determined by a broader generative model in which any causal information about an auditory event is integrated. In this case, it would be relevant for future experiments to isolate the primitive factors in visual cues that drive intensity expectations. Fassnidge and Freeman (2018) found that videos with moving patterns that had high 'motion energy' were more likely to induce an illusory auditory experience. In this experiment 'motion energy' was quantified with a computational model that captured the degree to which patterns of luminance changed over space and time. Videos with high 'motion energy' contain high amounts of flickering or movement. It could be that motion energy mimics the physical energy of an object generating sounds, and this cue informs our perception of intensity. More broadly, any visual cue that indicates the transfer of kinetic energy into sound holds the 'potential' to predict a sound's intensity. For example, we know that the velocity and mass of objects in a collision are two factors that determine the amount of energy in the resultant sound (Rienstra & Hirschberg, 2004). Accordingly, visual information about these two factors is predictive of the intensity of a forthcoming sound, and thus may influence subjective loudness. In addition, it is of interest to resolve whether intensity expectations are malleable enough to influence perceived loudness after a short period of learning. To do this, future studies need to determine whether ecologically irrelevant cues such as static geometric shapes can be associated with specific auditory intensities, and in turn, whether this can alter loudness percepts.

Current conceptualisations of subjective loudness do not adequately account for the influence of higher order information (M. Epstein & Florentine, 2009, 2012; Moore, 2014). We have demonstrated that causal information provided by 'seeing volume' influences the perception of loudness. The findings of this study may have interesting implications for improving the application of hearing models out in the 'real world', where visual cues are abundantly present. If it is possible to introduce a visual parameter into loudness models, this may increase the accuracy in which predicted loudness functions map onto the audiovisual experience of loudness. This finding may have particular relevance for aging populations. It is well known that aging populations often experience hearing loss (Fozard, 1990) and further, that aging populations demonstrate enhanced multisensory integration (Laurienti, Burdette, Maldjian, & Wallace, 2006). For aging populations the influence of visual stream information on perceived auditory intensity may be exaggerated. Consequently, the potential for loudness models to represent aging populations may be improved by further consideration of how visual stream information interacts with the perceived loudness of auditory input.

In conclusion, this is (to the best of our knowledge) the first study to demonstrate that visually created intensity expectations can regulate loudness judgements. This effect remained evident even after accounting for the possibility of post-perceptual response biases. In summary, we have demonstrated that when we *see* an action signaling a louder sound, our perception of loudness is affected in accordance with the visually-created expectation.

6 Visual Cues to Source Power & Loudness (EEG experiment)

Seeing the intensity of a sound-producing event modulates the amplitude of the initial auditory evoked response

Author contributions:
Conceptualisation: SL, DJM, TJW. Stimuli: SL. Methodology: SL, DJM, TJW. Programming: DJM. Data collection: SL. Data analysis and presentation: SL, DJM, TJW. Writing - original draft: SL. Writing – review and editing: SL, TJW, DJM. Supervision: TJW, DJM.

Published as:
Libesman, S., Mannion, D. J., & Whitford, T. J. (2020). Seeing the Intensity of a Sound-producing Event Modulates the Amplitude of the Initial Auditory Evoked Response. *Journal of Cognitive Neuroscience*, *32*(3), 426-434.

Preamble:

In the previous Chapter we demonstrated that visual information about the power of a sound source influenced the perceived loudness of the sound in a manner consistent with the visually created expectation. In the present Chapter we use electroencephalography (EEG) to explore whether these same visual cues to source power also influence the neurophysiological response to auditory input. This Chapter has been published as a Research article in the Journal of Cognitive Neuroscience. It is noteworthy that this article was published before the results of Chapter 5 had been finalised. As a consequence of this, there is no reference to the previous Chapter.

6.1 Abstract

An auditory event is often accompanied by characteristic visual information. For example, the sound level produced by a vigorous handclap may be related to the speed of hands as they move toward collision. Here, we tested the hypothesis that visual information about power of a sound source is capable of altering the subsequent neurophysiological response to auditory stimulation. To do this we used electroencephalography (EEG) to measure the response of the human brain ($n = 28$) to the audiovisual delivery of handclaps. Depictions of a weak handclap were accompanied by auditory handclaps at low (65 dB) and intermediate (72.5 dB) sound levels, whereas depictions of a vigorous handclap were accompanied by auditory handclaps at intermediate (72.5 dB) and high (80 dB) sound levels. The dependent variable was the amplitude of the initial negative component (N1) of the auditory evoked potential. We find that identical clap sounds (intermediate level; 72.5 dB) elicited significantly lower N1 amplitudes when paired with a video of a weak clap, compared to when paired with a video of a vigorous clap. These results demonstrate that intensity predictions can affect the neural responses to auditory stimulation at very early stages ($<$ 100ms) in sensory processing. Furthermore, the established sound–level dependence of auditory N1 amplitude suggests that such effects may serve the functional role of altering auditory responses in accordance with visual inferences. Thus, this study provides evidence that the neurally evoked response to an auditory event results from the combination of a person's beliefs with incoming auditory input.

6.2 Introduction

Hermann von Helmholtz, a pioneer in audition research, once described the experience of watching a plucked guitar string by writing *"When we strike a string, its vibrations are at first sufficiently large for us to see them, and its corresponding tone is loudest. The visible vibrations become smaller and smaller, and at the same time the loudness diminishes."* (Helmholtz, 1877, p. 10). Here, Helmholtz presented a clear demonstration that visual cues have the capacity to provide information about auditory intensity. Everywhere around us, we see the relationship between the characteristics of a physical action and the nature of the auditory consequence. For example, when someone is whispering they will narrowly open their lips, whereas a shout will involve a wide open mouth. Similarly, a soft clap will involve a slow, weak movement, whereas a loud clap will often involve a fast, powerful movement. Although it is clear that visual cues can provide information about the expected intensity of auditory events, it is unclear whether visual cues can modulate the responsiveness of the primary auditory cortex to auditory events. Addressing this question is the aim of this study.

Using ERPs acquired through EEG, we investigated how predictions regarding the expected intensity of sounds (specifically, handclaps) affected the evoked neurophysiological responses (specifically, the amplitude of the N1 component of the auditory-evoked potential). The N1 component is the negative peak that appears approximately 100 msec following the onset of a brief auditory stimulus. It has been found to have dominant origins in the auditory cortex (Pantev et al., 1995). An important feature of the N1 is that its amplitude is known to be intensity dependent; sounds of higher intensity elicit larger N1 amplitudes than sounds of lower intensity (Brocke et al., 2000; Dierks et al., 1999; Hegerl et al., 1994; Mulert et al., 2005; Rapin, Schimmel, Tourk, Krasnegor, & Pollak, 1966).

This study investigated the interaction between visual information and auditory intensity predictions by pairing auditory signals (sounds of hands clapping) with videos of a person performing handclaps. Two different videos were presented: a video of an actor producing a weak handclap and a video of an actor producing a vigorous handclap. These visuals were suggestive of generating either low or high auditory intensity, respectively. Handclap sounds of varying intensities were paired with these videos. Our hypothesis was that the degree of activation of the primary auditory cortex in response to an auditory event is the combination of (1) the intensity of the auditory signal at the ear and (2) the concomitant visual stream information, depicting the generation of the auditory signal at a particular intensity. Specifically, we propose that the N1 amplitude to a handclap at a given intensity would be greater when paired with a video depicting a vigorous handclap than when paired with a video depicting a weak handclap. Here, the amplitude of the N1 component generated from the received signal shifts toward the response that would be generated from the expected (visual) signal.

6.3 Methods

6.3.1 Participants

A total of 36 participants were recruited from a pool of students enrolled in an introductory psychology course at UNSW Sydney. Four participants did not produce data because of technical problems with data acquisition, and four participants were excluded for failing to identify (predefined) requisite number of "catch" trials (see Procedure section). Of the remaining 28 participants, 13 were women, and 25 were right-handed. The majority of participants (20 of 28) were between 17 and 20 years old. Participants received course credit for their involvement and gave informed and written consent in accordance with the experiment protocols approved by the Human Research Ethics Advisory Panel in the School of Psychology, UNSW Sydney (#2968). All participants were naïve to the purposes of the experiment.

6.3.2 Apparatus

Auditory stimuli were presented via an AudioFile device (Cambridge Research Systems) and over-ear headphones (Beyerdynamic; Model DT990 Pro). The sound level produced by the headphones was determined using an artificial ear, microphone, and analyzer (Brüel & Kjæer, Nærum, Denmark; models 4152, 4144, and 2250, respectively). All subsequently reported sound levels are in units of dB SPL as determined by this calibration method.

Visual stimuli were presented from a XLT2420T BenQ computer monitor (60 Hz, 1920 × 1080 resolution). Participants viewed the monitor in a well-lit room at a distance of 82 cm for a visual angle of 37.0° × 20.8°. The experiment was controlled using PsychoPy (Peirce, 2007, 2008).

The EEG was recorded using a BioSemi ActiveTwo system with 64 Ag–AgCl active electrodes placed per the extended 10–20 system, sampling at a rate of 2048 Hz. The Ag/Ag–Cl electrodes were connected to all 64 cap channels, with additional electrodes attached to the mastoids and nose and placed 1 cm from the outer canthi of both eyes and 1 cm under the left eye to monitor horizontal and vertical eye movements. Online referencing was to sensors located in the parietal region of the cap (Common Mode Sense active electrode, Driven Right Leg passive electrode) (The continuous EEG record for each participant is available at `https://doi.org/10.6084/m9.figshare.c.4286837.v1` ; Figshare). EEG data were processed using BrainVision Analyser (Version 2.1).

6.3.3 Stimuli

Auditory claps were produced by convolving an anechoic recording of a clap with a room impulse response. The impulse response was obtained from the Salford-BBC Spatially-sampled Binaural Room Impulse Responses database (Satongar et al., 2014) and characterized a frontally positioned source in an enclosed room at a distance of 1 m. The clap (obtained from `https://freesound.org/people/Anton/sounds/345`) was downsampled to the sampling rate of the impulse response (48 kHz) before convolution, and the resulting waveform was again downsampled to the sampling rate of the presentation device (44.1 kHz). Manipulations of clap level were produced by multiplications of this waveform. The sounds for each of the three audible levels are available at `https://doi.org/10.6084/m9.figshare.c.4286837.v3`.

Visual depictions of claps were produced by recording videos of the first author (with visible hands, arms, and torso) producing a handclap with either "weak" or "strong" levels of force. Compared with the "weak" clap, the top hand in the "strong" clap moved from a higher position, traveled more rapidly to the bottom hand, and produced greater vibration on collision. The recordings were made at a spatial resolution of 1920 × 1080 pixels using a Sony Cybershot RX100 digital camera. Videos were converted to grayscale, resampled to 960 × 540 pixels (giving a viewing angle of 18.5° × 10.4° on the presentation monitor), and temporally cropped such that the sequence had a duration of 54 frames (900 msec). The contact of the hands occurred at frame 27 (thus providing 433 msec of anticipatory motion). The pixel intensities in the resulting sequences were then each normalized to have a mean of 0.0 and a standard deviation of 0.45 before being clipped to be within a $[-1, +1]$ interval. This was performed to enforce that the two videos were similar in their overall distributions of pixel intensity. Example frames are shown in Figure 6.1, and videos are available at `https://doi.org/10.6084/m9.figshare.c.4286837.v3`.

6.3.4 Design

We used a within-subject design with a total of six cells, with three conditions each for "weak" and "strong" visual depictions. The three conditions for the "weak" clap visual depiction involved the clap sounds being delivered at three levels (silent, 65 dB, and 72.5 dB), and the three conditions for the "strong" clap visual depiction involved the clap sounds being delivered at three levels (silent, 72.5 dB, and 80 dB). We chose this design over a fully crossed alternative to retain the ecological audiovisual association;

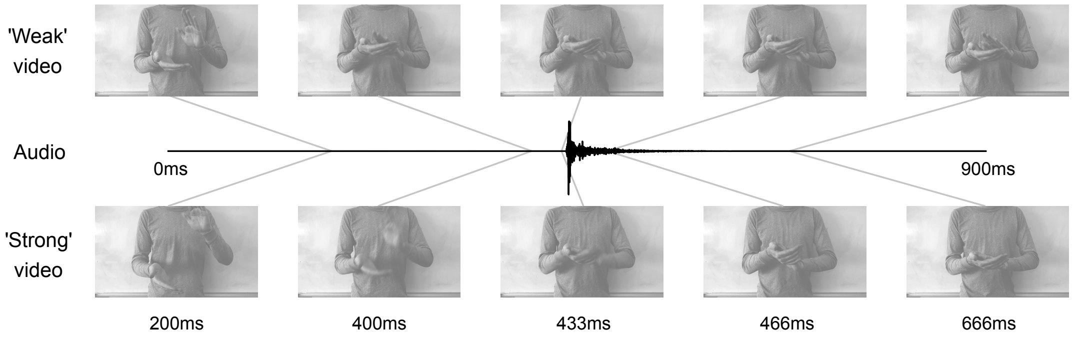

Figure 6.1. The time course of the two videos administered in this experiment.

over the course of the experiment, a “weak” visual clap was more likely to be paired with an auditory clap of a lower sound level than a “strong” visual clap.

6.3.5 Procedure

The experimental task was conducted in a single session, which lasted approximately one and a half hours. After being fitted with an EEG cap and electrodes, participants had their EEG continuously recorded as they completed the experimental protocol. This protocol was approximately 50 min in duration and consisted of 12 experimental runs. Each run contained 39 trials. Each trial began with the onset of the visual depiction of a clap. At the frame at which the hands collided, a parallel port signal was emitted, which simultaneously triggered the onset of auditory delivery and marked the EEG record. Following conclusion of the video sequence, there was an interval of random duration (uniformly sampled from between 3 and 5 sec) before the trial ended during which the screen was uniformly gray.

The first eight runs of the experiment consisted of trials of each of the four audiovisual conditions (weak-video, 65 dB sound; weak-video, 72.5 dB sound; strong-video, 72.5 dB sound; strong-video, 80 dB sound). The trials were presented in random order. Over the course of these eight runs, 72 trials of each of the four audiovisual conditions were presented.

The final four runs of the experiment consisted of trials of the two video-only conditions (weak-video, silent; strong-video, silent). Over the course of these four runs, 72 trials of each of the two video-only conditions were presented.

Each run also included three randomly interspersed “catch” trials, intended to allow the identification of participants who were not reliably attending to the clap events. These trials were identical to a randomly selected condition, with the exception that a small green cross was briefly presented (67 msec) following the collision of the hands in the video. Participants were required to demonstrate they had detected the green cross by means of a keypress. These catch trials were not included in the EEG analysis.

