Quality-controlled audio-visual depth in stereoscopic 3D media

A thesis presented for the degree of
Doctor of Philosophy

Jonathan Stuart Berry

School of Engineering and Computing Sciences
Durham University
United Kingdom
October 19, 2015

Abstract

BACKGROUND: The literature proposes several algorithms that produce "quality-controlled" stereoscopic depth in 3D films by limiting the stereoscopic depth to a defined depth budget. Like stereoscopic displays, spatial sound systems provide the listener with enhanced (auditory) depth cues, and are now commercially available in multiple forms.

AIM: We investigate the implications of introducing auditory depth cues to quality-controlled 3D media, by asking: "Is it important to quality-control audio-visual depth by considering audio-visual interactions, when integrating stereoscopic display and spatial sound systems?"

MOTIVATION: There are several reports in literature of such "audio-visual interactions", in which visual and auditory perception influence each other. We seek to answer our research question by investigating whether these audio-visual interactions could extend the depth budget used in quality-controlled 3D media.

METHOD/CONCLUSIONS: The related literature is reviewed before presenting four novel experiments that build upon each other's conclusions. In the first experiment, we show that content created with a stereoscopic depth budget creates measurable positive changes in audiences' attitude towards 3D films. These changes are repeatable for different locations, displays and content. In the second experiment we calibrate an audio-visual display system and use it to measure the minimum audible depth difference. Our data is used to formulate recommendations for content designers and systems engineers. These recommendations include the design of an auditory depth perception screening test. We then show that an auditory-visual stimulus with a nearer auditory depth is perceived as nearer. We measure the impact of this effect upon a relative depth judgement, and investigate how the impact varies with audio-visual depth separation. Finally, the size of the cross-modal bias in depth is measured, from which we conclude that sound does have the potential to extend the depth budget by a small, but perceivable, amount.

Declaration

The work in this thesis is based on research carried out in the Innovative Computing Group, the School of Engineering and Computing Sciences, University of Durham, UK. No part of this thesis has been submitted elsewhere for any other degree or qualification and it all my own work unless referenced to the contrary in the text.

Part of the work presented in this thesis has been documented in the following publications:

- J. S. Berry, D. A. T. Roberts, and N. S. Holliman. 3D sound and 3D image interactions: a review of audio-visual depth perception. In *Proceedings of Human Vision and Electronic Imaging XIX* - SPIE volume 9014, 2014.
- J. Berry, D. Budgen, and N. Holliman. Evaluating subjective impressions of quality controlled 3D films on large and small screens. *Journal of Display Technology*, 2015.

Prior to starting work on this thesis, the author conducted a particularly relevant preliminary study that is documented in the following publication:

- A. Turner, J. S. Berry, and N. Holliman. Can the perception of depth in stereoscopic images be influenced by 3D sound? In *Proceedings of Stereoscopic Displays and Virtual Reality Systems XXII* - SPIE Volume 7863, 2011.

"To God be the glory, great things he has done.

So loved he the world that he gave us his Son,
who yielded his life an atonement for sin
and opened the life gates that all may go in."

F. J. Crosby (1820-1915)

Acknowledgements

First and foremost, a special thanks should go to Prof. Nick Holliman for supervising the start of my PhD and remaining a close colleague since moving on from Durham University. Nick has been a constant source of encouragement, technical advice and professional support throughout my time as a research student at Durham University. I am also immensely grateful to Prof. David Budgen who took over supervision of my PhD after Nick's departure. David's input concerning the assembly and presentation of a clear narrative through this Thesis have been invaluable.

On a personal note, I would like to thank my wife, Janet Berry, for the tremendous support she has offered throughout the PhD. She has carried me through many tough times during the past four years. I would also like to thank my parents, Dr David Berry and Mrs Carol Berry. I "blame" them for instilling in me a deep desire to complete a PhD and explore the boundaries of human knowledge, having watched them both undertake and discuss their own academic research.

I have also received a significant amount of support from a wider group of friends. I'd like to thank members of the postgraduate house groups at St Nicholas' Church (Durham), as well as staff and students of St John's College (Durham), for their prayers and support. It's important to thank the "bois" of *Durham University Big Band*, especially those close friends who also play in *The Invitations*, for offering me space away from the PhD to play music and thus regather my sanity. I am also particularly grateful to a number of friends for proof reading parts of this thesis: Martin Dhenel, Edmund Waddelove, Clare Bliss, Ant Cooper, and Andrew Duckworth.

On a very practical level, St John's College provided partial funding for my travel to the *SPIE Electronic Imaging Symposium* in California, for which I am very grateful. It is also important that I acknowledge the work of Tommasso Selvetti at K-Array (Italy) who spent time selecting a pair of their KT-20 loudspeakers with well matched frequency response curves for use in our experiments. Finally, I acknowledge and thank the EPSRC for funding the PhD.

Abbreviations

2AFC two alternative forced choice.

3DTV stereoscopic three-dimensional television.

ANOVA analysis of variance.

BRIR binaural room impulse response.

BSHAA British Society of Hearing Aid Audiologists.

DLP digital light processing.

HCI human-computer interaction.

HRTF head related transfer function.

ILD interaural level difference.

ITD interaural time difference.

JOGL Java open graphics library.

MAD minimum audible depth.

MLE maximum likelihood estimation.

PDH pressure discrimination hypothesis.

PEST parameter estimation by sequential testing.

RAD relative auditory depth.

S3D stereoscopic three-dimensional.

SMPTE Society of Motion Picture and Television Engineers.

WFS wave field synthesis.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

"The danger with stereoscopic film-making is that if it is improperly done, the result can be discomfort. Yet, when properly executed, stereoscopic films are beautiful and easy on the eyes." So writes the film maker Lenny Lipton in his book, *The Foundations of Stereoscopic Cinema* (Lipton 1982). A whole body of scientific research has arisen around this desire to produce stereoscopic media "properly" and "quality-control" the stereoscopic depth cue. There is a smaller, but analogous, literature base concerning the development of good quality 3D spatial sound systems and content. Recent steps towards the commercialisation of 3D spatial sound systems have turned attention towards the use of stereoscopic media as a showcase for the technology (André et al. 2010; Evrard et al. 2011; Kuhlen et al. 2007; Rebillat et al. 2010; Springer et al. 2006). However, we should be hesitant to assume audio and visual perception can be treated as mutually exclusive entities. Studies from Psychology reveal a number of interactions that occur between audio and visual perception that could offer new ways of enhancing audio-visual media. This thesis reconciles this body of literature with the needs and interests of spatial sound and stereoscopic display systems engineers and content designers. This is done through novel experimentation that explores the quality-control of audio-visual depth when integrating both technologies.

1.1 Background summary

Humans view the world through two eyes. These eyes are in different positions, meaning the brain receives two images of the same scene, each using a different projection. The differences between each eye's image of a scene depends on the depth, or distance, of the scene's content. Objects that are nearby will appear at very different positions in each image, whereas objects that are far away will appear

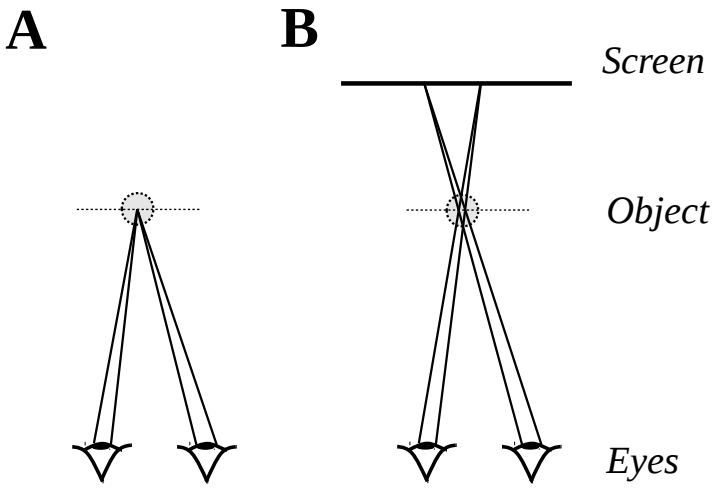

Figure 1.1: The conflict between vergence and accommodation for S3D displays. **A:** Vergence and accommodation in the real world - both focus and vergence meet at a single point in space. **B:** Vergence and accommodation when viewing S3D displays - vergence conflicts with focus, which remains on the screen.

at almost the same position in each image. By a process called *stereopsis*, the brain fuses these two images into a single image together with added depth sensation. This cue to the visual depth of an object, arising from the disparity between each eye's image, is called the *binocular cue*.

Stereoscopic three-dimensional (S3D) displays offer enhanced depth perception in images by conveying binocular depth cues to the viewer. At the heart of every S3D display is a mechanism that displays a different image to the left and right eyes. If the images shown by the display to the left and right eye consist of projections that are related to each eye's view of the natural world, then the brain will fuse the display's images into a single "3D" image. However, S3D displays do not offer a perfect replication of real-world viewing. For instance, when viewing an S3D display, the eyes focus upon the depth of the screen, but *verge* upon the content's binocular depth (*vergence* is the degree to which the viewing direction of the eyes simultaneously "toe-in" so as to intersect at the depth and position of the object of attention). This phenomenon is shown in Figure 1.1. It is called the vergence-accommodation conflict and can be uncomfortable because it is unnatural; in the real world we focus and verge upon the same point. This is one example of several human and technological factors that should be properly considered if S3D displays are to provide an enjoyable and comfortable viewing experience.