6.3.6 Analysis

For each participant, the EEG data were first rereferenced offline to an average of the mastoid electrodes. The continuous EEG was then band-pass filtered from 0.1 to 30 Hz (eighth-order zero-phase Butterworth IIR). ERPs were then extracted, where each ERP was 600 msec in duration and encompassed the 200 msec prior and 400 msec following the onset of the initiation of the clap sound. These ERPs were then baseline-corrected by subtracting the average of the 200 msec preonset signal, separately for each condition. Eyeblink artifacts were then corrected with the method of Gratton, Coles, and Donchin (1983) using an algorithm that involves the subtraction of ocular artifacts by creating a propagation factor that captures the relationship between ocular activity monitored with an electrooculogram (created with external electrodes) and EEG traces at each electrode. Electrodes PO7, P8, Oz, POz, P6, and O2 were leading to more than 75% of trials to be rejected for three participants, and so topographic interpolation was conducted on these electrodes for these participants. For each electrode, ERPs containing a voltage change between adjacent 200-msec intervals in excess of $200\mu V$ or a maximum gradient greater than $50\mu V$ were then excluded. On average, the “weak” visual (65 dB) retained 88% of trials (SD = 20%), the “weak” visual (72.5 dB) retained 87% of trials (SD = 21%), the “strong” visual (72.5 dB) retained 89% of trials (SD = 18%), and the “strong” visual (80 dB) retained 85% of trials (SD = 24%). The remaining ERPs were then averaged across trials, separately for each condition. Finally, the ERPs from the silent conditions were subtracted from the audiovisual conditions with the corresponding visual clap (“weak” or “strong”); this was designed to reduce the influence of any purely visual contributions to the ERPs and is a typical procedure in multisensory studies (Guthrie & Buchwald, 1991; Stekelenburg & Vroomen, 2007). The resulting ERPs are used in all subsequent analyses and are shown (averaged across participants) in Figure 6.2A.

The dependent variable was the amplitude of the N1 component of the auditory

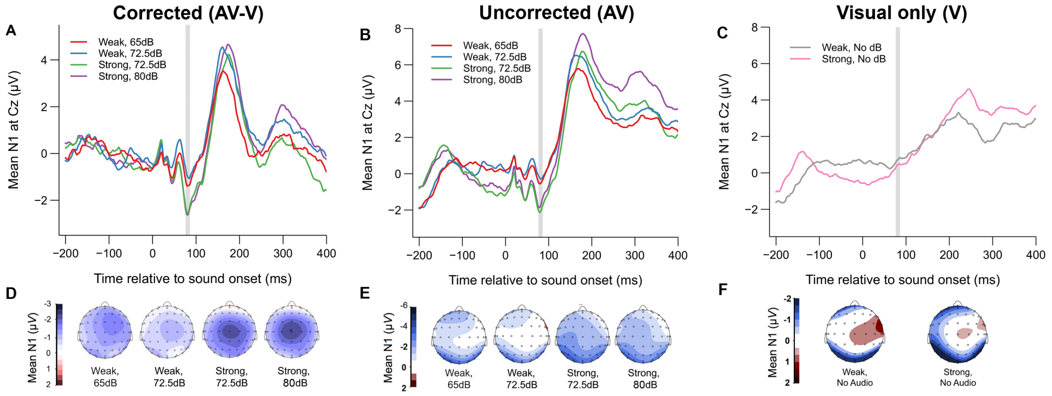

Figure 6.2. Differences between the AV–V, AV, and V–only waveforms in N1 amplitude. **(A)** Grand-averaged waveforms for the corrected (AV–V) condition. **(B)** Grand-averaged waveforms for the uncorrected (AV) condition. **(C)** Grand-averaged waveforms for the V–only condition. The 0 msec point on the x axis represents the onset of the auditory stimulus (which was silent in the V–only condition). All waveforms were recorded from electrode Cz, baselined to -200 to 0 msec prestimulus. **(D)**, **(E)**, and **(F)** represent the scalp topographies for the AV–V, AV, and V–only conditions, respectively. The time window used for the scalp topographies and statistical analysis was 76–86 msec.

evoked potential, which is typically elicited by binaural auditory stimulation and has a central topography that is maximal around Cz (Luck, 2005). These characteristics were observed in the current study, as shown in Figure 6.2D. Hence, all analyses were conducted on electrode Cz (as is common practice; for example: Oestreich et al., 2015; Vroomen & Stekelenburg, 2010). To determine the time window in which to evaluate the N1, we averaged the ERPs across all conditions and participants to produce a "collapsed localizer" waveform (Luck & Gaspelin, 2017). The time at which such a waveform was at its minimum (81 msec) was used to set an N1 evaluation window of 76–86 msec, common across all participants and conditions. The average voltage in this time window was then extracted from each participant and condition and was used as the measurement of the N1 amplitude. Differences between the N1 amplitude across conditions were evaluated using a paired-sample *t*-test.

6.4 Results

The aim of this study was to determine whether visual information about the intensity of a sound-generating event influences the neural processing of an associated auditory signal. We paired visual depictions of handclaps with weak or strong force with

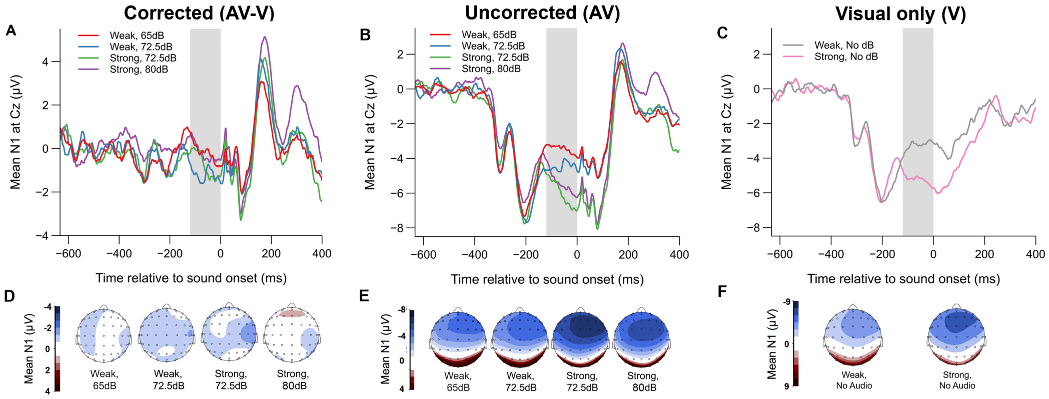

Figure 6.3. Extending our time window to 1033ms and setting the baseline to 200ms before the onset of the visual, we explored the anticipatory waveforms. **(A)** The grand average for each corrected AV–V condition. **(B)** The grand average for each uncorrected AV condition. **(C)** The grand average for each V–only condition. Across all three panels the waveforms are based on electrode Cz. The 0-msec point on the x-axis represents the onset of the auditory stimulus. The baseline was taken from -633ms–433ms when the visual stimulus appeared. The grey area indicates the time window of the pre-stimulus activation -120-0ms. The grand average scalp topography of the corrected waveforms **(D)**, the uncorrected waveforms **(E)** and the visual–only waveforms **(F)** for each condition. The time window for each scalp topography was -120-0ms

auditory claps of different sound levels. The key comparison in determining the effect of visual information on auditory processing was between the visual depictions of "weak" and "strong" claps, when paired with the same intensity clap sound (72.5 dB).

As shown in Figure 6.4, the mean N1 amplitude to 72.5-dB clap audio was larger when paired with the visually strong clap ($M = 2.320$) than when paired with the visually weak clap ($M = 1.107$). This difference ($M = 1.213$, $SEM = 0.585$) was statistically significant (paired sample, $t(27) = 2.073$, $p = .048$, $d = 0.392$), supporting the hypothesis that visual information about the intensity of an auditory event affects the amplitude of the auditory evoked potential.

The N1 amplitudes evoked by the claps at 72.5 dB were similar to the N1 amplitudes evoked by claps at 65 and 80 dB (which were respectively paired with the weak and strong clap visual), as shown in Figure 6.4A. The mean N1 amplitudes for the weak visual claps were comparable for the 65 dB ($M = 1.429$) and 72.5 dB ($M = 1.107$) auditory intensities, with this difference ($M = 0.322$, $SEM = 0.320$) not being statistically significant, $t(27) = 1.005$, $p = .324$, $d = 0.190$. Similarly, the mean N1 amplitudes for the strong visual claps were comparable for the 72.5 dB ($M = 2.320$)

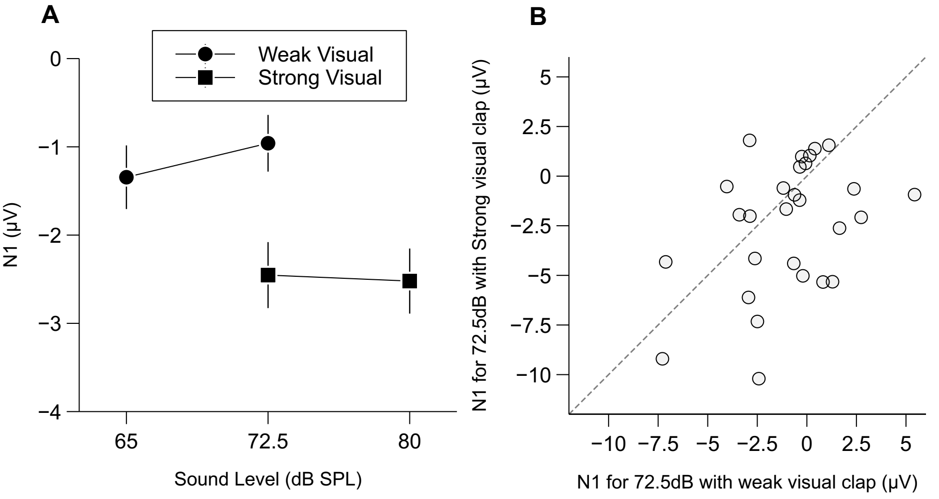

Figure 6.4. **(A)** The mean amplitude of participants N1 as a function of sound level, split between the weak-video and strong-video conditions. The SEM bars have been corrected to reflect error variance of a within subjects design, as is recommended (Cousineau, 2005). **(B)** Scatter plot of the mean N1 amplitude for 72.5 dB as a function of the weak and strong videos.

and 80 dB ($M = 2.346$) auditory intensities, with this difference ($M = 0.026$, $SEM = 0.435$) not being statistically significant, $t(27) = 0.059$, $p = .954$, $d = 0.011$.

Examining the conditions used to correct for any purely visual contribution to the ERPs, we find that the voltages evoked by the silent clap videos in the N1 time window were similar across the weak- and strong-video conditions. When sounds were absent, the mean N1 amplitudes were comparable for the weak clap ($M = 0.807$) and strong clap ($M = 0.277$) videos, with this difference ($M = 0.531$, $SEM = 0.526$) not being statistically significant, $t(27) = 1.010$, $p = .321$, $d = 0.191$.

We also conducted a post hoc exploratory analysis that examined the between-video differences (i.e., weak clap video vs. strong clap video) in anticipatory preclap activity across the corrected (AV–V), the uncorrected audiovisual, and the video-only blocks. As shown in Figure 6.3 B and C, the “strong clap” videos elicited a negative-going deflection ~120 msec preclap with a frontal topography that was not present in the “weak clap” videos. Comparing the uncorrected strong and weak clap videos (pooled across auditory stimuli), the between-video difference ($M = 1.78$, $SEM = 0.42$) was significant in the anticipatory period immediately preclap (paired sample, $t(27) = 4.24$, $p < .001$, $d = 0.567$). A similar result was observed when comparing the strong and weak clap videos in the V-only conditions: a negative-going deflection with a similar time course and topography was again observed in the strong clap condition (but not the weak clap condition), and the difference in prestimulus activity ($M = 2.13$, $SEM = 0.63$) was statistically significant, $t(27) = 3.371$, $p = .002$, $d = 0.691$. However, these prestimulus differences were not present in the "corrected" waveforms (which were the primary focus of our analysis), suggesting the subtraction of the visual-only condition was effective at eliminating prestimulus differences between the strong and weak clap conditions ($M = 0.07$, $SEM = 0.589$, $t(27) = 0.123$, $p = .902$, $d = 0.016$; see Figure 6.3A).

6.5 Discussion

The critical finding of this study was that expectations carried in the visual stream regarding the loudness of a physical event (a clap) significantly influenced the electrophysiological response to the associated auditory stimulus. The key comparison was between the "weak" and "strong" clap video clips when both were paired with the same clap sound level of 72.5 dB. The results revealed that the amplitude of the auditory N1 component was significantly larger when viewing the video depicting a "strong" clap compared with when viewing the video depicting a "weak" clap. This result was observed despite the fact that visual-evoked activity has been subtracted out of the waveforms and the auditory stimulus was physically identical in both conditions (i.e., a 72.5-dB clap). Consistent with additive models of multisensory interactions (Besle, Fort, Delpuech, & Giard, 2004; Besle, Fort, & Giard, 2004; Giard & Peronnet, 1999), this result indicates that the visual characteristics of an auditory event interact with auditory processing. Given the low latency of the N1 component (< 100 msec) and the fact the N1 is generated in the primary auditory cortex, this result suggests the multisensory interaction occurs at very early stages and in primary sensory regions. Furthermore, the direction of this effect favors our hypothesis; that the response of the auditory cortex generated from the received auditory signal (as indexed by the amplitude of the N1 component) is shifted toward the response that would be generated from the expected auditory signal.

Previously, it has been suggested that predictive cues are a mechanism for the reduction of uncertainty, thereby leading the brain to process a signal more efficiently and attenuating the N1 component (Besle, Fort, Delpuech, & Giard, 2004; Hughes et al., 2013; Schafer & Marcus, 1973). According to predictive coding theories, the brain is an active inference machine, and making accurate predictions aids the brain in dealing with uncertainty and inferring the most likely cause of a signal (Clark, 2013; Friston, 2005b). By this account, if a signal received in a bottom–up cortical area deviates from the top–down prediction, a prediction error will occur, which will increase the evoked neural response (Arnal & Giraud, 2012; SanMiguel, Saupe, & Schröger, 2013; van Laarhoven et al., 2017). However, in this study, it is unlikely that the observed differences between the "weak" and "strong" clap visuals when paired with a clap sound level of 72.5 dB were due to differences in uncertainty or prediction errors. The reason for this is twofold. First, temporal cues were held constant with both visual cues being edited to have the same amount of temporal predictability (i.e., both videos had 450 msec of anticipatory motion; note, however, that our assumption of equivalent temporal predictability is questionable, as we discuss further below). Second, the degree of sound-level uncertainty did not differ between different conditions: A weak visual cue had a 50% chance of producing either a 65- or 72.5-dB clap sound and a strong visual cue had a 50% chance of producing either a 72.5- or 80-dB clap sound. Factors beyond uncertainty that have also been found to modulate the N1 component are selective attention and variable ISIs (Näätänen & Picton, 1987). However, it is unlikely that differences in attention drove the effect found in this study as both clap visuals had equal task relevancy, and there was no significant difference in the evoked response in the N1 time window between the weak and strong visual control conditions (i.e., when the video was relayed silently). Furthermore, the ISIs did not differ between conditions. Thus, the differences in N1 amplitude we observed between the weak and strong visuals at 72.5 dB were unlikely to have been driven by extraneous factors (temporal predictability, volume uncertainty, ISI) but were instead driven by participants' visually based expectations about the sound influencing their neurophysiological response to the sound.