In this thesis, the display *depth* of a stimulus is defined as the component of its position along the axis perpendicular to the display screen. Depth is closely related

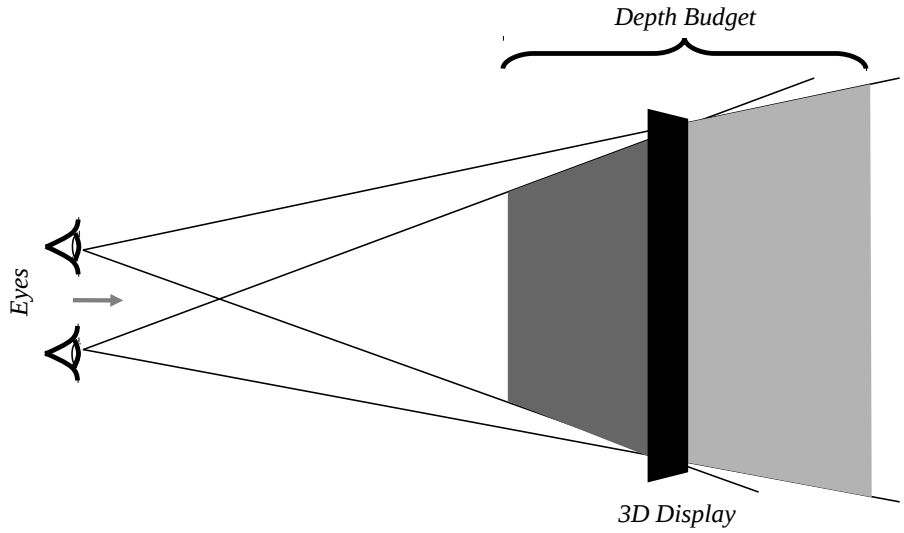

Figure 1.2: The inevitable depth budget in a S3D display that arises due to both human and technological factors.

to ego-centric distance if participants are placed on the axis that is perpendicular to, and centred upon, the screen. Therefore, we generally refer to an *ego-centric* concept of depth – where a nearer or a further depth refers respectively to a difference directed towards or away from the viewer. We do sometimes refer to a *screen-centric* concept of depth – where a greater or smaller depth refers to a position further from, or nearer to, the screen. The distinction between these two concepts of depth should be clear in the text.

The range of binocular depth that is comfortable to view on a given S3D display is called the "zone of comfort" (Shibata et al. 2011) or the "depth bracket" (Block and McNally 2013) or the "depth budget" (Holliman 2010), which is the term we use in this thesis. Its value needs to factor in the immense variation between humans in their ability to fuse comfortably S3D images. The smaller the range of depth used, the larger the set of people who will be able to view the content comfortably. This depth budget can be very restrictive and limiting, particularly when viewing small displays from close viewing distances. Hence, there would be significant benefit in finding a means of extending the range of depth a display can show in a manner that is comfortable to view.

In a similar way to vision, we are able to localise sound sources, due partly to the positioning of our ears – positioned within pinna (the externally visible part of the ear) on the left and right side of the head. However, for depth perception of sources directly in front of the head, the inter-ear (or inter-aural) differences have

a smaller role when compared with other aural cues including: loudness (pressure), reverberation and frequency spectrum differences. Loudness at the ear is most commonly considered to indicate auditory distance, but as it can be confused with loudness at the source, it is only really valuable if you have some prior knowledge of the source's volume. In this way it is analagous to familiar size in human vision. Reverberation, or more specifically the ratio of direct sound to reverberant sound, is able to provide an absolute cue to a source's depth without needing to have prior knowledge of the source.

Neither stereo nor surround sound provide a cue to a source's exact location in 3D space. Stereo systems present sound in the same horizontal and frontal plane, giving no sense of depth or elevation to the sound sources. Surround sound improves on stereo, but still presents sound in the same horizontal plane with very limited options for depth (in front/behind). 3D sound systems aim to create sound sources at all required locations in a continuous 3D space for a given set of soundscapes. These are used in simulation, film and gaming, creating a new level of realism and opening a new field of action-tasks and cues available to content designers (Cater et al. 2007).

There are various different approaches to building 3D sound systems. One approach, recently taken by Dolby (Sergi 2013), is to use large arrays of loudspeakers to improve the spatial fidelity of the sound fields that can be reproduced (although this can only ever be an approximation of true 3D sound). Another approach is analogous to the principle that lies behind S3D displays, in that the left and right ear channels are split, for example using headphones, and each ear is fed a processed audio stream with localisation cues including inter-aural differences. The simplest way to capture such audio-streams is to record using two microphones placed inside the pinna of a model (or real) head.

The brain is therefore required to combine an array of visual and auditory cues when forming a *cross-modal* perception of the world around us. If these cues conflict, the brain is left to assemble the best perception it can, and the result isn't always successful. A number of interesting illusions have been been observed when audition and vision appear to be in conflict with each other (e.g. McGurk and MacDonald 1976; Sekuler et al. 1997; Shams et al. 2002). Perhaps the most well known illusion is the *ventriloquist effect*, in which the apparent position of a sound is typically biased towards a different location than the actual source, because it appears a more visually rational source for the sound. In this situation, cues from the auditory and visual scenes conflict, resulting in the brain wrongly interpreting the auditory cues. The ventriloquist effect has been inverted under certain circumstances, resulting

in the brain interpreting the visual cues wrongly (Burr and Alais 2006; Phan et al. 2000). Audio-visual interactions have already been used to enhance conventional 2D media. One such example is taken from the film *Star Wars IV: A New Hope*, where a "swoosh" sound creates the perception of a door sliding open when in fact the door instantaneously disappears (Chion and Gorbman 1994). Furthermore, we all experience the ventriloquist effect when viewing conventional 2D media, as sounds appear to emanate from visual sources on the screen instead of the loudspeakers, which can be in a very different position.

There has been substantial research concerning content design for S3D displays, and to a lesser extent, spatial sound systems. This research sits firmly within the human-computer interaction (HCI) field of computer science, as it aims to improve the user experience offered by the technology. Researchers therefore need to reconcile work in psychology, specifically concerning the human visual and auditory systems, with the engineering and computing technologies involved in building displays and display content. The outcome of such research includes algorithms, design principles and guidance concerning the production of a quality user experience. When we refer to "quality-controlled depth", we refer to depth cues controlled in a manner that is informed by such research. For example, Jones et al. (2001) outline algorithms that control the binocular depth cue in a way that is informed by research concerning the depth budget or "zone of comfort".

It is important, at this early stage, to make the distinction between *relative* and *absolute* perception. Relative perception is concerned with distinguishing differences between multiple stimuli, whereas absolute perception requires the making of a singular judgement without any references. The two are closely linked, though in the natural environment tasks involving relative spatial perception are far more prevalent than tasks involving absolute spatial perception. Auditory depth cues such the loudness and frequency spectrum require prior knowledge of the source, and are therefore far better relative cues than absolute cues. How "flat" a 3D visual or auditory display appears is directly dependent upon our ability to perceive relative depth differences between content stimuli and the display.

1.2 Research questions

In the light of developments aimed at the commercialisation of spatial sound, interest has turned to the use of S3D media as a showcase for the technology (André et al. 2010; Evrard et al. 2011; Kuhlen et al. 2007; Rebillat et al. 2010; Springer et al. 2006). Integrating spatial sound and S3D displays may offer new ways of using the

audio-visual interaction. For example, a preliminary experiment executed by the author, and published prior to beginning the work outlined here (Turner et al. 2011), concluded that auditory depth can have an influence upon judgements of relative visual depth in an S3D display. There is the potential, therefore, for extending an S3D display's depth budget without using more binocular disparity. This was the starting point for the trail of research that is presented here, and which can be summarised by the following over-arching research question:

Is it important to quality-control audio-visual depth by considering audio-visual interactions when integrating S3D display and spatial sound systems?

There are three significant parts to this research question. The first part is concerned with the importance of quality-controlling audio-visual depth (the potential subjective and/or commercial value of quality-controlling audio-visual depth). The second is related to the impact of audio-visual interactions upon the quality-control of audio-visual depth. The final part addresses the application context of integrating S3D displays and spatial sound systems.

The approach we have taken to answer this research question begins with an experimental study that shows the importance of quality-controlling the binocular depth cue alone, by restricting it to particular depth budget. This study serves the purpose of motivating our later work, which explores the possibility of extending the depth budget using audio depth. Furthermore, it provides evidence that quality-controlling depth cues is a valuable focus for research. The other experiments we then report, address each part of the over-arching research question in reverse order:

- The third part was addressed by building a calibrated audio and visual display system for which we measured the minimum audible depth (MAD) we should expect participants to perceive (the minimum visible depth could be taken from literature).
- We then used the calibrated experimental setup, and the measurement of the setup's MAD, to design and run an experiment that would observe the impact of an audio-visual interaction upon a relative depth perception task. This addresses the second part of the over-arching research question.
- Finally we measure the size of the audio-visual bias (the impact of which we observed whilst addressing the second part of the over-arching research question) in order to discuss the cross-modal effect's potential significance and application to S3D media, thereby addressing the first part of the over-arching research question.

Our approach was broken into the following subsidiary research questions:

1. **What is there to be learnt from the literature?**

   There is a wealth of literature concerning visual depth perception in S3D displays and the engineering of display systems. There is notably less material concerning relative auditory depth (RAD) perception, particularly in application scenarios, and still less concerning audio-visual depth perception. We address this research question by first reviewing literature related to the background of the over-arching research question. This includes the psychology and physiology of the human visual and auditory system, as well as the engineering of S3D displays and spatial sound systems. We then broadly look at literature related to cross-modal interactions, particularly focussing upon audio-visual interactions in depth perception. Finally, we review in detail the preliminary experiment undertaken by the author, as it marks the beginning of our experimental research trail. The act of collating and reviewing literature concerning audio-visual depth perception is the first novel contribution of this thesis. A summary of this work was published by the author in the proceedings of the *Electronic Imaging* symposium run by *SPIE: The International Society for Optics and Photonics* (Berry et al. 2014).

2. **Does viewing S3D content with quality-controlled binocular cues create measurable positive changes in the audience's subjective attitudes towards S3D media?**

   We began our experimental research by assessing the importance of quality-controlling visual depth cues. The rationale behind this is partly explained above; by our approach to tackle the over-arching research question. A positive answer to this research question would suggest that it is important to quality-control the binocular depth cue. By extension, this suggests it could be important to quality-control other depth cues, particularly if there is a perceptual interaction between them and the binocular cue. For the purposes of answering this question, we have used media that quality-controls the binocular cue by implementing the research by Jones et al. (2001). This research proposes a mapping of binocular depth in the scene to a given S3D display depth budget. As already mentioned, the extension of this depth budget is one possible outcome of the effect observed in the preliminary experiment. Therefore, it seemed sensible to evaluate the use of a visual depth budget before exploring a possible way of extending it. We measured the changes in subjective attitudes by asking participants to rate their responses to five different

questions on a scale of 0-100. We then looked for statistically significant differences in their answers before and after viewing high quality S3D content with quality-controlled binocular cues. This evaluation of subjective impressions, specifically of high-quality S3D media with quality-controlled binocular cues, is a novel contribution of this thesis and was published in the *IEEE Journal of Display Technology* (Berry et al. 2015).