Hence, our study has provided a unique view on how predictive visual information may bias the neural response to an auditory event. Without employing differences in predictive accuracy, we have shown that the visual characteristics predicting the nature of an auditory event can bias the response of the auditory cortex to that event. Specifically, we have shown that auditory processing may be modulated by "seeing volume". There has been a great deal of work on quantifying the elements that influence loudness judgments. The intensity of an auditory signal is the most dominant determinant of loudness judgments (Stevens, 1955) yet frequency (Melara & Marks, 1990), timbre (Allen, 1971), duration (Miskolczy-Fodor, 1959) and reverberation (Zahorik & Wightman, 2001) have also been shown to influence perceived loudness. It is worth speculating that if the N1 amplitude modulation in this study is correlated with perceived intensity, our perception of loudness may be a synthesis of both auditory and visual information. Support for the notion that the amplitude of the N1 component reflects perceived loudness comes from the fact that the N1 component is extremely sensitive to the intensity of the received auditory signal (Brocke et al., 2000; Dierks et al., 1999; Hegerl et al., 1994; Mulert et al., 2005; Rapin et al., 1966). Furthermore, there is some behavioral evidence that the modulation of the auditory N1 is correlated with loudness estimates (Roussel, Hughes, & Waszak, 2014; A. Sato, 2008). If this is the case, "seeing volume" may be a new factor in which loudness may be modulated. This would indicate that the perception of auditory intensity is not only influenced by the direct properties of the auditory signal but also by the perceived relationship between the signal and visually causative events. This hypothesis is somewhat supported by the McGurk effect (McGurk & MacDonald, 1976) in which the brain fuses information from the auditory and visual streams to resolve similar but ambiguous stimuli. The difference in this study is that the merging of sensory stream information is not in relation to signal *identity* but rather to signal *intensity*.

We have argued that our finding that the 72.5-dB clap elicited a smaller N1 amplitude when paired with the "strong-video" (relative to the "weak-video") was due to the "strong-video" generating an expectation of a more intense sound. However, as flagged previously, an alternative "low-level" explanation is that the onset of the clap was more temporally predictable in the "weak-video" as the actors' hands were moving slower. The fact that increasing the temporal predictability of a sound has been shown to decrease the N1 amplitude that it elicits (Lange, 2011; Oestreich et al., 2015) is consistent with this explanation. Although disentangling the factors of temporal predictability and intensity expectations is challenging in the current paradigm without reducing the ecological validity of the stimuli, future studies could test the alternative hypothesis by, for example, replacing the clap videos with a slow-moving versus fast-moving bar. Such research would clarify the cause of the N1 differences observed in the current study.

To our knowledge, there have only been two other studies that have investigated the behavioral consequences of visually created intensity expectations on loudness judgments. Using both speech and nonspeech stimuli (clapping), Rosenblum and Fowler (1991) required people to rate the amount of perceived effort put into the generation of a sound and, second, to rate the loudness of a sound when paired with the same visual. When auditory stimuli were paired with a video of a sound emitter that was perceived to be putting in more effort, perceived loudness ratings also increased. Aylor and Marks

(1976) required participants to judge the relative loudness of narrowband noise transmitted through different barriers (row of hemlock trees, slat fence, acoustic tile barrier, or no barrier). Here, participants carried out two conditions, one blindfolded and one where the barrier obscuring the sound source was visible. In the blindfolded condition, there were no differences between loudness estimates for any of the barriers. In the condition in which participants had no blindfold, loudness ratings were relatively attenuated when the barriers did not completely visually obstruct the sound source (slat fence, no barrier). On the basis of this finding, it was suggested that when a sound source was occluded by a barrier, participants expected the barrier to diminish the loudness of the auditory stimulus, which in turn raised their loudness estimates. To confirm the relationship between the neural coding of auditory intensity and subjective loudness, future studies are needed to follow up whether psychophysical loudness judgments are influenced by a larger range of visual cues containing information about the intensity of auditory stimuli.

The finding of this study may have interesting implications for certain populations. First, there is evidence that aging populations undergo hearing loss (Fozard, 1990), aging populations (with and without hearing impairment) demonstrate altered N1 responses compared with younger participants (Tremblay, Piskosz, & Souza, 2003), and furthermore that aging populations demonstrate enhanced multisensory integration (Laurienti et al., 2006). It would be possible that the influence of visual cues on auditory processing may be exaggerated in an aging population who are more reliant on visual cues to support auditory information. Furthermore, the volume dependency of the N1 has been shown to be modulated in part by the serotonergic system, which has implicated its relevance in clinical disorders that are related to serotonin dysregulation (Hegerl, Gallinat, & Juckel, 2001). Low central serotonergic transmission is related to low loudness dependence of the auditory evoked signal, and high serotonergic transmission has been related to high loudness dependence of the auditory evoked signal (Hegerl et al., 2001). It would be interesting to investigate whether serotonin dysregulation extends beyond impacting the “loudness dependence” of the N1 component to impacting the modulation of the N1 based on expected loudness. Finally, people with schizophrenia have been specifically implicated as having deficits in auditory perception (Matthews et al., 2013) and an abnormal auditory N1 response when comparing the difference between audiovisual stimuli (with predictive visual cues) and auditory only stimuli (Stekelenburg, Maes, Van Gool, Sitskoorn, & Vroomen, 2013). As a consequence, people with schizophrenia who exhibit functional differences in the formation and adaption of top–down inferences may demonstrate an N1 response that is less influenced by intensity expectations.

A limitation of the current study is that we could not determine the boundaries of when predictive information about the intensity of a sound may not influence the elicited neural response. This means we cannot determine whether this effect is limited to ecologically valid visual stimuli or whether there are certain generalizable properties of visual collisions that can be distilled to create intensity expectations within abstract cues, and finally whether one can learn to associate an irrelevant cue with an intensity expectation. Nonetheless, this study has demonstrated that visually based intensity expectations regarding auditory intensity may bias the amplitude of the N1. Future studies will need to determine the boundaries of this effect.

Finally, in a post hoc analysis, we unexpectedly identified a difference in anticipatory activity between the "strong" and "weak" clap videos: The "strong clap" videos elicited a negative-going potential with a frontal topography from approximately 120 msec preclap. Although we are agnostic as to the underlying cause of the negative going deflection, one possibility is that it reflects a stimulus preceding negativity (SPN; Brunia, 1988; Van Boxtel & Böcker, 2004). The SPN is a slow, negative-going potential that is elicited in anticipation of affective stimuli. Although the most common stimuli used to elicit the SPN are those which generate a significant affective or physiological response (e.g., electric shocks, pictures of opposite-sex nudes), it is possible that the anticipation of a loud (aversive) sound in response to the "strong clap" video could have been sufficient to elicit an SPN. Regardless of the identity of this component, it is important to note that it is unlikely to be responsible for the between-condition differences in N1 amplitude we observed (i.e., the primary result of this study). That is, we subtracted out the activity elicited in the V-only condition from the AV condition to create the corrected-audiovisual waveform (AV−V), as is common in studies of this nature (Stekelenburg & Vroomen, 2007). As illustrated in Figure 6.3A, there were no systematic differences in prestimulus activity between the (corrected) "strong clap" and "weak clap" videos at 72.5 dB. This result suggests that the observed between condition differences in N1 amplitude between conditions were not the result of differences in prestimulus activity between the "strong" and "weak" clap videos.

In conclusion, this study has shown that the early evoked neurophysiological response to an auditory stimulus is dependent not only on the intensity of the stimulus but also on one's expectations regarding the intensity of the stimulus.

7 General Discussion

The primary aim of this thesis was to establish whether visual information about the likely intensity of an auditory event can affect the perceived loudness of that event and the associated activation of the auditory cortex. This aim was firstly addressed by using psychophysics to quantify the perceived loudness of sounds that were visually cued as being generated at different egocentric distances or with different amounts of power. It was also addressed by using EEG to quantify the evoked auditory response to sounds that were visually cued as being generated with different amounts of power. In this final chapter, I summarize the main findings of the empirical research conducted in this thesis (7.1), I discuss the implications for our understanding of audition (7.2), I provide directions for future research (7.3), and I conclude the thesis (7.4).

7.1 Summary

Chapter 3 investigated how loudness was affected by viewing a sound source at different egocentric distances. The aim was to determine whether visual cues to a sound source's distance can facilitate loudness constancy, which is the ability to perceive the invariant loudness of an auditory object when presented at different distances. When examining loudness constancy, estimates of sound intensity may be gauged either based on the loudness of a sound at-the-ear or on the estimated power of the sound source at its location. This chapter focused on the apparent loudness of sounds. We simulated a loudspeaker presenting sounds in anechoic conditions using a computer monitor and headphones. The simulation involved a loudspeaker that presented sounds at different distances in an open field that had rich monocular cues to depth. Using a Bayesian 2IFC procedure we demonstrated that participants' judgments of loudness were unaffected by the concomitant visual cues as to the sound source's distance. This failure to find an effect was replicated across 6 experiments in which we employed two methods of speaker presentation (i.e., speakers appearing frontally or speakers viewed by a panning camera) and four different auditory stimuli (250Hz tone, pink noise, a spoken utterance ‘Ba’ and a 250Hz tone that underwent an ecological delay based on the distance the sound traveled from its source). Contrary to a loudness constancy hypothesis, these experiments provided evidence that depth cues were not integrated into loudness percepts in anechoic conditions, when simulating rich monocular cues to depth.

Chapter 4 continued the investigation into how perceived loudness is affected by cuing the egocentric distance of the sound source. In this chapter we considered the influence of both visual distance information and auditory distance information on loudness constancy. While Chapter 3 required participants to estimate the apparent loudness of received sounds, Chapter 4 shifted the task demands to require participants to estimate the power of the distal sound source. Once again, the visual cue was of a loudspeaker that presented sounds at different egocentric distances, however now the loudspeaker was simulated to present within a concert hall with a reverberant sound field. Loudness judgements were extracted using a Bayesian 2IFC procedure. We generated auditory signals by convolving binaural room impulse responses (BRIRs) recorded at each presentation distance in the hall with white noise. This simulated the properties of auditory input that would be produced by the sound source at each position. Visual information was provided by photographs of the speaker at each presentation distance in the same hall. In the audio-only condition we found that partial loudness constancy was facilitated by the auditory cues alone. We did not find that introducing congruent visual information improved loudness estimates when compared against the audio-only condition. Moreover, when we manipulated the visual depiction of the sound source to appear closer or farther than the paired auditory signal, we found that visual information did affect loudness estimates. If visual information was used to accurately estimate source power, auditory input from a closer source would need to be attenuated relative to auditory input from a farther source. In contrast, however, we found that input from the comparison speaker was judged as being amplified when the visual depiction of the speaker was both further *and* closer than the auditory signal. Importantly, source loudness estimates were not attenuated when the speaker was visually shifted to appear closer than the paired auditory signal. This result suggests that visual information about a sound source’s distance, when simulated through the use of photographs, is not integrated into source loudness estimates in a way that would facilitate loudness constancy.

Chapter 5 considered the possibility that the perceived loudness of a sound may be influenced by visual information about the generative power of its source. We employed a two-pair forced choice task to examine this question. In each trial, there were two pairs of stimuli (i.e., four intervals). Each pair contained a probe sound and an anchor sound. The probe was always a handclap sound presented without visual information, and was identical in both pairs. One of the anchors always included visual information as to the likely power of the sound source by depicting either a 'strong' or a 'weak' hand-clap. The level of the sound paired with the video depicting the 'strong' clap was 75dB, while the level of the sound paired with the video depicting the 'weak' clap was 65dB. The other anchor did not include any visual information about the intensity of the handclap. In this anchor the sound was always 10dB above or below the sound presented in the visible anchor. From trial to trial the probe sound varied in intensity between the ranges defined by the two anchors. The task was to nominate the pair in which the probe sound was more similar in loudness to its anchor. In employing a task that required participants to identify perceptual similarity, we reduced the ability for post-perceptual demand characteristics to influence responses. We found that visual information that was suggestive of a higher intensity sound (i.e., the video of the 'strong' clap) increased the perceived loudness of that sound. This finding demonstrated the perception of loudness is influenced in accordance with visually created expectations.

Chapter 6 used EEG to measure the neurophysiological activity of the brain and explored whether visual cues to source power could influence the auditory-evoked response. Specifically, we measured the auditory evoked response to sounds paired with visual cues that were suggestive of the likely power of the sound source. To do this, we employed the same stimuli that were used in Chapter 5 (i.e., the videos of the 'strong' and 'weak' claps). The 'weak' clap visual cue was paired with clap sounds delivered at 65dB and 72.5dB, whilst the 'strong' clap visual was paired with clap sounds delivered at 72.5dB and 80dB. The activation of the auditory cortex was indexed by the amplitude of the N1 component of the auditory evoked potential. We found that the amplitude of the N1 component generated from *identical* clap sounds (i.e., at 72.5 dB) was greater when visual cues suggested that the sound source was more powerful. This result establishes that visual information can regulate the activation of the primary auditory cortex at very early stages (i.e., <100ms post sound) in accordance with visually created expectations.

7.2 Implications

The experiments presented as part of this thesis advance our understanding of how visual information about the generation of an auditory signal is handled by our sensory system. I now explore the implications of these findings for the understanding of how subjective loudness is affected by 1) visual cues to a sound source's distance and 2) visual cues to a sound source's power. Following this I will discuss why some visual information does seem to influence the subjective loudness of an auditory event, while other visual information does not. I will then explore how these findings may be contextualised within the broader principles of perceptual inference.

7.2.1 The coding of auditory intensity based on visual distance cues

Chapter 3 and 4 of this thesis demonstrate that visual information regarding the distance of a sound source does not support the phenomenon of loudness constancy. In Chapter 3, I provided evidence that visual information regarding the distance of a sound source does not influence apparent loudness estimates in an anechoic sound field. The follow up experiment in Chapter 4 employed two factors that have previously been found to facilitate a maximal degree of loudness constancy. These factors were placing the sound source in a reverberant sound field (Mershon et al., 1981), and requiring participants to estimate the *distal* power of the sound source (Mohrmann, 1939). In this chapter, I did not find evidence that congruent visual signals aided participants in estimating source power in a reverberant sound field. To further tease apart any potential influence of visual information on loudness estimates, I manipulated the visual cues so that the sound source appeared to be placed systematically closer or farther than the location of the sound source delivering the auditory signal. Critically, the results of this manipulation indicated that visual distance cues were not recruited in a manner that would increase the accuracy with which a source's power could be estimated.

A number of early studies that investigated loudness constancy produced findings that were either inconsistent with the results of this thesis (Mohrmann, 1939), or inconclusive (Shigenaga, 1965; von Fieandt, 1951). Mohrmann (1939) employed a method of adjustment in which participants were required to alter the intensity of a close reference speaker to approximate the intensity of a comparison speaker shown at different distances. The findings of Mohrmann (1939) suggested that participants demonstrated loudness constancy, and that the degree of constancy was higher when the speaker was visible as opposed to when it was not visible. von Fieandt (1951) and Shigenaga (1965) employed a similar method of adjustment and again, loudness constancy was demonstrated. However, of these two studies, only von Fieandt (1951) examined loudness constancy when both visual cues to source distance were provided and when they were not (i.e, loudness estimates performed in the dark). In von Fieandt's (1951) experiment the influence of vision on loudness estimates was difficult to assess as performance was similarly close to complete constancy in both the visual and non-visual conditions. It is noteworthy that, in all of these early experiments, visual cues were instantiated by physically moving the egocentric distance of the sound source. This meant that visual cues to the distance of the sound source covaried with auditory cues to the distance of the sound source. Accordingly, it is possible that auditory cues interacted with visual distance cues to give rise to partial loudness constancy. This is a plausible confound as the ratio of direct-to-reverberant energy in an auditory waveform varies as a function of sound source distance and this cue has been found to inform source loudness estimates (Kolarik et al., 2016; Zahorik & Wightman, 2001). With this in mind, of these three studies, only Mohrmann (1939) provided specific evidence that visual information influences source loudness estimates over and above auditory cues alone.

More recent research investigating whether loudness constancy is influenced by vision has attempted to better control for the effect of auditory cues on subjective loudness. In an experiment by Mershon et al. (1981), participants estimated the loudness of sounds delivered from a static location (through a hidden loudspeaker) whilst a silent but visible dummy loudspeaker moved between 3 distances (75, 225 and 375cm). When this task was conducted in a reverberant sound-field, participants demonstrated a loudness-constancy like effect in which loudness estimates increased as the apparent distance of the source increased. However, participants failed to demonstrate a loudness-constancy like effect when this task was conducted in an anechoic sound-field. More recently, Altmann et al. (2012) paired short noise bursts with the onset of a light source at varying distances in a dark anechoic room, and measured perceived loudness. This study found no evidence that the distance of the light source behaviourally influenced apparent loudness estimates. It is possible, though, that the limited availability of visual depth cues in this experiment played a role in this null effect. Berthomieu et al. (2019) attempted to re-address the question of whether viewing the depth of a source can influence the perceived loudness of auditory input by improving the quality of visual cues to source depth. They did this by employing a virtual reality environment and simulating sounds coming from a sound source at 5 different distances within three different sound fields (a sports hall, a concert hall and anechoic conditions). It was found that apparent loudness estimates were not influenced by the presence of visual cues depicting the distance of the sound source. The study of Berthomieu et al. (2019) provides one of the most well controlled investigations of visual cues on loudness constancy currently available. Despite this, a question that remained unanswered in their experiment was: does loudness constancy occur when the task demands shift from estimating the apparent loudness of a sound to estimating the distal power of the sound source? Chapter 3 extends the findings of Altmann et al. (2012) to suggest that visual cues to source distance do not facilitate loudness constancy in an anechoic open field when rich cues to depth are provided. Chapter 4 extends the conclusions of Berthomieu et al. (2019) to suggest that in a reverberant sound field visual cues to the distance of a sound source do not facilitate loudness constancy, and that this conclusion holds when participants make estimates of distal source power.

It could be argued that the conclusions drawn from Chapters 3 and 4 are constrained by our use of experimental designs that employed the simulation of depth using a computer monitor, a simulation that did not employ binocular cues. I would make two responses to this point. First, both Altmann et al. (2012) and Berthomieu et al. (2019) utilised binocular cues to depth (i.e., stereopsis) and both experiments returned results that are consistent with the results of the present thesis. This suggests that the loudness findings based on monocular cues to depth demonstrated in Chapter 3 and 4 are likely to apply to conditions in which binocular cues to depth are also present. Second, Chapter 4 utilised visual stimuli which have previously been used to procure estimates of sound source distance (Anderson & Zahorik, 2014). In the experiment of Anderson and Zahorik (2014) it was demonstrated that the same photographs depicting the sound source at different distances facilitated relatively accurate estimates of that source's distance. Given that Chapter 4 employed exactly the same stimuli that was used by Anderson and Zahorik (2014), we can assume this visual information was sufficient to accurately cue the distance of the sound source.

7.2.2 The coding of auditory intensity based on visual power cues

Chapters 5 and 6 of this thesis demonstrated that visual information about a sound source's power is integrated into both the subjective perception of its loudness and the auditory evoked activity of the brain. Using stimuli depicting an audiovisual handclap, Chapter 5 provided evidence that visual stimuli that cued the power of a sound source could influence the subjective experience of loudness. This fascinating finding is akin to the McGurk effect (McGurk & MacDonald, 1976), in which the auditory percept is constructed by fusing auditory input with visual stimuli that cued the phonetics of a sound. By utilising a novel experiment design, Chapter 5 is the first study (to my knowledge) to provide evidence that this result is driven by a true perceptual effect and not a response bias. In Chapter 6, using the same clap stimuli, I found that the activation of the auditory cortex is also regulated in accordance with the visually created expectation. This finding corroborates the results of Chapter 5, and suggests that visual cues as to a sound source's power can influence the primary sensory coding of its auditory intensity.

The only other study (to my knowledge) that has investigated the effect of visual cues to the power of a sound source on loudness, also found evidence that the perception of loudness is constructed through the combination of auditory input and visual information about the cause of auditory input. Using both videos and audio recordings of speech and non-speech stimuli (i.e., clapping), Rosenblum and Fowler (1991) required participants to 1) rate the amount of perceived effort that the actor put into generating the sound, and 2) rate the loudness of sounds that were paired with videos of sound-generating actions that appeared to vary in terms of the level of effort they required. The results indicated that the perceived loudness of the sounds increased when they were paired with visual cues that were perceived to require more effort. Unfortunately, the study design of Rosenblum and Fowler (1991) could not distinguish between post-perceptual decisions and genuine perceptual effects. This is problematic as the associated loudness of a visual cue may logically inform a participant's loudness 'decision', especially on uncertain trials. For example, we 'know' that a visual cue of a high-effort clap is related to the production of a louder sound than a low-effort clap. This may *cognitively* influence a participant's response when rating the loudness of two sounds, even in the case that they are perceptually experienced as being the same. We addressed this problem in the behavioural study of Chapter 5, which was designed specifically to account for such post-perceptual biases. We accounted for bias by shifting the task demands so that they could not be related to the task response. Specifically, we did not require participants to estimate which stimuli were more or less loud; in contrast we required participants to estimate which sounds were the most similar in loudness.

Chapter 5 and 6 provide unambiguous support for the notion that visual cues to the power of a sound source can influence subjective judgements of loudness, and that this influence is based on a genuine perceptual effect as opposed to a post-perception bias. Furthermore, these studies also suggest that visual information about the generative process involved in producing an auditory signal is integrated into the neural coding of auditory intensity. In the section below, I elaborate on how and why visual cues to the power of a sound source may specifically influence auditory perception.

7.2.3 When does visual information influence our coding of auditory intensity?

A natural question that emerges from the findings of this thesis is: why is subjective loudness influenced by causal cues to a sound source's power but not by causal cues to a sound source's depth? In the section below, I discuss four theories that attempt to explain why some visual signals do influence the neural and behavioral representations of auditory intensity, while others do not.

1. To construct functional representations of the world the brain encounters an inference problem. An identical signal arriving at a sensory receptor may be the result of multiple different environmental generators 'out there' in the world. A common suggestion is that in order to construct perceptual representations of our environment, the brain bridges the chasm of uncertainty by *actively inferring* the most likely external cause of received input. The synthesis of cues from multiple sensory modalities is one process that the brain uses to construct useful representations of external sources (e.g., Campbell, 2007). Importantly, it has been suggested that during multisensory integration we optimise the precision of our combined percept by weighting each stream of information based on its relative reliability, with more reliable information given a higher weighting (Alais & Burr, 2004; Alais et al., 2010; M. O. Ernst & Bülthoff, 2004). If we assume that our perceptual system also operates in this manner when synthesising information about auditory intensity, we can speculate that if a cue predicts auditory intensity more reliably, it too will have a greater weighting compared to an unreliable cue. There is reason to believe that cues to a sound source's *power*, and cues to its *distance* will differ with respect to how reliably they can predict the intensity of an auditory event. For instance, there are multiple parameters that can determine how much a signal will attenuate based on its distance from the listener. In anechoic conditions, every time the distance of a sound source is doubled, the auditory intensity decreases by approximately 6 dB (Coleman, 1963). In reverberant sound fields, however, the rate of attenuation can be significantly lower (Zahorik, 2002a). Likewise, whether a sound source projects its signal uniformly, or whether the sound is focused in a specific direction, can affect the degree of attenuation a signal will undergo (Kolarik et al., 2016), as can environmental factors such as the wind (Traunmüller & Eriksson, 2000). Additionally, objects in the sound field can also reduce and absorb some of a sound's energy before it arrives at the ear (Aylor, 1972; D. I. Cook & Van Haverbeke, 1974; Kurze, 1974). Collectively, these parameters alter the governing relationship between source distance and aural intensity. It is plausible that these parameters reduce the reliability with which information about the distance of a sound source may be utilised to infer either auditory intensity at-the-ear, or the power of the sound source at its location. In contrast, when the source distance is fixed, visual information about the power of a sound source may be a reliable predictor of the intensity of an auditory event (both at-the-ear and at its location). There is a clear association between certain cues as to source power, and the loudness of the resulting sound. An exemplar of this can be seen in the manipulation check of Chapter 5, in which 100% of participants responded that a silent video of the 'strong' clap would have resulted in the production of a louder sound than a silent video of the 'weak' clap. If cues to the power of the sound source provide more reliable estimates of a sound's intensity than cues to the distance of a sound source, theories of optimal integration would suggest that source power cues would acquire a higher weighting (Alais & Burr, 2004; Alais et al., 2010; M. O. Ernst & Bülthoff, 2004). In short, this theory posits that the reliability with which cues predict the intensity of an auditory event will determine the extent to which visual information influences perceived loudness.

2. A neurophysiological perspective may suggest that the mechanism by which vision influences the perception of loudness occurs through the interaction of bottom-up sensory input with top-down neural predictions (van Laarhoven et al., 2017). In the EEG and MEG literature it has been suggested that to enable the early integration of visual information in auditory areas, visual cues to the sound need a predictive 'head start' that is greater than 100ms (Aoyama et al., 2006; Senkowski et al., 2007; Vroomen & Stekelenburg, 2010). Furthermore, to measure the suppression of the evoked neural response to a sound, a precise predictive template as to the likely temporal onset of the forthcoming sound needs to be established (Elijah, Le Pelley, & Whitford, 2018; Hughes et al., 2013; van Laarhoven et al., 2017). This template possibly also requires precision with respect to predicting the intensity of the sound, which would ostensibly be established through the combination of source distance and source power cues. In Chapters 3 and 4, we visually manipulated sound source distance but failed to identify an effect on perceived loudness. In these experiments sounds were delivered from computer-simulated loudspeakers. Loudspeakers do not provide anticipatory temporal cues as to the onset of auditory input, nor do they provide anticipatory cues as to the power of a source. Thus, based on this visual information, it would be impossible to form a precise predictive template prior to the reception of the auditory input. In Chapters 5 and 6 participants were provided with visual signals (i.e., hands moving towards collision from a fixed distance) which cued the temporal onset of the sound, the distance of the sound source, and the power of the sound source. With these three cues it would presumably be feasible to use visual information to form a distinct template that predicts the auditory event. In summary, this theory posits that for sensory cues to be integrated into the neural coding of auditory intensity, a precise *anticipatory* template of forthcoming input is required.

3. There may be a specific perceptual mechanism dedicated to mirroring actions, and this mechanism may facilitate the capacity for certain visual cues to influence perceived loudness (Grèzes et al., 1999; Iacoboni et al., 1999; Konorski, 1967; Koski et al., 2002). Chapters 3 and 4 employed visual stimuli (i.e., loudspeakers) in which no physical actions were used to generate the auditory signal, and in these experiments visual cues did not influence perceived loudness. Chapters 5 and 6 employed visual stimuli (i.e., hands clapping) in which the auditory signal was generated by a physical action, and in these experiments the visual cues were found to influence both perceived loudness and auditory-evoked activity. To my knowledge, only one other study that has investigated how cuing of the power of a sound source influences loudness. This study, by Rosenblum and Fowler (1991), also employed stimuli that contained goal directed actions: namely speech, and clapping. Consequently, we cannot rule out the possibility that the effect of visual cues to sound source power on perceived loudness is driven by an action-specific mechanism. It has been suggested that viewing an action may evoke similar perceptual processes to conducting the action itself, and that this may aid in learning and imitation (e.g., Iacoboni et al., 1999). Consistent with this hypothesis, it has been found that parts of the auditory cortex will become active during lip reading (Calvert et al., 1997; Pekkola et al., 2005). Studies have also found that Broca's area is not only involved in viewing speech, imitating speech and silent lipreading, but also when viewing goal directed hand movements (Grèzes et al., 1999; Iacoboni et al., 1999; Koski et al., 2002). Given that all previous studies that have demonstrated an effect of visual information on perceived loudness have employed stimuli that depict gestural movements, it is possible that this effect is driven by a mechanism that functions by 'mirroring' the sound-generating event in some capacity.

4. Repeatedly pairing a visual-cue with specific auditory-input could generate a learned association, which may influence the perceived loudness of the auditory input. A limitation of Chapters 5 and 6 was that, in both experiments, the video of the 'strong' clap had a higher probability of being paired with a higher intensity sound than a video of the 'weak' clap. This design was necessary to avoid a prediction error in which the video depicting a powerful sound generating action is paired with 'unnaturally' soft auditory-input, and vice versa. However, this design feature opens up the possibility of a learned association being developed between the visual cue and the intensity of auditory input. It is plausible that our auditory system may have highly malleable prior expectations that update rapidly given a specific context. According to this theory, the specific content of the visual stimulus may be incidental; it is the learned stimulus-association that drives the effect of a given cue on perceived loudness. Similar to my first theory, this theory also involves the regulation of loudness based on the intensity distributions that are associated with a visual cue. However, while the first theory assumes that participants employ a generative model to determine the causal influences of source power (i.e., cues that are either innate, or learned over neurodevelopment), the present theory assumes that any stimuli can be associated with source power, and that these associations can be learned over the time course of an experiment.

7.2.4 The representation of loudness based on visual cues - what do we still need to resolve?

The results of this thesis suggest that visual information about a sound source's power is integrated into a representation of its associated auditory consequences. What requires further attention is whether this effect reflects the perceptual system estimating auditory events proximally at-the-ear, or distally at their source. We can equally imagine that either of these operations could be adaptive in extracting useful information from our environment. For example, imagine that you are walking through the jungle and in the distance you see a tremendous silverback gorilla screaming and beating its chest. Is there cause for alarm? It may be that perceiving the power of the distal source is useful as it provides us with an indication of this creature's strength and tone (i.e., high intensity sounds indicates a strong aggressive animal). Alternatively, it may be that perceiving the sound at-the-ear is also useful, as the auditory signal has attenuated in intensity as it has travelled from the source, and this provides some indication that the source is not an immediate threat because it is far away. To understand if our perceptual system is geared towards estimating auditory events at-the-ear or at their location, we have to determine whether we discount the variations in intensity caused by variations in the distance of a sound source.