3. **What is the MAD in our experimental setup?**

   A crucial factor in determining whether we would expect to observe an audio-visual effect was whether we would expect to perceive the corresponding audio depth cues. Very little was known about the loudspeakers used in the preliminary experiment, or how their positioning affected RAD perception performance. Before further cross-modal experimentation could be undertaken, it was therefore important to source a calibrated experimental setup. This began with finding small loudspeakers that reduced occlusion and interference whilst having well matched frequency response curves. The loudspeakers were then positioned using motorised rails controlled by a computer, so that the listener was unable to use anything other than auditory depth cues to distinguish consistently between them. Finally, we measured the MAD that participants can perceive in our calibrated experimental setup. This sensory threshold and calibrated experimental setup directly informed the design of our later cross-modal experimentation. Both the measurement of the MAD within a TV viewing scenario and the evaluation of its implications for engineers and content designers, are novel contributions of this thesis.

4. **Does auditory depth influence our perception of relative depth in S3D images?**

   The preliminary experiment suggested that audio depth can influence our perception of depth in S3D images. However, the preliminary experiment lacked a calibrated experimental setup and there were weaknesses in the experimental design. The preliminary experiment's results are therefore used only to motivate our work, and not to answer our over-arching research question. We ran a new experiment, similar in design to the preliminary experiment, that gave us a more robust answer to this research question. The new experimental design included a more effective method of capturing qualitative data to support our results and the experimental setup had been refined substantially whilst addressing the previous research question. The new setup included loudspeakers with frequency response matched curves positioned within a calibrated arrange-

ment for which we know the MAD. The new experimental design also included the implementation of a screening test for RAD perception. The design of this screening test was based upon our previous measurements of the MAD. The implementation of the screening test and the evaluation of its results, along with the other improvements reported here, are novel contributions of this thesis.

5. **How does the cross-modal effect vary with auditory depth?**

   When considering whether an audio-visual effect could be applicable to the quality-control of depth in S3D media, it is important to understand the nature of the effect. We were interested in how the effect varied with auditory depth, or more specifically the depth separation between the audio cues and the visual cues. We might expect that the effect's strength would increase as the audio-visual separation increases. But as the separation between audio and visual cues becomes very large, the brain may cease to consider them as one stimulus, meaning that the effect could have audio-visual separation limits. We answered both this research question and the previous research question using the same experiment. Participants viewed two images of the same cross-modal stimuli, one after the other. Whilst the visual depth of the stimulus did not change, the auditory depth did. Participants were asked to respond by identifying which stimulus appeared to be nearer to them. A cross-modal effect could be said to have observed if, in the significant majority of cases, participants believed that the nearer stimulus was the one with the nearer auditory component. Four different auditory depth changes were investigated in order to address specifically this research question. From our data set we then drew the psychometric function showing how the effect varied with audio-visual separation, and thus make conclusions concerning the effect's audio-visual separation limits. This measurement of the effect's psychometric function is a novel contribution of the thesis.

6. **How large is the cross-modal bias?**

   At this point in our work we had observed the impact of the cross-modal effect, an example of the inverted ventriloquist effect, upon a particular relative depth judgement task. This is different to measuring the size of the cross-modal spatial bias, which is a significant factor in deciding how applicable the effect is. A measurement of the cross-modal bias would tell us how much perceived depth the effect could add to a display's depth budget. If the effect usefully extends the depth budget, then we have clear evidence that it is important to consider audio-visual interactions when quality-controlling depth in integrated

S3D display and spatial sound systems. A new cross-modal experimental task, based upon the task used in our previous cross-modal experimentation, has been used to measure the bias size for the same four different audio-visual separations used in answering the previous research question. The measurement of this value and the consideration of its implications for quality-controlling depth in S3D media are the final novel contributions of this thesis.

1.3 Thesis structure

Each of the above questions is addressed by a separate chapter in this thesis. It was decided to write up each experiment individually in this way, because the results of each study offer a contribution to the motivation and design of later studies, making it the natural way to organise the thesis in a narrative form. A summary of this narrative, including inter-dependencies between chapters, is shown in Figure 1.3. We believe each experimental chapter offers novel contributions to the literature, and these are presented in the chapter's conclusions and then drawn together in the thesis conclusions.

We begin in Chapter 2 by reviewing the science of vision and audition in S3D displays and spatial sound systems, before looking in depth at the literature related to the audio-visual interaction in Chapter 3. With an understanding of the subject's background and the related literature, Chapter 4 then reviews the preliminary experiment undertaken by the author that marks the start of our research narrative. We then turn to the visual and auditory senses separately. In Chapter 5 we motivate further research seeking to quality-control depth by evaluating subjective impressions of S3D media that implements research we intend to use in our own cross-modal experimentation. In Chapter 6 we measure the minimum audible depth difference between our two loudspeakers, which will inform our choice of audio-visual depth separations in our cross-modal experimentation and help us screen the participants in our experiment for RAD perception. In Chapter 7 we report on new cross-modal experimentation seeking to confirm and expand upon the results from our preliminary experiment. We then extend the experimental task design in Chapter 8 so that we can measure the size of the audio-visual depth bias induced by the effect and thus reflect on its practical applicability. In the final chapter we draw together all our conclusions and discuss them within the context of the thesis and the research questions that were presented in the previous section.

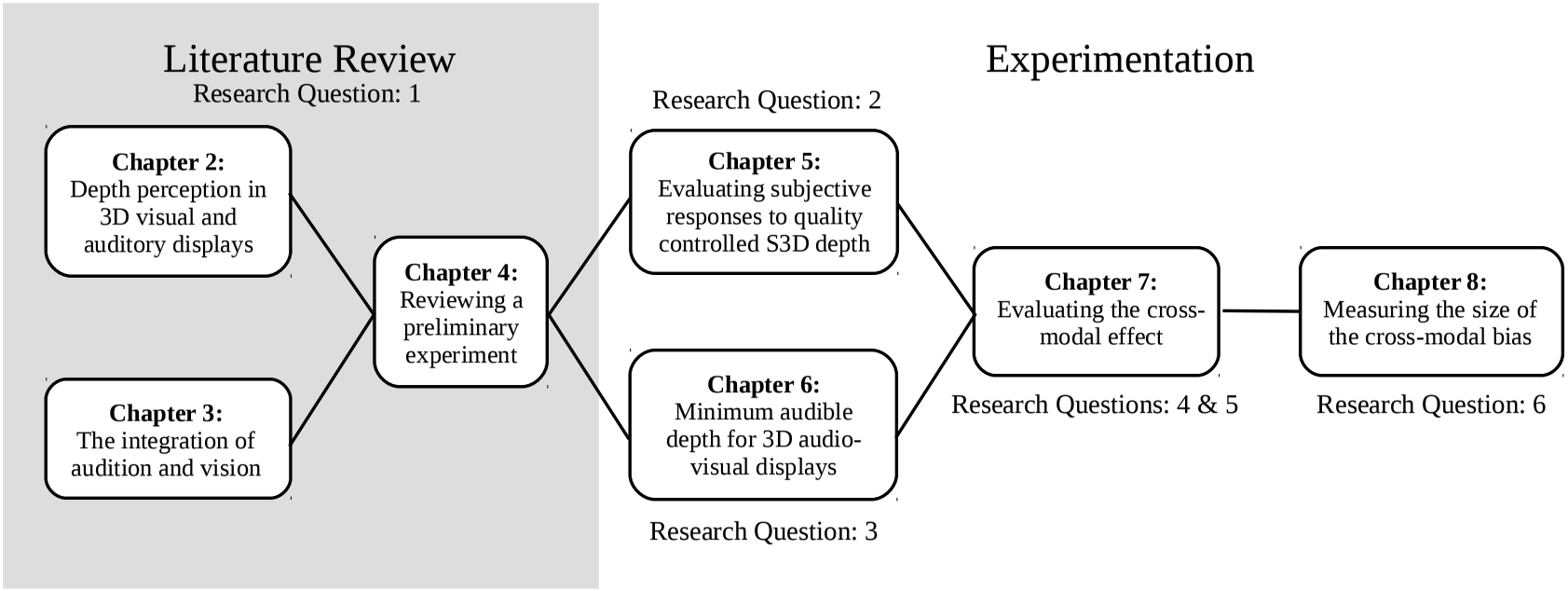

Figure 1.3: The thesis narrative. The literature review is comprised of three chapters. Each experimental study is written up as a seperate chapter, because the results and conclusions of each study contribute to the design and motivation of the later studies. The diagram also indicates where each subsidiary research question is addressed.

CHAPTER 2

Depth perception in 3D visual and auditory displays

This research began by seeking to answer the first research question detailed in Section 1.2: *What is there to be learnt from the literature?* As mentioned earlier, the literature review has been split into three chapters. This chapter addresses the broad background to the research, whilst Chapter 3 focuses on literature relating to cross-modal interactions. We finish in Chapter 4 with a detailed review of a preliminary experiment published previously by the author. The aim of these review chapters is to introduce some of the key concepts and studies in depth perception research, specifically concerning cross-modal interactions between audition and vision. They are therefore written with a particular interest in the application of this field to enhancing 3D audio-visual display systems and content.

This chapter is split into two primary parts: depth perception and the engineering of 3D displays. Each part is then further split into two subsidiary parts: vision and audition. We therefore begin by discussing visual depth perception in Section 2.1, including the cues to visual depth and the acuity of visual depth perception, before addressing auditory depth perception in Section 2.2. We then turn to the engineering of S3D displays in Section 2.3 before finishing by reviewing the design of spatial sound systems in Section 2.4. A summary of the background is not included in this chapter, as one was included in the previous chapter in Section 1.1.

2.1 Visual depth perception

Our ability to perceive visual depth in the natural environment is dependent upon a number of different cues. Understanding these cues is vital to understanding cross-modal interactions in depth perception. Coren and Ward (1989) identify twelve

Colour	Aerial Perspective Object Shading Texture Gradient	Pictorial
Size	Retinal Familiar	
Position	Interposition Height in the Plane Linear Perspective Motion Parallax	
	Stereopsis	Non-Pictorial
Physiological	Accommodation Eye Convergence	

Table 2.1: The twelve different cues to visual depth in the natural environment as categorised by Coren and Ward (1989). The classification into pictorial cues as outlined by Goldstein (2007) is also shown.

different cues used in visual depth perception in the natural environment and categorise them under four different headings: colour, size, position and physiological. Table 2.1 shows the breakdown of these cues that are discussed further below.