In the experiments of Chapters 5 and 6, the distance of the sound source remained fixed. As a result, these chapters could not disentangle whether our perceptual system constructs our perception of loudness based on the intensity expected at-the-ear or at-the-source. That is, the visual cues depicting the 'strong' clap at a fixed location would produce a higher intensity signal regardless of whether the estimation is aimed at-the-ear or at-the-source; likewise visual information depicting the 'weak' clap would produce a lower intensity signal regardless of whether the estimation is aimed at-the-ear or at-the-source. Without manipulating source distance, any causal inferences about the intensity of the signal at-the-ear or at the source would predict a perceptual shift of loudness in the same direction. In Chapters 3 and 4 we investigated how visual cues to source distance influence loudness estimates. To experience loudness constancy as facilitated by visual distance cues, more distant sound sources need to be experienced as relatively amplified in loudness. Neither Chapter 3 or 4 found any evidence of this. Thus, these chapters suggest that our perceptual system estimates the proximal input of each auditory event at-the-ear and not at its source. Importantly, the experiments of Chapter 3 and 4 included source distance cues *without* providing cues to the power of the sound source. It is possible that cuing the power of a sound source is a requirement for the integration of source distance information. In light of this, the results of this study provide support for the more limited conclusion that intensity is represented at-the-ear when source power cues are absent.

An additional question that emerges from Chapters 5 and 6 is whether the effect of visual cues to source power on perceived loudness is 'relative' or 'absolute'. If it is a 'relative' effect, this would imply that the biasing effect of visual information on perceived loudness is relative to the intensity of the sound it is paired with. In the context of our experiment, for example, is it possible that the 'weak' clap video predicted a distribution of sounds around 60 dB. If we hypothesise a 'relative' effect, then a sound less intense than 60 dB should have its perceived loudness amplified by the visual cues, while a sound above 60 dB should have its perceived loudness attenuated. On the other hand, if we hypothesise an 'absolute' effect, then the 'weak' clap video would always attenuate the perceived loudness of a sound presented at any intensity. Further research is needed to disambiguate these two possibilities.

7.3 Future research

The empirical work described in this thesis demonstrated how visual information about the generation of auditory signals can influence the subjective and neurophysiological processing of auditory intensity. These findings open up multiple hypotheses that require further consideration. I will now briefly highlight four future research directions that offer promising opportunities to develop our understanding of how visual information influences auditory processing.

As mentioned previously, given that the results of this program suggest that source power cues influence perceived loudness, but source distance cues do not, it would be useful to revisit the question of whether the estimation of auditory events occurs proximally or distally *when source power cues are present.* To address this question it would be necessary to present a sound source which is visually cued as having certain amounts of power (e.g., a video of a person performing a 'strong' and 'weak' handclap) at different distances, and compare it to a condition in which the same sounds are presented in the absence of visual power cues (e.g., a video of a loudspeaker). If the intensity of the auditory event is estimated at-the-ear, loudness estimates will be unaffected by the distance of the sound source in both conditions. However, if loudness constancy is expressed when source power cues are present, we hypothesise that a 'flatter' loudness-distance function would result for the conditions in which visual power cues are present. That is, despite the sound source's distance altering the intensity of the signal at the ear, we hypothesise that when source power cues are present, participants would experience relatively invariant loudness across the presenting distances.

Isolating the specific causal features that can influence perceived loudness is also critical to extending the conclusions of this research program. In order to achieve this, it is again useful to consider the primitive components in the generative process of an auditory signal. In section 2.2, I presented the influence graph (Figure 2.1) that provided the scaffolding for the hypotheses tested in this thesis. Now that we have presented evidence that source power cues influence loudness, it is possible to create a more detailed generative model to further specify the causal influences that go towards determining the power of the sound source. Any visual information that indicates the transfer of kinetic energy into sounds holds the 'potential' to predict the power of that sound. We know that the velocity and mass of objects in a collision are two factors that determine the amount of kinetic energy that is transferred into the intensity of the resultant sound (Rienstra & Hirschberg, 2004). These two features (velocity and mass) can thus be visually manipulated to create differing predictions of the intensity of an auditory event. Related to this, a study by Fassnidge and Freeman (2018) demonstrated that silent videos that depicted moving patterns with high 'motion energy' were more likely to induce an illusory auditory experience than patterns with low 'motion energy'. Here 'motion energy' was quantified with a computational model that captured the degree to which patterns of luminance changed over space and time. Videos with high 'motion energy' contain high amounts of flickering or movement. It could be that 'motion energy' mimics the movement of objects transferring kinetic energy into auditory energy. In this case, we would hypothesise that visual cues that contain higher 'motion energy'- which mimic the physical vibrations of a sound source with greater power - would increase the perceived loudness of sounds compared to visual cues that contain low 'motion energy'. Conducting an experiment of this nature offers the opportunity to address two prospective questions put forward in section 7.2.3, as 'motion energy' stimuli 1) do not have to be action dependent and 2) can be generated without an anticipatory period that temporally cues the onset of the sound. The exploration of motion energy and other such features will help us better understand what visual cues regarding the causes of sounds are used to predict a sound's intensity, and to what degree they influence perceived loudness.

On the other hand, it may be that visual cues do not need to relay information about the ecological cause of auditory input to influence loudness perception, but rather that any visual stimulus may be conditioned to have the same influence. To test this, it would be necessary to associate visual stimuli that do not contain cues as to the cause of an auditory event with different sound intensities. This could be carried out using a visual stimulus such as the hand of a clock that triggers a sound when the hand reaches a certain point. Here the cue to auditory intensity may be related by the colour of the clock hand. One colour (e.g., blue) would have a higher likelihood of being paired with a higher intensity sound (72.5 dB, 80 dB), while the other colour (e.g., red) would have a higher likelihood of being paired with a lower intensity sound (65 dB, 72.5 dB). If loudness expectations are established through short term associations formed over the course of an experiment, then the presentation of the blue hand should lead to a sound being judged as louder than when the red hand is presented, even when the sound itself is the same intensity in both conditions (i.e., both paired with the 72.5 dB sound). A key question of interest in this experiment would be the time scale over which the learned association develops. A second question of interest would be whether this learned association is context dependent. For example, it may be that while we learn that the blue hand predicts a louder noise in a silent laboratory, the same is not true if we go outside to the noisy street, where it is presumably adaptive to establish new priors that are dependent on the new acoustic parameters of this environment. It is, of course, also possible that we can learn new sound intensity associations, whilst still having less mutable or even innate prior expectations related to the causal nature of sound sources that have been established over a longer developmental period.

In the neurophysiological experiments of this thesis I have focused on the N1 component, but other auditory evoked components may provide further insights into the neural integration of auditory intensity. The P50 and P2 are components of the auditory evoked potential of particular interest. The p50 is a positive deflection of the auditory evoked potential that occurs 30-50ms after the onset of a sound and is thought to reflect sensory gating (Lijffijt et al., 2009; White & Yee, 2006). The P2 another positive deflection of the auditory evoked potential that occurs 170-200ms after the onset of a sound and like the N1 it is also thought to be sound intensity dependent (Adler & Adler, 1989; Paiva et al., 2016). Examining the neural correlates of when loudness will be influenced by higher order cues will help us develop our understanding of how intensity is represented in the auditory cortex.

Lastly, as described in section 7.2.4 we are yet to determine whether the effects observed in Chapters 5 and 6, are indicative of an 'absolute' or 'relative' shift in auditory intensity coding. As previously discussed, in order to disentangle these possibilities, it would be necessary to pair visual cues that indicate the power of the sound source (i.e., a 'weak' and 'strong' clap) with a wider distribution of sound levels. If visual cues to source power influence loudness in a 'relative' sense, we would expect sounds that are of a lower intensity than predicted by the visual cue to have their perceived loudness amplified, while sounds that are of a higher intensity than predicted by the visual cue would have their perceived loudness attenuated. In contrast, if visual cues to source power influence loudness in an 'absolute' sense, we would expect perceived loudness to be reduced in the 'weak' clap conditions and amplified in the 'strong' clap conditions at all sound levels.

7.3.1 Potential implications

Psychoacoustic research has effectively charted how the low-level physical features of sound-waves relate to the experience of subjective loudness (Moore, 2012). However, these 'low level' conceptualisations of loudness do not account for the influence of higher order information (M. Epstein & Florentine, 2009, 2012; Moore, 2014). The results of this thesis may have interesting implications for improving the application of models of loudness out in the 'real world', where visual cues are abundantly present. We have demonstrated that information provided by 'seeing volume' (i.e., where there are visual cues to expected auditory intensity) can influence the perception of loudness. If it is possible to introduce a visual parameter into loudness models, this may increase the accuracy with which functions that predict subjective loudness will map onto the actual experience of loudness. Such a development would have particular relevance for aging populations. It has been well established that aging populations often experience hearing loss (Fozard, 1990). It has also been established that aging populations demonstrate enhanced multisensory integration (Laurienti et al., 2006). Taken together, these findings suggest that the effect of visual stream information on perceived auditory intensity may be heightened in aging populations. If this is the case, loudness models – which are commonly used to inform both the design and use of audiograms and hearing aids – could be improved by considering a visual parameter in the coding of loudness (Florentine & Zwicker, 1979; Moore & Glasberg, 2004; Moore et al., 1997, 2010).

A second potential 'real-world' implication for this research relates to hallucinatory experiences in psychotic populations. People with schizophrenia have been found to show deficits in auditory perception (Matthews et al., 2013). Furthermore, the expression of hallucinatory symptoms in this population has been suggested to relate to the abnormal regulation of top-down expectations on sensory input, particularly in the auditory domain (Powers, Kelley, & Corlett, 2016; Schmack, Rothkirch, Priller, & Sterzer, 2017; Sterzer et al., 2018). It has been suggested that in some situations people with schizophrenia overweight incoming sensory input and underweight top-down inferences (Corlett, Frith, & Fletcher, 2009; Dima et al., 2009; P. C. Fletcher & Frith, 2009; Friston, Stephan, Montague, & Dolan, 2014; Sterzer, Mishara, Voss, & Heinz, 2016), and in other situations do the opposite (Cassidy et al., 2018; Ćurčić-Blake et al., 2013; Friston, 2005a; Powers, Mathys, & Corlett, 2017). In an attempt to reconcile these contradictory hypotheses, it has been proposed that top-down inferences may be over or under weighted depending on the complexity of the inference, with simple low-level inferences being underweighted and more complex higher level inferences being overweighted (Kwisthout, Bekkering, & Van Rooij, 2017; Sterzer et al., 2018). The paradigms of Chapters 5 and 6 provide a platform for developing our understanding of how information as to the generative causes of a sensory signal will regulate *both* the subjective and neurophysiological coding of auditory events in people suffering from psychotic disorders such as schizophrenia. On the basis of the literature two competing hypotheses are possible: either populations with schizophrenia will code the intensity of auditory events in a way that is (a) less biased by visual source power cues due to the weaker regulation of top down generative information, or (b) more biased by visual source power cues due to the enhanced regulation of top down generative information. Exploring whether visual information about the causal generators of an auditory signal is over or under employed by psychotic populations may provide insight into the experience of auditory-verbal hallucinations, in which people perceive sounds in the absence of appropriate generative cues.

7.4 Conclusion

Intensity is a fundamental property of the auditory signal which plays a key role in influencing both our perceptual and neurophysiological response to sounds. The data presented in this thesis suggest that visual cues as to the nature of a sound-producing event can regulate both the subjective perception of loudness and the evoked response of the auditory cortex. These results demonstrate that perceived loudness and its neurophysiological correlates are not only driven by low-level auditory information (such as aural intensity), but also by high-level predictions *about* the expected intensity of the sound. Taken together, these findings emphasise the functional role of causal information in disambiguating our perceptual representations of auditory events.

References

Adler, G., & Adler, J. (1989). Influence of stimulus intensity on aep components in the 80-to 200-millisecond latency range. *Audiology*, *28*(6), 316–324.

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

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

Allen, G. D. (1971). Acoustic level and vocal effort as cues for the loudness of speech. *The Journal of the Acoustical Society of America*, *49*(6B), 1831–1841.

Allison, R. S., Gillam, B. J., & Vecellio, E. (2009). Binocular depth discrimination and estimation beyond interaction space. *Journal of Vision*, *9*(1), 10–10.

Altmann, C. F., Matsuhashi, M., Votinov, M., Goto, K., Mima, T., & Fukuyama, H. (2012). Visual distance cues modulate neuromagnetic auditory N1m responses. *Clinical Neurophysiology*, *123*(11), 2273–2280.

Altmann, C. F., Ono, K., Callan, A., Matsuhashi, M., Mima, T., & Fukuyama, H. (2013). Environmental reverberation affects processing of sound intensity in right temporal cortex. *European Journal of Neuroscience*, *38*, 3210–3220.

Anderson, P. W., & Zahorik, P. (2014). Auditory/visual distance estimation: accuracy and variability. *Frontiers in Psychology*, *5*, 1097.

Aoyama, A., Endo, H., Honda, S., & Takeda, T. (2006). Modulation of early auditory processing by visually based sound prediction. *Brain Research*, *1068*(1), 194–204.

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398.

Arnold, D. H., Johnston, A., & Nishida, S. (2005). Timing sight and sound. *Vision Research*, *45*, 1275–1284.

Aylor, D. E. (1972). Sound transmission through vegetation in relation to leaf area density, leaf width, and breadth of canopy. *The Journal of the Acoustical Society of America*, *51*(1B), 411–414.

Aylor, D. E., & Marks, L. E. (1976). Perception of noise transmitted through barriers. *The Journal of the Acoustical Society of America*, *59*(2), 397–400.

Baird, J. C. (1970). *Psychophysical analysis of visual space: International series of monographs in experimental psychology* (Vol. 9). Elsevier.

Barfield, W., & Rosenberg, C. (1995). Judgments of azimuth and elevation as a function of monoscopic and binocular depth cues using a perspective display. *Human Factors*, *37*(1), 173–181.

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

Barth, D. S., Goldberg, N., Brett, B., & Di, S. (1995). The spatiotemporal organization of auditory, visual, and auditory-visual evoked potentials in rat cortex. *Brain Research*, *678*(1-2), 177–190.

Berliner, J., & Durlach, N. (1973). A perceptual-anchor model for context coding in intensity perception. *The Journal of the Acoustical Society of America*, *54*(1), 336–336.

Berthomieu, G., Koehl, V., & Paquier, M. (2019). Loudness and distance estimates for noise bursts coming from several distances with and without visual cues to their source..

Besle, J., Fort, A., Delpuech, C., & Giard, M.-H. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, *20*(8), 2225–2234.

Besle, J., Fort, A., & Giard, M.-H. (2004). Interest and validity of the additive model in electrophysiological studies of multisensory interactions. *Cognitive Processing*, *5*(3), 189–192.

Bilecen, D., Seifritz, E., Scheffler, K., Henning, J., & Schulte, A.-C. (2002). Amplitopicity of the human auditory cortex: an fMRI study. *Neuroimage*, *17*(2), 710–718.

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, *14*(10), 693–707.