The majority of the cues in Table 2.1 are also classed as "pictorial cues" (Goldstein 2007). These cues are used for perceiving depth in 2D displays. The development of S3D displays has enabled content creators to utilise the binocular (or stereopsis) cue and, to a limited extent, the physiological cues as well, thus enhancing our ability to perceive depth in the content. Despite this, perception of depth in S3D displays is not perfect, due to a vergence-accommodation conflict discussed in further detail in Section 2.3.3.

In the real world, visual depth perception is generally accurate for distances less than 20 m, but in virtual environments it is typically subject to significant underestimations (Napieralski et al. 2011). These underestimations occur for all distances, including the near field (within arms reach), action space (within 30 m) and vista space (greater than 30 m). It is perhaps particularly interesting that there is no significant difference between action task performance in depth for high and low quality virtual environments, but there is a significant difference in verbal responses (Kunz et al. 2009). This is possibly because different neurological streams are used to give verbal reports and actions. Our performance in visual depth perception is by no means perfect, and by no means fully understood.

Figure 2.1: Aerial perspective tells us that the bluer mountains with less detail must be further away.

2.1.1 Colour cues

As light travels through air, part of it is reflected in random directions by oxygen and nitrogen molecules according to a process known as "Rayleigh scattering". This process scatters more light with shorter wavelengths at the blue end of the visible spectrum, than longer wavelengths at the red end of the visible spectrum. When viewing objects that are further away, more of this blue light is scattered into the viewer's line of sight, causing distant objects to appear bluer (Smith 2005). The scattering also reduces the contrast, and thus detail, in the view. This effect, which can therefore be used as a cue to depth, is shown in Figure 2.1 and is called **Aerial Perspective**.

The shadows formed by shining a directional light upon a colour consistent surface are able to provide cues to the surface's orientation and shape variance in 3D space. By comparing the colour shade of two points upon a colour consistent surface, viewers can often perceive a change in depth. This cue to depth is called **Object Shading**.

A texture is a pattern that appears on an object's surface. Many objects have a texture of sorts, such as the grain in a wooden object, or the yarn pattern in woven fabric. Such a texture will give cues to depth through other pictorial cues like linear perspective (Section 2.1.3), shading (Section 2.1.1) and familiar size (Section 2.1.2). The cue to depth given by a texture is called the **Texture Gradient**.

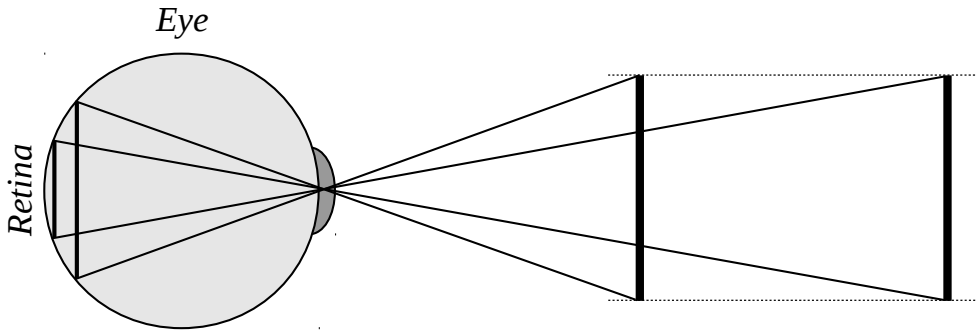

Figure 2.2: Two bars of the same length, but different depths, are projected onto the retina of the eye at different sizes.

2.1.2 Size cues

Objects that are closer to the viewer appear larger due to the size of image formed upon the retina. Figure 2.2 shows how two bars of the same size but different depths are projected on to the retina at different sizes. This **Retinal Size** can be used as a cue to depth, particularly when combined with **Familiar Size**. If the size of an object is familiar, then retinal size at any point in time can be compared with the familiar size to make a depth judgement. Familiar size is therefore called a relative depth cue.

2.1.3 Position cues

If an object occludes another object in the scene, the latter is assumed to be further away from the viewer. This can be seen in Figure 2.1 where the nearest mountain ridge occludes the further mountain ridge. This depth cue is called **Interposition**.

Linear Perspective is a result of the Retinal Size cue when viewing parallel lines that change in depth. As two parallel lines increase in depth, the distance between them appears to decrease. The effect is shown in Figure 2.3. Linear perspective is most effective when viewing objects with defined edges, such as a regular table or a brick building.

For objects resting on a plane that extends across the viewer and away from the viewer in depth (such as the earth's surface) the object's **Height in the Plane** can indicate its depth. If two ships are seen on the sea, the one nearer the horizon, and thus higher in the plane, is deemed to have the greater depth. This cue is also used when perceiving the depth of objects in other planes, such as plates on a table, or words on a page of a book.

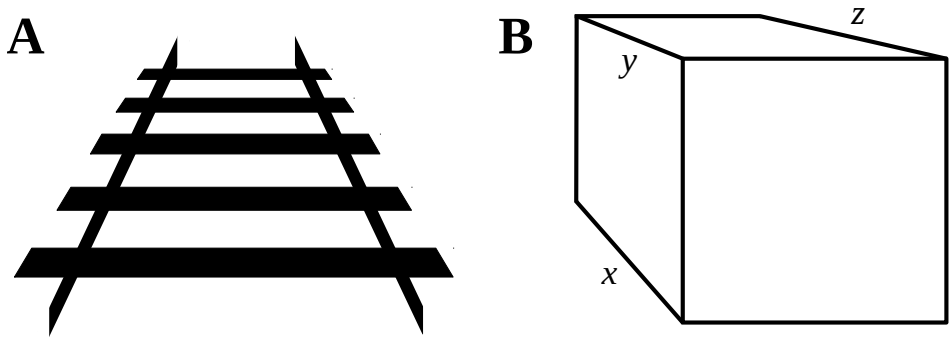

Figure 2.3: Two examples of linear perspective. **A:** Note how the converging train tracks result in the appearance of depth change. **B:** Edges *x, y* and *z* converge, giving the cuboid a greater sense of depth than if the edges were parallel.

As a viewer moves from side to side whilst focusing on a particular depth, all objects nearer than that focus point will appear to move in the opposite direction to the viewer's movements, and all objects further than that focus point will appear to move with the viewer. This effect is particularly noticeable as you look out of a window and compare the apparent movement of objects inside with the apparent movement of objects outside, as you move from side to side. Motion within a scene can also cue depth since further objects will appear to move slower than nearer objects (e.g. aeroplanes the fly higher appear to move slower). So far, all cues discussed have only required static scenes, but this cue, called **Motion Parallax**, requires movement of the scene content or viewer to be applicable.

Stereopsis is also classed by Coren and Ward (1989) as a position cue. Whilst the cues discussed so far form the basis of depth perception in 2D displays, the addition of the stereopsis cue forms the basis for depth perception in S3D displays. Humans view the world around them through two different eyes, each supplying the brain with a slightly different view. The difference between the two retinal images the brain receives from the eyes is called the binocular disparity and includes key information about the depth of the objects being viewed. These two images are combined by the brain in a process called stereopsis (Cumming and DeAngelis 2001; Patterson 1992) to create a single perceived image with embedded cues to depth. Many depth judgements depend upon the use of stereopsis when the other pictorial cues are weak. It is possible to perceive visual depth using just pictorial cues, or just binocular disparity, although using both of these significantly improves and quickens spatial perceptions.

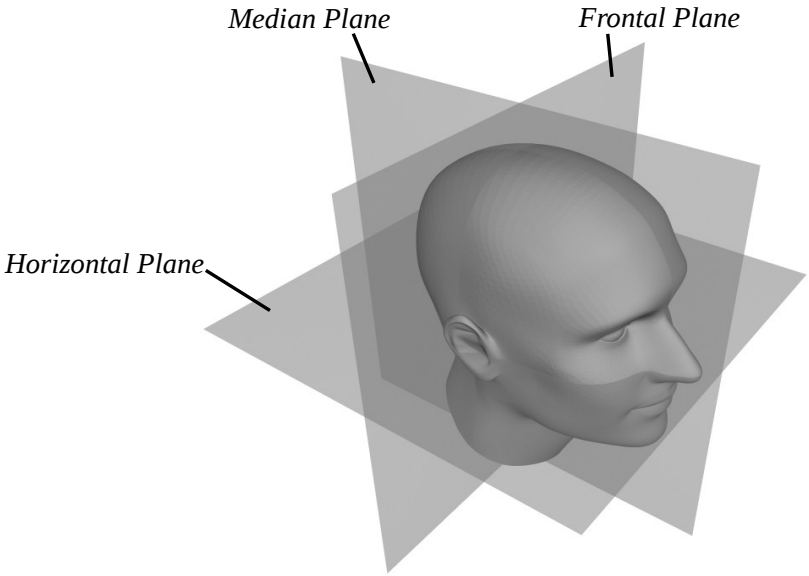

Figure 2.4: Showing the three orthogonal auditory planes that we refer to throughout this thesis. The median plane splits evenly between the left and right ear, whilst the frontal plane splits the space in front and behind the ears and the horizontal plane splits space above and below the ears.

2.1.4 Physiological Cues

Physiological changes in the viewer may also provide cues to depth. **Eye Accommodation** is the change in focus of the eye's lens, and will give a sense of depth. For instance as you refocus from viewing objects beyond a pane of glass to viewing a mark on the glass, you will be aware of a depth change and a change in eye vergence. **Eye Vergence** is the simultaneous movement of both eyes in opposite directions to maintain binocular focus upon an object. As an object moves closer, the eyes rotate towards each other to maintain a single focused picture of the object; this is increasing vergence. Eye vergence can be sensed, so is capable of cuing depth.