Blakemore, S.-J., Wolpert, D., & Frith, C. (2000). Why can't you tickle yourself? *Neuroreport*, *11*(11), R11–R16.

Brandt, J. F. (1972). Effects of stimulus bandwidth on listener judgments of vocal loudness and effort. *The Journal of the Acoustical Society of America*, *52*(2B), 705–707.

Brandt, J. F., Ruder, K. F., & Shipp Jr, T. (1969). Vocal loudness and effort in continuous speech. *The Journal of the Acoustical Society of America*, *46*(6B), 1543–1548.

Brechmann, A., Baumgart, F., & Scheich, H. (2002). Sound-level-dependent representation of frequency modulations in human auditory cortex: A low-noise fMRI study. *Journal of Neurophysiology*, *87*(1), 423-433. (PMID: 11784760)

Brocke, B., Beauducel, A., John, R., Debener, S., & Heilemann, H. (2000). Sensation seeking and affective disorders: characteristics in the intensity dependence of acoustic evoked potentials. *Neuropsychobiology*, *41*(1), 24–30.

Bronkhorst, A. W., & Houtgast, T. (1999). Auditory distance perception in rooms. *Nature*, *397*, 517–520.

Brunia, C. (1988). Movement and stimulus preceding negativity. *Biological Psychology*, *26*(1-3), 165–178.

Bruno, N., & Cutting, J. E. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General*, *117*(2), 161.

Brunswik, E. (1944). Distal focussing of perception: Size-constancy in a representative sample of situations. *Psychological Monographs*, *56*(1), i.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

Burr, D., Silva, O., Cicchini, G. M., Banks, M. S., & Morrone, M. C. (2009). Temporal mechanisms of multimodal binding. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1663), 1761–1769.

Butler, R. A., Levy, E. T., & Neff, W. D. (1980). Apparent distance of sounds recorded in echoic and anechoic chambers. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(4), 745.

Calcagno, E. R., Abregu, E. L., Eguía, M. C., & Vergara, R. (2012). The role of vision in auditory distance perception. *Perception*, *41*(2), 175–192.

Callan, D. E., Jones, J. A., Munhall, K., Kroos, C., Callan, A. M., & Vatikiotis-Bateson, E. (2004). Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of Cognitive Neuroscience*, *16*(5), 805–816.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., . . . David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*(5312), 593–596.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, *10*(11), 649–657.

Calvert, G. A., Spence, C., & Stein, B. E. (2004). *The handbook of multisensory* *processes* (G. A. Calvert, C. Spence, & B. E. Stein, Eds.). Massachusetts, USA: MIT press.

Campbell, R. (2007). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1001–1010.

Carlson, V. (1977). Instructions and perceptual constancy judgments. *Stability and constancy in visual perception: Mechanisms and Processes*, 217–254.

Cassidy, C. M., Balsam, P. D., Weinstein, J. J., Rosengard, R. J., Slifstein, M., Daw, N. D., ... Horga, G. (2018). A perceptual inference mechanism for hallucinations linked to striatal dopamine. *Current Biology*, *28*(4), 503–514.

Chapanis, A., & McCleary, R. (1953). Interposition as a cue for the perception of relative distance. *The Journal of General Psychology*, *48*(2), 113–132.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Coleman, P. D. (1962). Failure to localize the source distance of an unfamiliar sound. *The Journal of the Acoustical Society of America*, *34*(3), 345-346.

Coleman, P. D. (1963). An analysis of cues to auditory depth perception in free space. *Psychological Bulletin*, *60*(3), 302.

Collins, J. K. (1976). Distance perception as a function of age. *Australian Journal of Psychology*, *28*(2), 109–113.

Cook, D. I., & Van Haverbeke, D. F. (1974). *Tree-covered land-forms for noise control* (Vol. 263). Nebraska, USA: Forest Service, US Department of Agriculture.

Cook, M. (1978). The judgment of distance on a plane surface. *Perception & Psychophysics*, *23*(1), 85–90.

Corlett, P. R., Frith, C. D., & Fletcher, P. C. (2009). From drugs to deprivation: a Bayesian framework for understanding models of psychosis. *Psychopharmacology*, *206*(4), 515–530.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to loftus and masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–45.

Creem-Regehr, S. H., Willemsen, P., Gooch, A. A., & Thompson, W. B. (2005). The influence of restricted viewing conditions on egocentric distance perception: Implications for real and virtual indoor environments. *Perception*, *34*(2), 191–204.

Culling, J. F., & Edmonds, B. A. (2007). Interaural correlation and loudness. In *Hearing–from sensory processing to perception* (pp. 359–368). Berlin, Germany:

Springer.

Ćurčić-Blake, B., Liemburg, E., Vercammen, A., Swart, M., Knegtering, H., Bruggeman, R., & Aleman, A. (2013). When Broca goes uninformed: reduced information flow to Broca's area in schizophrenia patients with auditory hallucinations. *Schizophrenia Bulletin*, *39*(5), 1087–1095.

Da Silva, J. A. (1982). Scales for subjective distance in a large open field from the fractionation method: Effects of type of judgment and distance range. *Perceptual and Motor Skills*, *55*(1), 283–288.

Da Silva, J. A. (1985). Scales for perceived egocentric distance in a large open field: Comparison of three psychophysical methods. *The American Journal of Psychology*, 119–144.

Davis, E. T., & Hodges, L. F. (1995). Human stereopsis, fusion, and stereoscopic. *Human Stereopsis, Fusion, and Stereoscopic Virtual Environments. In Barfield, Woodrow and Furness, III, Thomas A.(Ed). Virtual Environments and Advanced Interface Design*, 145–174.

DeLucia, P. R. (1991). Pictorial and motion-based information for depth perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 738.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.

Dierks, T., Barta, S., Demisch, L., Schmeck, K., Englert, E., Kewitz, A., … Poustka, F. (1999). Intensity dependence of auditory evoked potentials (AEPs) as biological marker for cerebral serotonin levels: effects of tryptophan depletion in healthy subjects. *Psychopharmacology*, *146*(1), 101–107.

Dima, D., Roiser, J. P., Dietrich, D. E., Bonnemann, C., Lanfermann, H., Emrich, H. M., & Dillo, W. (2009). Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modelling. *Neuroimage*, *46*(4), 1180–1186.

Dunn, B. E., Gray, G. C., & Thompson, D. (1965). Relative height on the picture-plane and depth perception. *Perceptual and Motor Skills*, *21*(1), 227–236.

Ehrenstein, W. H., Arnold-Schulz-Gahmen, B. E., & Jaschinski, W. (2005). Eye preference within the context of binocular functions. *Graefe's Archive for Clinical and Experimental Ophthalmology*, *243*, 926–932.

Eisler, H. (1981). Sensations, correlates and judgments: Why physics? *Behavioural and Brain Sciences*, *4*, 193–194.

Elijah, R. B., Le Pelley, M. E., & Whitford, T. J. (2018). Act now, play later: Temporal expectations regarding the onset of self-initiated sensations can be modified with behavioral training. *Journal of Cognitive Neuroscience*, *30*(8), 1145–1156.

Ellermeier, W., & Faulhammer, G. (2000). Empirical evaluation of axioms fundamental to Stevens's ratio-scaling approach: I. Loudness production. *Perception & Psychophysics*, *62*(8), 1505–1511.

Epstein, M. (2007). An introduction to induced loudness reduction. *The Journal of the Acoustical Society of America*, *122*(3), EL74–EL80.

Epstein, M., & Florentine, M. (2009). Binaural loudness summation for speech and tones presented via earphones and loudspeakers. *Ear and Hearing*, *30*(2), 234–237.

Epstein, M., & Florentine, M. (2012). Binaural loudness summation for speech presented via earphones and loudspeaker with and without visual cues. *The Journal of the Acoustical Society of America*, *131*(5), 3981–3988.

Epstein, W. (1966). Perceived depth as a function of relative height under three background conditions. *Journal of Experimental Psychology*, *72*(3), 335.

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, *40*(4), 481–492.

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

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

Ernst, S. M. A., Verhey, J. L., & Uppenkamp, S. (2008). Spatial dissociation of changes of level and signal-to-noise ratio in auditory cortex for tones in noise. *Neuroimage*, *43*(2), 321–328.

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience*, *22*(13), 5749–5759.

Fassnidge, C. J., & Freeman, E. D. (2018). Sounds from seeing silent motion: Who hears them, and what looks loudest? *Cortex*, *103*, 130–141.

Fechner, G. T. (1860/1966). *Elements of psychophysics* (Vol. 1). New York, USA: Holt, Rinehart and Winston New York.

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

Fisher, G. H. (1968). Agreement between the spatial senses. *Perceptual and Motor Skills*, *26*(3), 849–850.

Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Bell System Technical Journal*, *12*(4), 377–430.

Fletcher, H., & Munson, W. A. (1937). Relation between loudness and masking. *The Journal of the Acoustical Society of America*, *9*(1), 1–10.

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48–58.

Florentine, M. (2011). Loudness. In M. Florentine, A. N. Popper, & R. R. Fay (Eds.), *Loudness* (pp. 1–15). New York, USA: Springer.

Florentine, M., Buus, S., & Poulsen, T. (1996). Temporal integration of loudness as a function of level. *The Journal of the Acoustical Society of America*, *99*(3), 1633–1644.

Florentine, M., Namba, S., & Kuwano, S. (1986). English definition of loudness, noisiness and annoyance and the comparison between UK, USA and Japan. In *Proceedings of the international conference on noise control engineering* (pp. 831–834).

Florentine, M., & Zwicker, E. (1979). A model of loudness summation applied to noise-induced hearing loss. *Hearing Research*, *1*(2), 121–132.

Foley, J. M. (1980). Binocular distance perception. *Psychological Review*, *87*(5), 411.

Fowler, C. A., & Rosenblum, L. D. (1991). The perception of phonetic gestures. *Modularity and the Motor Theory of Speech Perception*, 33–59.

Fox, P. T., & Raichle, M. E. (1986). Focal physiological uncoupling of cerebral blood flow and oxidative metabolism during somatosensory stimulation in human subjects. *Proceedings of the National Academy of Sciences*, *83*(4), 1140–1144.

Fozard, J. L. (1990). Vision and hearing in aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (Vol. 3, pp. 143–156). Massachusetts, United States: Academic Press.

Freides, D. (1975). Reply to Rudel and Teuber. *Psychological Bulletin*, *82*(6), 948.

Friston, K. J. (2005a). Hallucinations and perceptual inference. *Behavioral and Brain Sciences*, *28*(6), 764–766.

Friston, K. J. (2005b). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836.

Friston, K. J. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301.

Friston, K. J. (2012). Predictive coding, precision and synchrony. *Cognitive Neuroscience*, *3*(3-4), 238–239.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148–158.

Gamble, E. A. (1909). Minor studies from the psychological laboratory of wellesley college: Intensity as a criterion in estimating the distance of sounds. *Psychological Review*, *16*(6), 416.

Gardner, M. B. (1968). Proximity image effect in sound localization. *The Journal of the Acoustical Society of America*, *43*(1), 163–163.

Gebhard, J., & Mowbray, G. (1959). On discriminating the rate of visual flicker and auditory flutter. *The American journal of psychology*, *72*(4), 521–529.

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*(5), 473–490.

Gibson, E. J., & Bergman, R. (1954). The effect of training on absolute estimation of distance over the ground. *Journal of Experimental Psychology*, *48*(6), 473.

Gibson, E. J., Bergman, R., & Purdy, J. (1955). The effect of prior training with a scale of distance on absolute and relative judgments of distance over ground. *Journal of Experimental Psychology*, *50*(2), 97.

Gibson, E. J., Gibson, J. J., Smith, O. W., & Flock, H. (1959). Motion parallax as a determinant of perceived depth. *Journal of Experimental Psychology*, *58*(1), 40.

Gibson, J. J. (1950). The perception of visual surfaces. *The American Journal of Psychology*, *63*(3), 367–384.

Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition.* Psychology Press.

Gigerenzer, G., & Strube, G. (1983). Are there limits to binaural additivity of loudness? *Journal of Experimental Psychology: Human Perception and Performance*, *9*(1), 126–136.

Gilinsky, A. S. (1951). Perceived size and distance in visual space. *Psychological Review*, *58*(6), 460.

Glave, R., & Rietveld, A. (1975). Is the effort dependence of speech loudness explicable on the basis of acoustical cues? *The Journal of the Acoustical Society of America*, *58*(4), 875–879.

Godey, B., Schwartz, D., De Graaf, J., Chauvel, P., & Liegeois-Chauvel, C. (2001). Neuromagnetic source localization of auditory evoked fields and intracerebral evoked potentials: a comparison of data in the same patients. *Clinical Neurophysiology*, *112*(10), 1850–1859.

Goodfellow, L. D. (1934). An empirical comparison of audition, vision, and touch in the discrimination of short intervals of time. *The American Journal of Psychology*, *46*(2), 243–258.

Gratton, G., Coles, M. G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, *55*(4), 468–484.

Grèzes, J., Costes, N., & Decety, J. (1999). The effects of learning and intention on the neural network involved in the perception of meaningless actions. *Brain*, *122*(10), 1875–1887.

Guilford, J. P. (1954). *Psychometric methods*. New York, USA: McGraw-Hill.

Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology*, *28*(2), 240–244.

Hagen, M. A. (1974). Picture perception: Toward a theoretical model. *Psychological Bulletin*, *81*(8), 471.

Hagenmuller, F., Heekeren, K., Meier, M., Theodoridou, A., Walitza, S., Haker, H., ... Kawohl, W. (2016). The loudness dependence of auditory evoked potentials (LDAEP) in individuals at risk for developing bipolar disorders and schizophrenia. *Clinical Neurophysiology*, *127*(2), 1342–1350.

Hall, D. A., Haggard, M. P., Summerfield, A. Q., Akeroyd, M. A., Palmer, A. R., & Bowtell, R. W. (2001). Functional magnetic resonance imaging measurements of sound-level encoding in the absence of background scanner noise. *The Journal of the Acoustical Society of America*, *109*(4), 1559-1570.

Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *The Journal of the Acoustical Society of America*, *69*(6), 1763–1769.

Hart, H. C., Hall, D. A., & Palmer, A. R. (2003). The sound-level-dependent growth in the extent of fMRI activation in Heschl's gyrus is different for low-and high-frequency tones. *Hearing Research*, *179*(1-2), 104–112.

Hart, H. C., Palmer, A. R., & Hall, D. A. (2002). Heschl's gyrus is more sensitive to tone level than non-primary auditory cortex. *Hearing Research*, *171*(1-2), 177–190.

Hegerl, U., Gallinat, J., & Juckel, G. (2001). Event-related potentials: Do they reflect central serotonergic neurotransmission and do they predict clinical response to serotonin agonists? *Journal of Affective Disorders*, *62*(1-2), 93–100.

Hegerl, U., Gallinat, J., & Mrowinski, D. (1994). Intensity dependence of auditory evoked dipole source activity. *International Journal of Psychophysiology*, *17*(1), 1–13.

Hellman, R. P., & Zwislocki, J. (1963). Monaural loudness function at 1000 cps and interaural summation. *The Journal of the Acoustical Society of America*, *35*(6), 856–865.