2.2 Auditory localisation and depth perception

Our ability to locate a sound source is due to a number of different cues that arise from the shape of the human head and the acoustical environment in which the source and listener are placed. Many of these cues can be explained by the simplistic *Duplex Theory* outlined in Section 2.2.1, with the finer details encapsulated in the head related transfer function (HRTF) and the binaural room impulse response (BRIR) discussed in Section 2.2.2. A detailed evaluation of cues to auditory depth

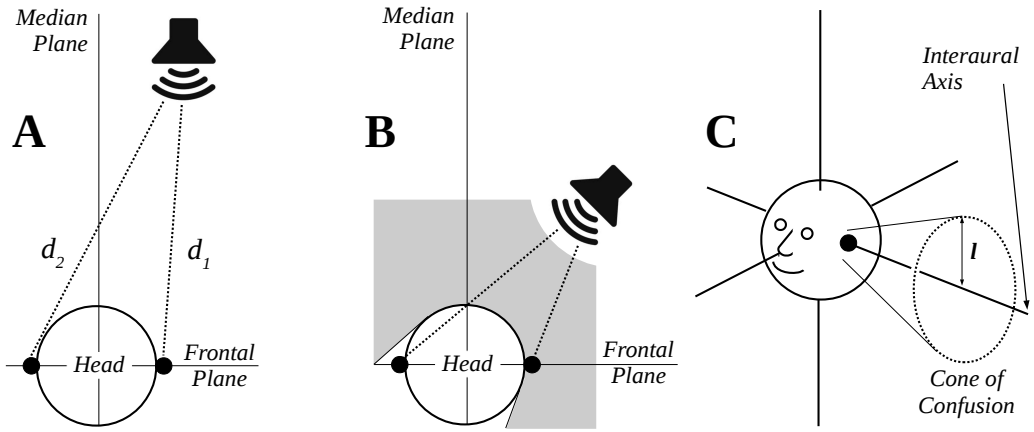

Figure 2.5: The Duplex Theory. **A:** Interaural time differences (ITDs) arise because $d1 \neq d2$ when the source is not located on the median plane. **B:** Interaural level differences (ILDs) arise from the shading effect caused by the head, allowing only reflected and diffracted sound to reach the contralateral (further) ear. **C:** Any set of sound sources placed on the cone of confusion (dashed line) will be indistinguishable due to them sharing the same ITD and ILD values. This holds for any value of l perpendicular to the interaural axis. The Cone of Confusion is a significant failing of the Duplex Theory.

is made in Section 2.2.3. Figure 2.4 outlines some terminology used in this section. For a detailed review of auditory depth perception, refer to the review *Auditory distance perception in humans: a summary of past and present research* by Zahorik et al. (2005).

2.2.1 The Duplex Theory

The Duplex Theory (Strutt 1907; Kapralos et al. 2008) is one of the earliest attempts at understanding human sound localisation. The theory is centred around modelling the head as a sphere to explain a number of different auditory localisation cues.

Figure 2.5A shows how the ear separation results in a path length difference for a sound travelling to each ear (unless the sound source lies on the median plane). Because of this path length difference, sound reaches the ipsilateral ear (the ear nearer to the sound source) before reaching the contralateral ear (the ear further from the sound source). This time difference, referred to as the ITD, is a key cue to sound localisation.

In Figure 2.5B the contralateral ear is shaded from the sound by the head, allowing only reflected or diffracted sound to reach the ear. Because of this there is an intensity (or volume level) difference between the sound at each ear. This ILD is also a key cue to sound localisation.

The Duplex Theory remains a simplistic model with failings, despite describing two of the most significant auditory localisation cues (ITD and ILD). Perhaps the most significant of these failings is the "cone of confusion" shown in Figure 2.5C. Any set of auditory sources in free space placed a constant perpendicular distance l from the interaural axis share the same ITD and ILD values, causing the auditory sources to be indistinguishable. This remains true for all values of l perpendicular to any point on the interaural axis, such that the auditory sources are positioned outside of the head. A more detailed understanding of the head and ear shape is required to explain our ability to distinguish between sources on the cone of confusion.

2.2.2 HRTF and BRIR

Later work undertaken by Batteau in the 1960's addressed the impact of body and ear shape on our auditory localisation ability (Batteau 1967; Brungart and Rabinowitz 1999; Kapralos et al. 2008). He discovered that humans can use sound interactions with the head, upper torso, shoulders and pinna of each ear to cue auditory localisation. These interactions are captured in the HRTF, a response that characterises how an ear receives sound from a point in space. HRTFs are theoretically calculated by solving the wave equation whilst considering the interactions with the body. In practice this is too complex, so various simplifications are made (the Duplex Theory is an extreme example of such a simplification). The function maps a sound in free space, to a sound as heard by a single ear (there is a different HRTF for left and right ears), and is dependent upon the sound's frequency, azimuth, elevation angle and distance.

There are other external factors that facilitate auditory depth perception which are not captured by the HRTF. One of the most significant external cues is the acoustical environment. A BRIR represents the response of a particular acoustical environment and listener to sound energy. The listener-specific part of the BRIR is defined in terms of the HRTF. BRIRs are typically measured using small microphones placed into a listener's ears.

2.2.3 Auditory depth cues

Of the depth cues covered by the models in sections 2.2.1 and 2.2.2, the most significant are (Coleman 1963):

- Loudness
- Reverberation

- Frequency spectrum
- Interaural differences (ITD and ILD)

The most quoted and widely known cue to auditory depth is loudness - specifically, the loudness at the listener. Unfortunately the loudness at the listener is often confused with the loudness at the source, resulting in a poor ability to judge absolute auditory depth in free space (Coleman 1962). This cue is therefore stronger if the listener has prior knowledge about the source, making loudness more suitable as a relative cue than absolute cue.

Another significant factor, which further complicates this field of research, is reverberation; our ability to detect auditory depth is directly dependent upon the environment in which the detection occurs. More specifically, the energy ratio of direct and reverberant sound (D/R) provides an absolute cue to the source's depth (Bronkhorst and Houtgast 1999), contrasting with lateral localisation which can be degraded by echoic environments (Moore and King 1999). Listeners can also discriminate differences in the D/R cue (Larsen et al. 2008), so the cue is also used in RAD perception, as we discuss at the end of this section.

The frequency spectrum of a source provides a relative depth cue, in a similar way to loudness. As sound propagates through air the proportion of low frequency content (relative to the high frequency content) increases, resulting in a different frequency spectrum that is dependent upon depth (von Bekesy 1938). Coleman (1968) reported that the frequency spectrum plays a dual role in auditory depth perception. He concluded that a greater proportion of high frequency content can indicate a closer sound at distances greater than a few feet, and a further sound at distances less than a few feet. He also confirmed that altering the frequency spectrum of the sound resulted in a different perception of the source's depth, but the use of this as a depth cue relies on knowledge of the source's frequency spectrum at other depths.

Inter-aural differences, explained by the duplex theory, also have an impact on perception of near-source depth (Coleman 1963). The nearer the source is to the head, the larger the ITD and ILD. Inter-aural differences are only of use in the near field, as the shading effect and path length difference quickly tends to zero in the far field (Nielsen 1991). It has been shown that, for a source on the inter-aural axis (see Figure 2.5C), the ILD can vary by as much as 20 dB for distances between 17.5 cm and 87.5 cm (Hartley and Fry 1921). ITD is not such a strong depth cue, although simple geometry is able to show that the ITD decreases with greater distance (Brungart and Rabinowitz 1999; Duda and Martens 1998).

Zahorik (2002) investigated how these cues are combined to give an overall perception of auditory depth. He hypothesises a framework in which each cue creates its own estimate of depth, which is then weighted according to the strength of the cue's content and the consistency of the cue's estimate when compared with other cue estimates. His experimental results show that the manner in which listeners weight two principle depth cues does not vary with depth change, but only with acoustical and directional changes.

Some recent studies have sought to understand cue combination specifically for RAD perception (Akeroyd et al. 2007; Kolarik et al. 2013a, b). These studies investigate how the loudness level and (D/R) cue are combined to perceive relative depth. These will be the dominating cues in our experimental setup. Akeroyd et al. (2007) found that RAD perception improved if both cues were available. This was confirmed by Kolarik et al. (2013b) who also concluded that performance was better with just the loudness cue than with just the D/R cue. However, they also found that performance using both cues was generally better for environments with longer reverberation times. Our experiment has therefore been performed in a semi-reverberant environment.

2.2.4 Acuity of auditory depth perception

A number of studies have psychophysically examined the acuity of absolute auditory depth perception (Speigle and Loomis 1993; Nielsen 1993, 1991; Bronkhorst and Houtgast 1999; Bronkhorst 2002; Zahorik 2002; Fontana and Rocchesso 2008) and RAD perception (Edwards 1955; Simpson and Stanton 1973; Strybel and Perrott 1984; Ashmead et al. 1990; Peter Barnecutt 1998; Volk et al. 2012). Despite RAD perception arguably being practised more in every day scenarios (sounds are rarely heard in isolation), the literature addressing absolute auditory depth perception far outweighs that addressing RAD perception, with the references given for absolute depth perception being just selected highlights of a much larger literature corpus. Such an unbalance is evident in the review of the field by Zahorik et al. (2005).

Nielsen (1991) discusses an experiment to assess the acuity of depth perception of sources on the median plane. Participants were played auditory stimuli through a series of speakers, and asked to judge where in the room the stimuli source was located. They recorded their judgement on a quantised map of the area around them. The speakers were placed on the median plane at distances of 1 m, 2 m, 3.5 m and 5 m. Each speaker was raised above the one in front to avoid any auditory degradation due to shading. Participants undertook the experimental task

in a small curtained cubicle at the centre of a room, so that the positions of the speakers and the room layout were unknown to them, while the acoustics were not notably degraded. The results show a large amount of variability between between individuals in their ability to hear absolute auditory depth, though there was a clear sense of learning demonstrated, suggesting that RAD perception is stronger than absolute auditory depth perception. The observed limit in dynamics of perceived distances is evidence of an acoustical horizon which is explored further by Bronkhorst and Houtgast (1999).

Bronkhorst and Houtgast (1999) investigated the effect of reverberation upon depth perception in an echoic environment using virtual acoustics (see Section 2.4). Listeners were presented with bursts of pink noise convolved with a BRIR that simulated the experimental room. They were then asked to rate the apparent depth of the sound source on a quantised scale. They found that the perceived depth was determined by three primary variables: virtual source distance, number of reflections and the relative level of reflections. When 27 or more reflections were used they, like Nielsen, observed an acoustic horizon at approximately 2 m. They present a mathematical model that expresses the perceived auditory depth of a sound as a function of the ratio between direct and reflected energies. They show that this model can accurately predict performance and explain the acoustical horizon through the use of an integration window for determining the energy of the direct sound.