Hellström, Å. (1979). Time errors and differential sensation weighting. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(3), 460.

Helmholtz, V. H. (1877). *On the sensations of tone (english translation aj ellis, 1885, 1954).* New York: Dover.

Hendrix, C., & Barfield, W. (1995). Relationship between monocular and binocular depth cues for judgements of spatial information and spatial instrument design. *Displays*, *16*(3), 103–113.

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision*, *4*(12), 1–1.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593-1623.

Howard, I. P., & Rogers, B. J. (2002). *Seeing in depth, vol. 2: Depth perception.* University of Toronto Press.

Howard, I. P., & Templeton, W. B. (1966). *Human spatial orientation.* London, UK: John Wiley & Sons.

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological Bulletin*, *139*(1), 133.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, *286*(5449), 2526–2528.

Irwin, R., Hinchcliff, L. K., & Kemp, S. (1981). Temporal acuity in normal and hearing-impaired listeners. *Audiology*, *20*(3), 234–243.

Jaekl, P., Seidlitz, J., Harris, L. R., & Tadin, D. (2015). Audiovisual delay as a novel cue to visual distance. *PloS One*, *10*, e0141125.

Jäncke, L., Shah, N., Posse, S., Grosse-Ryuken, M., & Müller-Gärtner, H.-W. (1998). Intensity coding of auditory stimuli: an fMRI study. *Neuropsychologia*, *36*(9), 875–883.

Jesteadt, W. (1980). An adaptive procedure for subjective judgments. *Perception & Psychophysics*, *28*(1), 85–88.

Jogan, M., & Stocker, A. A. (2014). A new two-alternative forced choice method for the unbiased characterization of perceptual bias and discriminability. *Journal of Vision*, *14*(3), 20–20.

Julesz, B. (1986). Stereoscopic vision. *Vision Research*, *26*(9), 1601–1612.

Kaas, J., & Collins, C. (2004). The resurrection of multisensory cortex in primates: Connection patterns that integrate modalities. *The Handbook of Multisensory Processes*, 285–293.

Keen, K. (1972). Preservation of constant loudness with interaural amplitude asymmetry. *The Journal of the Acoustical Society of America*, *52*(4B), 1193–1196.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol.*, *55*, 271–304.

Kim, W. S., Ellis, S. R., Tyler, M. E., Hannaford, B., & Stark, L. W. (1987). Quantitative evaluation of perspective and stereoscopic displays in three-axis manual tracking tasks. *IEEE Transactions on Systems, Man, and Cybernetics*, *17*(1), 61–72.

Kingdom, F. A., & Prins, N. (2010). *Psychophysics: A practical introduction.* London, UK: Academic Press.

Klein, E., Swan, J. E., Schmidt, G. S., Livingston, M. A., & Staadt, O. G. (2009). Measurement protocols for medium-field distance perception in large-screen immersive displays. In *2009 ieee virtual reality conference* (pp. 107–113).

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

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, *43*(24), 2539–2558.

Kolarik, A. J., Moore, B. C., Zahorik, P., Cirstea, S., & Pardhan, S. (2016). Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics*, *78*(2), 373–395.

Konorski, J. (1967). *Integrative activity of the brain; an interdisciplinary approach.* Chicago, USA: University of Chicago Press.

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*(16), 2729–2737.

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.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244.

Koski, L., Wohlschläger, A., Bekkering, H., Woods, R. P., Dubeau, M.-C., Mazziotta, J. C., & Iacoboni, M. (2002). Modulation of motor and premotor activity during imitation of target-directed actions. *Cerebral Cortex*, *12*(8), 847–855.

Künnapas, T. (1968). Distance perception as a function of available visual cues. *Journal of Experimental Psychology*, *77*(4), 523.

Kurze, U. J. (1974). Noise reduction by barriers. *The Journal of the Acoustical Society of America*, *55*(3), 504–518.

Kwisthout, J., Bekkering, H., & Van Rooij, I. (2017). To be precise, the details don't matter: on predictive processing, precision, and level of detail of predictions. *Brain and Cognition*, *112*, 84–91.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, *35*(3), 389–412.

Lange, K. (2011). The reduced N1 to self-generated tones: An effect of temporal predictability? *Psychophysiology*, *48*(8), 1088–1095.

Langers, D. R., van Dijk, P., Schoenmaker, E. S., & Backes, W. H. (2007). fMRI activation in relation to sound intensity and loudness. *Neuroimage*, *35*(2), 709–718.

Lasota, K. J., Ulmer, J. L., Firszt, J. B., Biswal, B. B., Daniels, D. L., & Prost, R. W. (2003). Intensity-dependent activation of the primary auditory cortex in functional magnetic resonance imaging. *Journal of Computer Assisted Tomography*, *27*(2), 213–218.

Laurienti, P. J., Burdette, J. H., Maldjian, J. A., & Wallace, M. T. (2006). Enhanced multisensory integration in older adults. *Neurobiology of Aging*, *27*(8), 1155–1163.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, *63*(8), 1279–1292.

Lehiste, I., & Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American English. *The Journal of the Acoustical Society of America*, *31*(4), 428–435.

Levin, C. A., & Haber, R. N. (1993). Visual angle as a determinant of perceived interobject distance. *Perception & Psychophysics*, *54*(2), 250–259.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477.

Lewald, J., & Guski, R. (2004). Auditory-visual temporal integration as a function of distance: no compensation for sound-transmission time in human perception. *Neuroscience Letters*, *357*, 119–122.

Lijffijt, M., Lane, S. D., Meier, S. L., Boutros, N. N., Burroughs, S., Steinberg, J. L., . . . Swann, A. C. (2009). P50, n100, and p200 sensory gating: relationships with behavioral inhibition, attention, and working memory. *Psychophysiology*, *46*(5), 1059–1068.

Lockwood, A. H., Salvi, R. J., Coad, M. L., Arnold, S. A., Wack, D. S., Murphy, B., & Burkard, R. F. (1999). The functional anatomy of the normal human auditory system: responses to 0.5 and 4.0 khz tones at varied intensities. *Cerebral Cortex*, *9*(1), 65–76.

Loomis, J. M., Da Silva, J. A., Fujita, N., & Fukusima, S. S. (1992). Visual space perception and visually directed action. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(4), 906.

Loomis, J. M., Klatzky, R. L., Philbeck, J. W., & Golledge, R. G. (1998). Assessing auditory distance perception using perceptually directed action. *Perception & Psychophysics*, *60*(6), 966–980.

Loomis, J. M., & Philbeck, J. W. (1999). Is the anisotropy of perceived 3-d shape invariant across scale? *Perception & Psychophysics*, *61*(3), 397–402.

Luck, S. J. (2005). *An introduction to the event-related potential technique.* Massachusetts, USA: MIT Press.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157.

Lütkenhöner, B., & Klein, J.-S. (2007). Auditory evoked field at threshold. *Hearing Research*, *228*(1-2), 188–200.

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*(3), e4638.

Marks, L. E. (1978). Binaural summation of the loudness of pure tones. *The Journal of the Acoustical Society of America*, *64*(1), 107–113.

Marks, L. E. (1979). Sensory and cognitive factors in judgments of loudness. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(3), 426.

Marks, L. E., & Florentine, M. (2011). Measurement of loudness, part I: Methods, problems, and pitfalls. In M. Florentine, A. N. Popper, & R. R. Fay (Eds.), *Loudness* (pp. 17–56). Springer.

Matthews, N., Todd, J., Mannion, D. J., Finnigan, S., Catts, S., & Michie, P. T. (2013). Impaired processing of binaural temporal cues to auditory scene analysis in schizophrenia. *Schizophrenia Research*, *146*(1-3), 344–348.

May, B. J., Little, N., & Saylor, S. (2009). Loudness perception in the domestic cat: reaction time estimates of equal loudness contours and recruitment effects. *Journal of the Association for Research in Otolaryngology*, *10*(2), 295–308.

Mayhew, J. E., & Frisby, J. P. (1981). Psychophysical and computational studies towards a theory of human stereopsis. *Artificial Intelligence*, *17*(1-3), 349–385.

McCann, B. C., Hayhoe, M. M., & Geisler, W. S. (2018). Contributions of monocular and binocular cues to distance discrimination in natural scenes. *Journal of Vision*, *18*(4), 12–12.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746.

Melara, R. D., & Marks, L. E. (1990). Interaction among auditory dimensions: Timbre, pitch, and loudness. *Perception & psychophysics*, *48*(2), 169–178.

Mendel, M. I., Sussman, H. M., Merson, R. M., Naeser, M. A., & Minifie, F. D. (1969). Loudness judgments of speech and nonspeech stimuli. *The Journal of the Acoustical Society of America*, *46*(6B), 1556–1561.

Mershon, D. H., Desaulniers, D. H., Amerson, T. L., & Kiefer, S. A. (1980). Visual capture in auditory distance perception: Proximity image effect reconsidered. *Journal of Auditory Research*, *20*(2), 129–136.

Mershon, D. H., Desaulniers, D. H., Kiefer, S. A., Amerson Jr, T. L., & Mills, J. T. (1981). Perceived loudness and visually-determined auditory distance. *Perception*, *10*(5), 531–543.

Mershon, D. H., & King, L. E. (1975). Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception & Psychophysics*, *18*(6), 409–415.

Miskolczy-Fodor, F. (1959). Relation between loudness and duration of tonal pulses. I. Response of normal ears to pure tones longer than click-pitch threshold. *The Journal of the Acoustical Society of America*, *31*(8), 1128–1134.

Mohr, C. M., King, W. M., Freeman, A. J., Briggs, R. W., & Leonard, C. M. (1999). Influence of speech stimuli intensity on the activation of auditory cortex investigated with functional magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *105*(5), 2738–2745.

Mohrmann, K. (1939). Lautheitskonstanz im entfernungswechsel. *Zeitschrift fur Psychologie*, *145*(1), 146-199.

Moore, B. C. (2012). *An introduction to the psychology of hearing.* Cambridge, UK: Emerald.

Moore, B. C. (2014). Development and current status of the "Cambridge" loudness models. *Trends in Hearing*, *18*, 1–29.

Moore, B. C., & Glasberg, B. R. (1996). A revision of Zwicker's loudness model. *Acta Acustica United with Acustica*, *82*(2), 335–345.

Moore, B. C., & Glasberg, B. R. (2004). A revised model of loudness perception applied to cochlear hearing loss. *Hearing Research*, *188*(1-2), 70–88.

Moore, B. C., & Glasberg, B. R. (2007). Modeling binaural loudness. *The Journal of the Acoustical Society of America*, *121*(3), 1604–1612.

Moore, B. C., Glasberg, B. R., & Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, *45*(4), 224–240.

Moore, B. C., Glasberg, B. R., & Stone, M. A. (2010). Development of a new method for deriving initial fittings for hearing aids with multi-channel compression: Cameq2-hf. *International Journal of Audiology*, *49*(3), 216–227.

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

Mulert, C., Jäger, L., Propp, S., Karch, S., Störmann, S., Pogarell, O., ... Hegerl, U. (2005). Sound level dependence of the primary auditory cortex: Simultaneous measurement with 61-channel EEG and fMRI. *Neuroimage*, *28*(1), 49–58.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133–137.

Myers, A. K., Cotton, B., & Hilp, H. A. (1981). Matching the rate of concurrent tone bursts and light flashes as a function of flash surround luminance. *Perception & Psychophysics*, *30*(1), 33–38.

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology*, *24*(4), 375–425.

Norman-Haignere, S., & McDermott, J. (2018, September 16). Learned object-specific invariances are revealed by the effect of intensity on environmental sound recognition. *PsyArXiv*. Retrieved from `https://doi.org/10.31234/osf.io/x7d82`

O'Connor, N., & Hermelin, B. (1972). Seeing and hearing and space and space and time. *Perception & Psychophysics*, *11*(1), 46–48.

Oestreich, L. K., Mifsud, N. G., Ford, J. M., Roach, B. J., Mathalon, D. H., & Whitford, T. J. (2015). Subnormal sensory attenuation to self-generated speech in schizotypy: electrophysiological evidence for a 'continuum of psychosis'. *International Journal of Psychophysiology*, *97*(2), 131–138.

Ozimek, E., & Zwislocki, J. J. (1996). Relationships of intensity discrimination to sensation and loudness levels: Dependence on sound frequency. *The Journal of the Acoustical Society of America*, *100*(5), 3304–3320.

Paiva, T. O., Almeida, P. R., Ferreira-Santos, F., Vieira, J. B., Silveira, C., Chaves, P. L., ... Marques-Teixeira, J. (2016). Similar sound intensity dependence of the n1 and p2 components of the auditory erp: Averaged and single trial evidence. *Clinical Neurophysiology*, *127*(1), 499–508.

Pantev, C., Bertrand, O., Eulitz, C., Verkindt, C., Hampson, S., Schuierer, G., & Elbert, T. (1995). Specific tonotopic organizations of different areas of the human auditory cortex revealed by simultaneous magnetic and electric recordings. *Electroencephalography and Clinical Neurophysiology*, *94*(1), 26–40.

Pantev, C., Hoke, M., Lehnertz, K., & Lütkenhöner, B. (1989). Neuromagnetic evidence of an amplitopic organization of the human auditory cortex. *Electroencephalography and Clinical Neurophysiology*, *72*(3), 225–231.

Patten, M. L., & Clifford, C. W. (2015). A bias-free measure of the tilt illusion. *Journal of Vision*, *15*(8), 1–14.

Peirce, J. W. (2007). PsychoPy–Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13.

Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., & Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. *Neuroreport*, *16*(2), 125–128.

Pfingst, B. E., Hienz, R., Kimm, J., & Miller, J. (1975). Reaction-time procedure for measurement of hearing. I. Suprathreshold functions. *The Journal of the Acoustical Society of America*, *57*(2), 421–430.

Pickles, J. O. (2013). *An introduction to the physiology of hearing.* Leiden, Netherlands: Brill.

Pickles, J. O., & Corey, D. P. (1992). Mechanoelectrical transduction by hair cells. *Trends in Neurosciences*, *15*(7), 254–259.

Pirenne, M. H. (1970). *Optics, painting & photography.*

Plumert, J. M., Kearney, J. K., Cremer, J. F., & Recker, K. (2005). Distance perception in real and virtual environments. *ACM Transactions on Applied Perception (TAP)*, *2*(3), 216–233.

Pollack, I. (1952). On the measurement of the loudness of speech. *The Journal of the Acoustical Society of America*, *24*(3), 323–324.

Powers, A. R., Kelley, M., & Corlett, P. R. (2016). Hallucinations as top-down effects on perception. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 393–400.

Powers, A. R., Mathys, C., & Corlett, P. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596–600.

Rapin, I., Schimmel, H., Tourk, L. M., Krasnegor, N. A., & Pollak, C. (1966). Evoked responses to clicks and tones of varying intensity in waking adults. *Electroencephalography and Clinical Neurophysiology*, *21*(4), 335–344.

Reisberg, D., Mclean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), (pp. 97–113). New Jersey, United States: Lawrence Erlbaum Associates, Inc.

Reite, M., Zimmerman, J. T., Edrich, J., & Zimmerman, J. E. (1982). Auditory evoked magnetic fields: response amplitude vs. stimulus intensity. *Electroencephalography and Clinical Neurophysiology*, *54*(2), 147–152.

Reiterer, S., Erb, M., Grodd, W., & Wildgruber, D. (2008). Cerebral processing of timbre and loudness: fMRI evidence for a contribution of Broca's area to basic auditory discrimination. *Brain Imaging and Behavior*, *2*(1), 1–10.

Renner, R. S., Velichkovsky, B. M., & Helmert, J. R. (2013). The perception of egocentric distances in virtual environments-a review. *ACM Computing Surveys (CSUR)*, *46*(2), 1–40.

Richards, D. G., & Wiley, R. H. (1980). Reverberations and amplitude fluctuations in the propagation of sound in a forest: implications for animal communication. *The American Naturalist*, *115*(3), 381–399.

Richards, W., & Miller, J. F. (1969). Convergence as a cue to depth. *Perception & Psychophysics*, *5*(5), 317–320.

Richardson, L., & Ross, J. (1930). Loudness and telephone current. *The Journal of General Psychology*, *3*(2), 288–306.

Rienstra, S. W., & Hirschberg, A. (2004). An introduction to acoustics. *Eindhoven University of Technology*, *18*, 19.

Riesz, R. (1933). The relationship between loudness and the minimum perceptible increment of intensity. *The Journal of the Acoustical Society of America*, *4*(3), 211–216.

Ritter, M. (1979). Perception of depth: Processing of simple positional disparity as a function of viewing distance. *Perception & Psychophysics*, *25*(3), 209–214.

Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1598), 2159–2168.

Rockland, K. S., & Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology*, *50*(1-2), 19–26.

Rogers, B., & Graham, M. (1979). Motion parallax as an independent cue for depth perception. *Perception*, *8*(2), 125–134.

Rogers, S. P., & Gogel, W. C. (1975). Relation between judged and physical distance in multicue conditions as a function of instructions and tasks. *Perceptual and Motor Skills*, *41*(1), 171–178.

Röhl, M., Kollmeier, B., & Uppenkamp, S. (2011). Spectral loudness summation takes place in the primary auditory cortex. *Human Brain Mapping*, *32*(9), 1483–1496.

Röhl, M., & Uppenkamp, S. (2012). Neural coding of sound intensity and loudness in the human auditory system. *Journal of the Association for Research in Otolaryngology*, *13*(3), 369–379.

Rosenblum, L. D., & Fowler, C. A. (1991). Audiovisual investigation of the loudness-effort effect for speech and nonspeech events. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(4), 976.

Roussel, C., Hughes, G., & Waszak, F. (2014). Action prediction modulates both neurophysiological and psychophysical indices of sensory attenuation. *Frontiers in Human Neuroscience*, *8*, 115.

Rowland, B., Stanford, T., & Stein, B. (2007). A Bayesian model unifies multisensory spatial localization with the physiological properties of the superior colliculus. *Experimental Brain Research*, *180*(1), 153–161.

Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception & Psychophysics*, *54*(3), 406–416.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

SanMiguel, I., Saupe, K., & Schröger, E. (2013). I know what is missing here: electrophysiological prediction error signals elicited by omissions of predicted "what" but not "when". *Frontiers in Human Neuroscience*, *7*, 407.

Sato, A. (2008). Action observation modulates auditory perception of the consequence of others' actions. *Consciousness and Cognition*, *17*(4), 1219–1227.

Sato, Y., Toyoizumi, T., & Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. *Neural Computation*, *19*(12), 3335–3355.

Satongar, D., Lam, Y., & Pike, C. (2014). Measurement and analysis of a spatially sampled binaural room impulse response dataset. In (Vol. 2, p. 1775-1782). International Institute of Acoustics and Vibrations. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84922648536&partnerID=40&md5=b89a3c58870e69e4de17662313e5c26e (cited By 3)

Schafer, E. W., & Marcus, M. M. (1973). Self-stimulation alters human sensory brain responses. *Science*, *181*(4095), 175–177.

Scharf, B. (1961). Complex sounds and critical bands. *Psychological Bulletin*, *58*(3), 205.

Scharf, B. (1978). Loudness. *Handbook of Perception*, *4*, 187–242.

Scharf, B., & Fishken, D. (1970). Binaural summation of loudness: Reconsidered. *Journal of Experimental Psychology*, *86*(3), 374.

Schmack, K., Rothkirch, M., Priller, J., & Sterzer, P. (2017). Enhanced predictive signalling in schizophrenia. *Human Brain Mapping*, *38*(4), 1767–1779.

Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1791.

Senkowski, D., Saint-Amour, D., Kelly, S. P., & Foxe, J. J. (2007). Multisensory processing of naturalistic objects in motion: a high-density electrical mapping and source estimation study. *Neuroimage*, *36*(3), 877–888.

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences*, *14*(9), 425–432.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature*, *408*(6814), 788.

Shams, L., Ma, W. J., & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport*, *16*(17), 1923–1927.

Shigenaga, S. (1965). The constancy of loudness and of acoustic distance. *Bulletin of the Faculty of Literature of Kyushu University*, *9*, 289–333.

Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science*, *145*(3638), 1328–1330.

Sigalovsky, I. S., & Melcher, J. R. (2006). Effects of sound level on fMRI activation in human brainstem, thalamic and cortical centers. *Hearing Research*, *215*(1-2), 67–76.

Silva, I., & Florentine, M. (2006). Effect of adaptive psychophysical procedure on loudness matches. *The Journal of the Acoustical Society of America*, *120*(4), 2124–2131.

Sinai, M. J., Ooi, T. L., & He, Z. J. (1998). Terrain influences the accurate judgement of distance. *Nature*, *395*(6701), 497–500.

Soeta, Y., & Nakagawa, S. (2009). Sound level-dependent growth of N1m amplitude with low and high-frequency tones. *Neuroreport*, *20*(6), 548–552.

Stebbins, W. C., & Miller, J. M. (1964). Reaction time as a function of stimulus intensity for the monkey 1. *Journal of the Experimental Analysis of Behavior*, *7*(4), 309–312.

Stekelenburg, J. J., Maes, J. P., Van Gool, A. R., Sitskoorn, M., & Vroomen, J. (2013). Deficient multisensory integration in schizophrenia: an event-related potential study. *Schizophrenia Research*, *147*(2-3), 253–261.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*(12), 1964–1973.

Stekelenburg, J. J., & Vroomen, J. (2012). Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events. *Frontiers in Integrative Neuroscience*, *6*, 26.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., ... Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry*, *84*(9), 634–643.

Sterzer, P., Mishara, A. L., Voss, M., & Heinz, A. (2016). Thought insertion as a self-disturbance: an integration of predictive coding and phenomenological approaches. *Frontiers in Human Neuroscience*, *10*, 502.

Stevens, S. S. (1955). The measurement of loudness. *The Journal of the Acoustical Society of America*, *27*(5), 815–829.

Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *The American Journal of Psychology*, *69*(1), 1–25.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural and social prospects*. NY, USA: Routledge.

Stevens, S. S., & Greenbaum, H. B. (1966). Regression effect in psychophysical judgment. *Perception & Psychophysics*, *1*(5), 439–446.

Stevens, S. S., & Poulton, E. C. (1956). The estimation of loudness by unpracticed observers. *Journal of Experimental Psychology*, *51*(1), 71.

Strainer, J. C., Ulmer, J. L., Yetkin, F. Z., Haughton, V. M., Daniels, D. L., & Millen, S. J. (1997). Functional MR of the primary auditory cortex: an analysis of pure tone activation and tone discrimination. *American Journal of Neuroradiology*, *18*(4), 601–610.

Sugita, Y., & Suzuki, Y. (2003). Audiovisual perception: Implicit estimation of sound-arrival time. *Nature*, *421*, 911.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise.

The Journal of the Acoustical Society of America, *26*(2), 212–215.

Surdick, R. T., Davis, E. T., King, R. A., & Hodges, L. F. (1997). The perception of distance in simulated visual displays: A comparison of the effectiveness and accuracy of multiple depth cues across viewing distances. *Presence: Teleoperators & Virtual Environments*, *6*(5), 513–531.

Suzuki, Y., & Takeshima, H. (2004). Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America*, *116*(2), 918–933.

Svarverud, E., Gilson, S. J., & Glennerster, A. (2010). Cue combination for 3D location judgements. *Journal of Vision*, *10*(1), 5–5.

Takeshima, H., Suzuki, Y., Fujii, H., Kumagai, M., Ashihara, K., Fujimori, T., & Sone, T. (2001). Equal-loudness contours measured by the randomized maximum likelihood sequential procedure. *Acta Acustica United with Acustica*, *87*(3), 389–399.

Thaerig, S., Behne, N., Schadow, J., Lenz, D., Scheich, H., Brechmann, A., & Herrmann, C. S. (2008). Sound level dependence of auditory evoked potentials: simultaneous EEG recording and low-noise fMRI. *International Journal of Psychophysiology*, *67*(3), 235–241.

Thomas, S. M., & Jordan, T. R. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(5), 873.

Thompson, W. B., Willemsen, P., Gooch, A. A., Creem-Regehr, S. H., Loomis, J. M., & Beall, A. C. (2004). Does the quality of the computer graphics matter when judging distances in visually immersive environments? *Presence: Teleoperators & Virtual Environments*, *13*(5), 560–571.

Thurlow, W. R., & Jack, C. E. (1973). Certain determinants of the "ventriloquism effect". *Perceptual and Motor Skills*, *36*(3_suppl), 1171–1184.

Todd, J. T., & Akerstrom, R. A. (1987). Perception of three-dimensional form from patterns of optical texture. *Journal of Experimental Psychology: Human perception and performance*, *13*(2), 242.

Todorović, D. (2008). Is pictorial perception robust? the effect of the observer vantage point on the perceived depth structure of linear-perspective images. *Perception*, *37*(1), 106–125.

Todorović, D. (2009). The effect of the observer vantage point on perceived distortions in linear perspective images. *Perception & Psychophysics*, *71*(1), 183–193.

Traer, J., & McDermott, J. H. (2016). Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National academy of*

Sciences of the United States of America, *113*, E7856–E7865.

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, *107*(6), 3438–3451.

Tremblay, K. L., Piskosz, M., & Souza, P. (2003). Effects of age and age-related hearing loss on the neural representation of speech cues. *Clinical Neurophysiology*, *114*(7), 1332–1343.

Tumarkin, A. (1968). Evolution of the auditory conducting apparatus in terrestrial vertebrates. In *Ciba foundation symposium-hearing mechanisms in vertebrates* (pp. 18–40).

Van Boxtel, G. J., & Böcker, K. B. (2004). Cortical measures of anticipation. *Journal of Psychophysiology*, *18*(2/3), 61–76.

van Laarhoven, T., Stekelenburg, J. J., & Vroomen, J. (2017). Temporal and identity prediction in visual-auditory events: Electrophysiological evidence from stimulus omissions. *Brain Research*, *1661*, 79–87.

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*, *102*(4), 1181–1186.

Vasama, J., Mäkelä, J. P., Tissari, S. O., & Hämäläinen, M. S. (1995). Effects of intensity variation on human auditory evoked magnetic fields. *Acta Oto-laryngologica*, *115*(5), 616–621.

Viemeister, N. F., & Plack, C. J. (1993). Time analysis. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Human psychophysics* (pp. 116–154). New York, USA: Springer.

Vishwanath, D., Girshick, A. R., & Banks, M. S. (2005). Why pictures look right when viewed from the wrong place. *Nature neuroscience*, *8*(10), 1401–1410.

von Fieandt, K. (1951). Loudness invariance in sound perception. *Acta Psychologica Fennica*, *1*, 9–20.

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*, *22*(7), 1583–1596.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Wagner, E., Florentine, M., Buus, S., & McCormack, J. (2004). Spectral loudness summation and simple reaction time. *The Journal of the Acoustical Society of*

America, *116*(3), 1681–1686.

Walsh, V., & Kulikowski, J. (1998). *Perceptual constancy: Why things look as they do.* New York, USA: Cambridge University Press.

Warren, R. M. (1981). Measurement of sensory intensity. *Behavioral and Brain Sciences*, *4*(2), 175–189.

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

Westman, J. C., & Walters, J. R. (1981). Noise and stress: a comprehensive approach. *Environmental Health Perspectives*, *41*, 291–309.

White, P. M., & Yee, C. M. (2006). P50 sensitivity to physical and psychological state influences. *Psychophysiology*, *43*(3), 320–328.

Woods, D. L., Herron, T. J., Cate, A. D., Yund, E. W., Stecker, G. C., Rinne, T., & Kang, X. (2010). Functional properties of human auditory cortical fields. *Frontiers in Systems Neuroscience*, *4*, 155.

Worden, F. G. (1971). Hearing and the neural detection of acoustic patterns. *Behavioral Science*, *16*(1), 20–30.

Wright, A., Davis, A., Bredberg, G., Ülehlová, L., & Spencer, H. (1987). Hair cell distributions in the normal human cochlea: a report of a European working group. *Acta Oto-Laryngologica*, *104*(sup436), 15–24.

Wu, B., He, Z. J., & Ooi, T. L. (2007). Inaccurate representation of the ground surface beyond a texture boundary. *Perception*, *36*(5), 703–721.

Wu, B., Ooi, T. L., & He, Z. J. (2004). Perceiving distance accurately by a directional process of integrating ground information. *Nature*, *428*(6978), 73–77.

Yang, T., & Kubovy, M. (1999). Weakening the robustness of perspective: Evidence for a modified theory of compensation in picture perception. *Perception & Psychophysics*, *61*(3), 456–467.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, *30*(3), 555–568.

Yeomans, J. S., & Frankland, P. W. (1995). The acoustic startle reflex: neurons and connections. *Brain Research Reviews*, *21*(3), 301–314.

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Zahorik, P. (2002a). Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, *111*(4), 1832–1846.

Zahorik, P. (2002b). Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America*, *112*(5), 2110–2117.

Zahorik, P. (2009). Perceptually relevant parameters for virtual listening simulation of small room acoustics. *The Journal of the Acoustical Society of America*, *126*(2), 776–791.

Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *ACTA Acustica united with Acustica*, *91*(3), 409–420.

Zahorik, P., & Wightman, F. L. (2001). Loudness constancy with varying sound source distance. *Nature Neuroscience*, *4*(1), 78–83.

Zimmer, K. (2005). Examining the validity of numerical ratios in loudness fractionation. *Perception & Psychophysics*, *67*(4), 569–579.

Zouridakis, G., Simos, P. G., & Papanicolaou, A. C. (1998). Multiple bilaterally asymmetric cortical sources account for the auditory N1m component. *Brain Topography*, *10*(3), 183–189.

Zwicker, E., Flottorp, G., & Stevens, S. S. (1957). Critical band width in loudness summation. *The Journal of the Acoustical Society of America*, *29*(5), 548–557.

Zwicker, E., & Zwicker, U. T. (1991). Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system. *Journal of the Audio Engineering Society*, *39*(3), 115–126.