Zahorik (2002) also uses virtual acoustics to assess auditory depth perception in an experiment addressing the discrepancy between perceived distance and actual distance. He noted that on the whole listeners under-estimate distances, though for near sources would often over-estimate. He fitted a power function to the data, of the form:

$$\psi_p = k{\psi_r}^a.$$

Where ψ_p is the perceived depth of the source, ψ_r is the actual depth of the source and k and a are constants. For perfect acuity $a = 1$ and $k = 1$. The average value of a across all listeners and stimulus conditions was approximately 0.39 and the average value of k was approximately 1.32. The fact that 0.39 is substantially lower than the veridical value of 1 supports the evidence for an acoustical horizon. Zahorik also noticed that consistent patterns of error arose in the listeners judgements across a variety of stimulus conditions including direction and source signal.

Most of the papers investigating RAD perception are concerned with the pressure discrimination hypothesis (PDH) which states that RAD perception is limited by the availability of discriminable differences in pressure, or loudness (Ashmead et al.

1990). Coleman (1963) noted that the apparent amplitude difference in dB of a sound played from a distance r compared to a reference distance r_0 is given by

$$20 \log \left(\frac{r}{r_0} \right)$$

We know the relative distance change ($\Delta r = r - r_0$ so $r = \Delta r + r_0$), which we can substitute into the equation to get

$$20 \log \left(\frac{r_0 + \Delta r}{r_0} \right) = 20 \log \left(1 + \frac{\Delta r}{r_0} \right).$$

The sensory threshold for wideband noise amplitude difference has been found to be between 0.3 to 0.5 dB (Miller 1947). Taking 0.4 dB as the sensory threshold for pressure difference and rearranging the equation gives

$$\frac{\Delta r}{r} = 10^{\frac{0.4}{20}} - 1 = 0.047$$

So the PDH predicts a MAD of about 5% of the reference distance (Strybel and Perrott 1984). Applying this to a television viewing distance of 2 m yields a MAD of about 10 cm. Various studies have presented evidence in agreement and disagreement with this result.

A number of earlier experiments yielded results that disagreed with the PDH (Edwards 1955; Simpson and Stanton 1973; Strybel and Perrott 1984), particularly at shorter reference distances. All of these studies used the method of limits experimental design to identify the sensory threshold. An auditory stimulus was played repeatedly whilst being moved from the reference position either toward or away from the participant. The participant then responded with either "towards" or "away" as soon as they were sure of the direction in which the stimulus was being moved. The positions of the stimulus at the point of response were recorded and the mean value taken as the threshold. All experiments gave results suggesting that as the reference distance decreased the threshold distance as a percentage of the reference distance increased. Strybel and Perrott (1984) investigated RAD perception over a large range of reference distances from 0.49 m to 48.76 m. They found that performance did roughly agree with the PDH for distances of 6.09 m to 48.76 m, but acuity still dropped off for smaller reference distances.

Ashmead et al. (1990) were the first to measure a result in agreement with the PDH for reference distances of 1 m and 2 m. Their experiment was undertaken in anechoic room using a single loudspeaker on a sliding platform. The two alternative

forced choice (2AFC) two-down/one-up adaptive experimental design set it apart from previous studies. The method of limits, used in previous studies, may have encouraged a more conservative judgement of the threshold due to participants being instructed to minimise response errors by waiting until they were sure of their answer. Participants were required to identify the nearer of two auditory stimuli at different depths, played sequentially from the same loudspeaker with a 1.5 s quiet gap, during which the speaker was moved to the new position. The first depth difference being tested was 10% of the reference distance and depths could be varied in 1% steps between 10% and 0% inclusive. With every two consecutive correct answers the depth difference decreased by 1% , but with every one wrong answer the depth difference increased by 1%. The result was a convergence upon the depth difference that yielded the threshold value of 70.7% correct responses from the participant. This aspect of the method has been explained in more detail by Levitt (1971). When a decrease in depth difference is followed consecutively by an increase, or vice versa, a "reversal" is said to have occurred. This procedure was repeated for 20 reversals before participant thresholds were calculated by averaging across the depth differences at each reversal (excluding the first five reversals that were treated as warm ups). The study demonstrated the importance of loudness in auditory depth perception by comparing results with a control case in which the loudness cue was removed using appropriate loudspeaker amplitude adjustments. In the control case response accuracy was significantly worse, though still significantly better than chance, suggesting that loudness is a key cue to depth, but not the only cue.

Volk et al. (2012) have measured the MAD using a wave field synthesis (WFS) system to position and play the auditory stimuli (see Section 2.4.3). Impulses of uniform exciting noise were used with a Gaussian grating in a 2AFC 2-down/1-up method combined with parameter estimation by sequential testing (PEST) for the step size adaptation. They measured a MAD of 5% at 0.5 m and 2% at 1 m, although for larger reference distances of 2 m and 10 m they measured 14% and 11%. These measurements do not appear to agree with the previous literature, suggesting that acuity improves for greater reference distances.

2.3 S3D displays

S3D displays add the binocular cue (see Section 2.1.3) to a scene using a mechanism that is capable of conveying different images to each eye. If the two correct angular views of the display's content are conveyed to the two eyes, the brain can perform

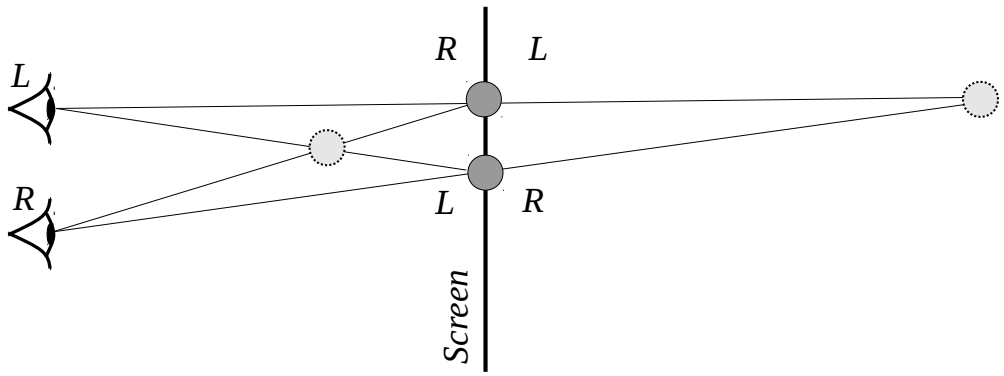

Figure 2.6: Showing both negative and positive screen parallax, and the resulting perception of depth. In the case where the object appears in front of the screen, the left and right eye images of the object cross over creating negative parallax (the left image of the object on the screen is viewed by the right eye, and the right image of the object on the screen is viewed by the left eye). In the case where the object appears behind the screen, the left eye and right eye images do not cross over creating positive parallax (the left image is viewed by the left eye and the right is viewed by the right eye).

stereopsis fusing the two images to obtain depth cues from the binocular disparity. The various different mechanisms for doing this are discussed below.

Jones et al. (2001) lists the benefits of S3D displays as:

- improved perception of depth relative to the display's surface
- improved spatial localisation
- improved perception of structure in visually complex scenes
- improved perception of surface curvature
- improved motion judgement
- improved perception of surface material

The horizontal difference between corresponding points in images received by the left and right eye is referred to as the screen parallax (Seuntiëns 2006). An object that is placed in the plane of the screen has zero screen parallax, whereas an object placed in front of the screen has negative parallax and an object placed behind the screen has positive parallax (shown in Figure 2.6). Depth can therefore be controlled by altering the amount of screen parallax, though the amount of depth actually perceived by the viewer depends upon more than just the screen parallax. The viewing distance, the viewing angle, and factors relating to the human visual system, such as interpupillary distance (Dodgson 2004), also effect the amount of perceived depth.

2.3.1 S3D display types

Holliman et al. (2011) have reviewed the various different forms of 3D displays, splitting the field primarily into two-view displays and multi-view displays. Two-view displays convey just two different images to the viewer; one for each eye. This popular form of S3D display can create the two images in a time sequential manner, or a parallel manner. Time sequential displays alternate between each eye's view, whereas parallel displays show both views at the same time.

There are multiple ways of splitting the two images between the two eyes. Wavelength selective displays do this using a colour anaglyph, colouring each image (often with cyan or red) so that they can be correctly filtered by coloured glasses. A similar technique uses polarised light and filters to split the two channels. Time sequential displays alternate between the two views at a rate quicker than the flicker fusion threshold (which Hecht and Smith (1936) measured to be approximately 58 Hz for bright stimuli). The views are then separated using *Shutter Glasses* that block the opposing eye's view as the screen alternates between each eye's image. Stereoscopes and head mounted displays have separate displays for each eye, which may either be directly placed in front of the eye, or linked to the eye using mirrors. "Waveguide Technology" projects images along pieces of glass using total internal reflection (Levola 2006, 2007). All of these methods require the use of glasses or some other sort of eye wear.

Auto-stereoscopic displays remove the need for any eye wear, using parallax barriers or directed back-lighting to send the light from each image to different points in space (Dodgson 2005, 2013). As such, most autostereoscopic displays require the viewer to postion their head in a "sweet spot" where each eye can see the correct image. A key technology used to overcome this problem is head-tracking, which enables the projection of the images into the viewing space to be dependent upon the position of the viewer (Dodgson 2006). Parallax barriers or lenticular lens arrays can be used to create repeated viewing regions, reducing the need for head tracking and allowing multiple people to view the display at the same time. Such displays still require the viewer to position themselves in a sweet spot, but due to the existence of multiple sweet spots, they provide more freedom to move around.

As a viewer changes their position relative to a two-view S3D display, the display's content appears to also move such that any occlusion in the scene remains constant; the technology has no "look-around" capability. Multi-view displays are able to show several pairs of images so that the occlusion in a scene is changed by changing the viewer's position relative to the display – they are able to look around

objects in the scene. These multi-view displays can be either full or horizontal parallax displays, where full parallax displays not only enable the viewer to look around objects in the horizontal plane, but also to look over objects in the vertical plane.

Multi-view displays are usually auto-stereoscopic and created using either head trackers or lens arrays. For each view, another two images have to be stored, increasing the amount of data required. This puts a significant technological limitation on multi-view displays, particularly when using high definition pictures. Many of the commercial multi-view displays currently available create multiple views at the expense of picture quality.

There are also non-stereoscopic 3D displays that use light emitting, light scattering or light relaying regions capable of occupying a volume rather than a surface in space. These "Volumetric Displays" create content that can be viewed from all angles by all people, but the technology still faces many significant limitations. The research field considered in this paper is specifically concerned with S3D images, so we will not consider these displays in further detail.

2.3.2 Limitations of S3D displays

S3D displays have many limitations that we need to be aware of. The recent surge of interest in S3D displays has brought with it concern regarding the possible detrimental effects that viewing might have upon the eyes. These detrimental effects often occur due to the limitations placed upon content and hardware by the commercially driven market. The many different factors that cause eye strain when viewing S3D images are listed by Shibata et al. (2011) as:

- eye wear
- crosstalk/ghosting
- misalignment between images
- inappropriate head orientation
- vergence-accommodation conflict
- flicker or motion artefacts
- visual-vestibular conflict

Many people find the eye wear irritative and uncomfortable to wear over extended periods of time. As well as this, the mass distribution of eye wear adds an unwanted financial and organisational cost for both businesses and viewers.

Crosstalk, also known as ghosting or leakage, is the leakage of the left eye's image into the right eye's view and vice-versa (Woods 2011). The same concept

can be found in 3D audio displays and is discussed in Section 2.4.2. Crosstalk has been shown to significantly reduce the magnitude of perceived depth in S3D images of both complex and simplistic scenes (Tsirlin et al. 2011, 2012). It can be mathematically defined as the leakage to signal ratio, typically expressed as a percentage. There are two types of crosstalk: *system crosstalk* which is entirely dependent upon the hardware, and *viewer crosstalk* which is dependent upon the perception of content. Crosstalk can be minimised using several different optical techniques such as apodisation.

Vertical misalignment (or vertical disparity) between images is not commonly part of the natural viewing experience, so can cause significant amount of eye strain. However, small amounts of vertical disparity can be used as a depth cue in certain situations, and parallels can be drawn with various situations in the real world where such has been shown to be the case (Read and Serrano-Pedraza 2009).

Different viewers perceive the same S3D content in different ways. This is a significant limitation of the technology. Content creators have to consider this carefully in order to offer the optimum viewing experience to the optimum number of viewers. Depending on which source is used, it is believed that approximately 5-12% of people are stereo-blind and unable to see binocular depth at all (Richards 1970).

2.3.3 The depth budget

The "depth budget" (Holliman 2010), otherwise referred to as the "depth bracket" (Block and McNally 2013) or "zone of comfort" (Shibata et al. 2011), is the depth available to content creators in creating S3D images that are comfortable to view by some appropriate majority of viewers (Jones et al. 2001). As the screen parallax increases, it become increasingly harder to fuse the images comfortably, perhaps due to the vergence-accommodation conflict mentioned in Section 2.3.2.

The vergence-accommodation conflict arises because the eyes focus upon the screen but attempt to verge upon a point out of the screen (shown in Figure 1.1). So vergence, but not accommodation, is useful as a depth cue when viewing S3D displays. The "zone of clear single binocular vision" is the set of vergence and focal stimuli that a viewer can see clearly while still fusing the two images (Shibata et al. 2011). This is different to the depth budget because images become uncomfortable to view long before the depth reaches the limit of clear single binocular vision.

The depth budget can vary widely for different individuals, but content is usually made to be viewed by the significant majority. As a result, the chosen depth budget is often smaller than needed for many viewers, in order to ensure that the majority

of people can comfortably watch the content. Nintendo have tackled this problem for 3D gaming on their 3DS device by adding a "Depth Slider" which allows the viewer to set a depth that is comfortable for them to view (Nintendo 2011). This comes with its own problems however, as users often need to be educated in its use and purpose.

2.3.4 Evaluating the S3D experience

The added creative medium of S3D depth has been employed with varying degrees of success. It is therefore important to evaluate the S3D experience through audience (or user-centred) research. Such studies are able to identify production goals, and indicate the progress that has been made towards them.

The study by Seuntiëns et al. (2005) argues that using 2D image quality models are not sufficient to evaluate 3D images. This is because the attributes they incorporate, such as noise, blur, colour or brightness, do not account for the added value of depth, which can be degraded by attributes such as keystone distortion, shear distortion or crosstalk. They present a study proposing the use of *viewing experience* and *naturalness* as evaluative concepts in order to better reflect the added value of depth in S3D images. In this study S3D images were degraded using various amounts of additive noise and shown to participants who rated them according to naturalness and viewing experience. The ratings of viewing experience gave significant effects for the amount of noise, the image shown, and whether or not the image was 3D or 2D. Naturalness yielded significant effects for the amount of noise in the image and whether or not the image was 3D or 2D. No interactions were found between any of the effects. The study therefore concluded that both naturalness and viewing experience account for the added value of depth.

Whilst the use of binocular cues may impact positively upon viewing experience and naturalness, they also have a negative impact upon other factors such as visual comfort, fatigue and sickness (Lambooij et al. 2009; Ukai and Howarth 2008; Nojiri et al. 2004). The prolific film-maker Lenny Lipton writes in his 1982 book *The Foundations of Stereoscopic Cinema* , "The danger with stereoscopic film-making is that if it is improperly done, the result can be discomfort. Yet, when properly executed, stereoscopic films are beautiful and easy on the eyes." Whilst this may not be the case for all people, improved visual comfort is undoubtedly a goal of high quality S3D and thus an important part of many audience-centred studies.

The study by Pölönen et al. (2012) has assessed the subjective responses of 85 participants to a S3D cinema viewing of the Hollywood blockbuster *Avatar*. The

participants filled out a series of questionnaires, including the Simulator Sickness Questionnaire, before and after watching the film. The post-viewing questionnaires included questions about viewing experience, naturalness and comfort. Results from this experiment could then be compared with a similar previous experiment in which participants viewed the film *U2 3D*. They found that approximately 10% of viewers may feel sick after a relatively long presentation, and that visual strain and sickness was roughly the same for the 165 minute long *Avatar* film and the 85 minute long *U2 3D* film. Viewing experience and naturalness both had average response values of approximately 7.5 out of 10. No reference measurements for these values were taken before the viewing.

A small group of studies undertaken by Obrist et al. (2011, 2012, 2013) have sought insight into audience response to stereoscopic three-dimensional television (3DTV). Data collection was run in a shopping mall over a three day period. During this time, 229 participants contributed towards results concerning *sickness* and 471 participants towards results concerning *presence* in the 3DTV viewing experience. A further 639 participants contributed towards results addressing children's' responses when watching 3DTV. They found that 88% of the participants who took part in the sickness study reported some symptoms of sickness. This sickness was influenced by gender and usually related to the visual system. The presence study found that presence was influenced by previously experienced discomfort, whether or not the viewer was standing or sitting and whether or not it was the first S3D viewing experience. The results from the children's study were very positive, with 71% of the participants saying they "like [S3D] very much" compared to just 5% holding a neutral or worse opinion and 73% of participants said they would like to watch 3DTV at home.

Both the uncontrolled environment and the rapid evaluation methods required in a shopping mall were identified as limitations by these three studies. Though perhaps the most overlooked aspect of these studies is the actual 3D content shown to the participants. The only information we are given about this content is the title, the length, the fact that they were produced by an unspecified industrial partner and a single 2D image from one of the films. We would expect the content to impact the viewing experience as significantly as the technology used to display the content, about which we are given much more detailed information.

The study by Richardt et al. (2011) reports an attempt to mathematically model the viewing comfort based upon parameters of the stereoscopic content. Specifically, their model is based upon a left-right check for consistent pixels between the left and right images. They validate their model with a perceptual study, in which they

show that subjective ratings of image quality are strongly correlated with the output of the model. The computational model they develop predicts the viewing comfort of stereoscopic images without the neeed for costly and lengthy perceputal studies.

2.4 3D auditory displays

Spatial (or 3D) sound systems aim to create sound sources at all required locations in a continuous 3D space for a given set of soundscapes. These are used in simulation, film and gaming, creating a new level of realism and opening a new field of action-tasks and cues available to content designers (Cater et al. 2007). They have been reviewed by both Kapralos et al. (2008) and André et al. (2010), and can be arranged into three categories: Section 2.4.1 discusses systems built from loudspeaker arrays, Section 2.4.2 addresses crosstalk cancelling systems and Section 2.4.3 looks at WFS.

2.4.1 Loudspeaker arrays

Loudspeakers are able to approximate 3D sound in many situations through the use of large speaker arrays and the technique of amplitude panning. Dolby Atmos, for the cinema, is a commercial example of a just such a spatial sound system (Sergi 2013). Amplitude panning is the arranging of loudspeakers and speaker amplitudes in order to simulate the directional properties of the ILD (Kapralos et al. 2008). At a simplistic level, the location of the sound source should be perceived as an amplitude-weighted combination of contributing loudspeakers. Using the labelling in Figure 2.7, this combination can be approximated by the stereophonic law of sines (Blumlein 1933):

$$\frac{Sin(\beta)}{Sin(\alpha)} = \frac{g_1 - g_2}{g_1 + g_2}$$

Ville Pulkki has worked extensively on the creation of sound fields using amplitude panning, beginning with his vector based reformulation of the technique entitled Vector Based Amplitude Panning (Pulkki 1997). This technique gives equations for virtual sound source positioning that are simple and computationally efficient, allowing 2 and 3 dimensional sound fields to be created from any number of arbitrarily placed loudspeakers. This technique has been explored and tested further in his later papers on the “Localisation of Amplitude Panned Virtual Sources” (Pulkki and Karjalainen 2001; Pulkki 2001).

It should be noted that in some simplistic soundscapes where sound need only be located discretely in a small number of places (smaller than or equal to the number

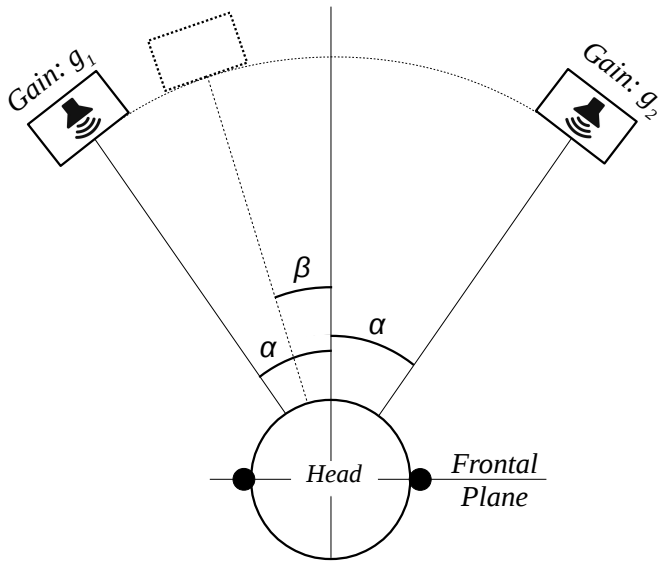

Figure 2.7: The technique of Amplitude Panning uses gain factors of loudspeakers that are equidistant from the user to create a virtual sound source (the dashed box). In this diagram β is dependent upon α and the two gain factors g_1 and g_2 as given in the law of sines.

of loudspeakers available in the given sound system) amplitude panning and large speaker arrays are unnecessary. By making each speaker a separate sound source and physically placing the speaker in the correct location, some 3D soundscapes can be recreated using small numbers of independently driven loudspeakers. Such simplistic sound spaces are rare in a continuous world, but may be usefully implemented for experimentation in this field of study (Turner et al. 2011).

2.4.2 Crosstalk cancelling systems

These 3D sound systems artificially utilise the cues discussed in Section 2.2 to give the impression of a 3D soundscape. This is done through calculation and application of HRTF or BRIR (see Section 2.2.2). Because the HRTF and BRIR is different for each ear, crosstalk between left and right channels must be minimised. Crosstalk is the leakage of the left channel into the right ear and vice versa (Kapralos et al. 2008). The simplest way to minimise crosstalk is to use headphones, though over extended periods of time these can become irritating like the use of glasses for 3D S3D displays. A significant amount of research has addressed crosstalk cancellation for a pair of stereophonic loudspeakers. The first crosstalk canceller for loudspeakers was implemented by Atal and Schroeder (1963) and was built upon the concept of

destructive interference of waves. A delayed and inverted version of the crosstalk from the left speaker is added to the right speaker, and vice versa. This introduces a distortion which is removed by a second round of similar crosstalk cancellation. The distortions added by crosstalk cancellation get smaller for each round, allowing a complete equation to be formulated (Kyriakakis et al. 1999). More recent advances in the field include Edgar Choueiri's development of the BACCH filters (Choueiri 2011). These BACCH filters eradicate any spectral colouration from the signal that is typically added by other crosstalk cancellation systems.

The dependence of the HRTF or BRIR upon position and orientation poses a difficulty for these systems, because the sounds being played in each ear should change as the head changes location or orientation. However, parallels can be drawn between this technology and multi-view or auto-stereoscopic S3D displays, in which a change in head location should cause a recalculation of the image view. Head tracking can be used to solve both of these problems. Further complications arise when trying to use binaural sound systems to match the position of auditory stimuli to visual stimuli in S3D images (discussed further in Section 3.1).

Perhaps a more significant failing of loudspeaker based crosstalk cancellation systems is the need for an acoustic sweet-spot in the same way that glasses-free S3D displays require viewing from a sweet-spot. Movement as small as 74–100 mm can result in the crosstalk cancellation collapsing, destroying the 3D effect (Mouchtaris et al. 2000). This still poses serious limitations upon the commercial viability of crosstalk cancellation systems.

Various comparisons have been made between real acoustics and virtual acoustics conveyed using a crosstalk cancellation display. Zahorik et al. (1995) compared loudspeaker sounds and virtual sounds in the free field using a small, acoustically unobtrusive headphone system. The results showed that, under certain conditions of virtual synthesis, it was not possible to discriminate between the real and virtual sound positions. A very similar experiment, undertaken by Kulkarni and Colburn (1998), investigated how accurately the HRTF have to be reproduced to achieve sounds that cannot be discriminated from real sounds. This experiment took place in an anechoic space and used small tubes to create an acoustically unobtrusive headphone system. The smoothed HRTF were constructed from a truncated Fourier series of the HRTF log magnitude spectrum. Once again the virtual and real acoustics were deemed indistinguishable, even for a surprisingly large amount of smoothing (just 32 terms of the Fourier series). H.A.Lagendijk and Bronkhorst (2000) also performed a comparison, using a slightly different setup. Instead of placing small headphones in the ears, they used a frame to mount a headphone a couple of cen-

timetres away from the ear. Once again they found real and virtual sound sources to be indistinguishable for the free field, but declared that further work was needed to compare the two systems in a reverberant space.

2.4.3 Wave field synthesis systems

WFS is a method of virtual sound source production (Berkhout et al. 1993; Boone et al. 1995; Kapralos et al. 2008) based upon Huygen's principle, that states:

Any wave front can be regarded as a superposition of elementary spherical waves.

Virtual wave fronts can therefore be synthesised from a set of wave fronts emitted by a large array of individually driven and closely spaced loudspeakers. The technique was initially developed by Berkhout (1988), using the mathematical basis of the Kirchhoff-Helmholtz integral which can be applied and interpreted as:

Sound pressure in a volume free of sound sources is completely deterministic if the sound pressure and velocity at all points on the volume surface are also deterministic.

Spatial perception of virtual auditory sources presented using WFS does not depend upon the listener's position or orientation. This makes WFS a uniquely attractive solution to spatialised audio for theatres and other multi-purpose auditoriums. The listeners, of which one can be catered for as easily as many, are free to move around the listening area enveloped by a wave field with natural time and space properties.

Despite this, there are certain characteristics of WFS systems that have hindered their widespread commercial availability. Due to the huge increase in the number of loudspeakers required by adding the third dimension, WFS systems typically restrict all sound sources to lie in a single plane (Boone 2001). Also, the highest frequency achievable by the system is inversely proportional to the spacing between loudspeakers (Verheijen 1998). Smaller spacings require a larger number of smaller loudspeakers that will cost more, placing another limit on the commercial viability of WFS.

In Section 1.1 we provide a summary of the background, and in Section 3.5 we discuss the relevant conclusions that can be drawn from the literature.

CHAPTER 3

The integration of audition and vision

The previous chapter detailed the wider background of this thesis, assessing each modality separately through a review of: auditory depth perception, visual depth perception and the engineering of spatial sound systems and S3D displays. However, the focus of our over-arching research question is upon *audio-visual* depth. We now examine studies that address the integration of audition with vision, for the purpose of building state-of-the-art display systems or understanding the perceptual effects that can occur between the different modalities. This chapter therefore continues to answer our first research question, "What is there to be learnt from the literature?"

The field of audio-visual interactions is difficult to structure into a clear taxonomy. The taxonomy we have chosen is shown in Figure 3.1. This is centred around our interest in the inversion of the ventriloquist effect to extend an S3D display's depth budget. We would like audition to influence vision, and we are less interested in whether vision can influence audition. The taxonomy also focuses upon the "third dimension", with this being depth, which is enhanced by the technologies we are exploring.

In Section 3.1 we discuss previous work undertaken to combine spatial sound systems with S3D displays. Such 3D audio-visual displays are the application scenarios for the work presented in this thesis. We then begin to examine the literature that addresses audio-visual interactions, with an introduction in Section 3.2. Our work is interested in assessing whether audition can influence vision, so in Section 3.3 we review studies where this has been observed. We continue in Section 3.4 by looking at auditory-visual interactions in depth perception. Finally, in Section 3.5, we draw conclusions from the literature that we have reviewed so far.

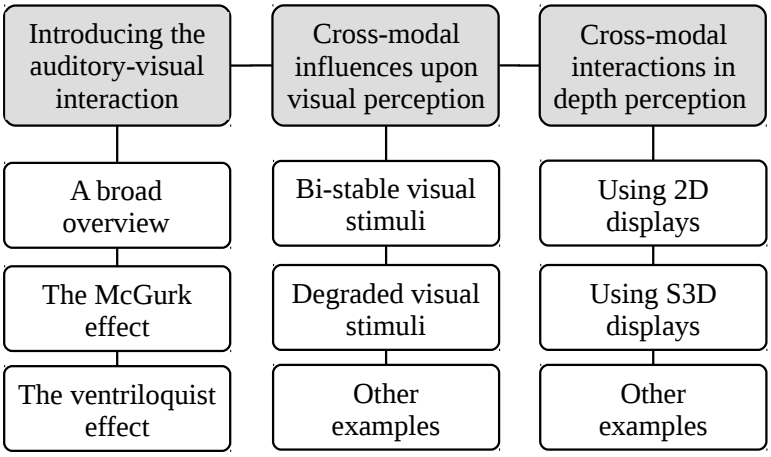

Figure 3.1: It is difficult to structure the literature addressing audio-visual interactions into a clear taxonomy. We have chosen to focus upon two particular interests of ours: that of audition influencing vision, and of audio-visual interactions in depth perception. We find that examples in literature addressing these two interests can be further broken down into three categories. Each category is addressed by a section in this chapter.

3.1 Integrating S3D displays and spatial sound systems

A number of recent projects have worked to develop integrated 3D audio and visual display systems (Kuhlen et al. 2007; Rebillat et al. 2010; Springer et al. 2006; Evrard et al. 2011). The majority of such attempts use WFS spatial sound systems because of their commercial availability since 2003 (Springer et al. 2006), as well as their independence from the listener's position. In order to achieve a spatially coherent audio-visual virtual environment, a multiview display must be used that allows people to look around the display's content. In a two-view S3D display the occlusion in an image cannot vary, so as the viewer changes position, the visual content also changes position, resulting in a disparity between visual and audio content. Such an unnatural effect is not conducive to an immersive environment and would require designing a location-dependent sound system instead of a location-dependent visual display.

The paper by Springer et al. (2006) describes a system that uses a 2D WFS sound field and a user-tracking multi-view S3D display. True spatial sound fields are very difficult to create with WFS and most systems create a 2D sound field in azimuth and depth instead. As discrepancies between the visual source and the auditory source have been shown to be unnoticed within a deviation of 22° in the vertical plane (de Bruijn and Boone 2003), this should not cause a significant problem.

Springer *et al.* developed two different scenarios to test the system: the "Billiard" scenario and the "Forest Brook and Stones" scenario. The billiard scenario was used to verify the system's audio-visual synchronisation. The user could hit one of two balls with a "virtual cue". Sounds were made as the cue hit a ball, when a ball hit a cushion and when a ball hit a ball. This suited the WFS system well, because a ball's movement is confined to the 2D plane of the table. The Forest Brook and Stones scenario placed the user in front of a forest scene with a brook flowing from left to right. Users positioned a S3D cursor above the brook or above the forest floor behind the brook. By clicking they were able to drop a stone from the cursor's position which made a dull thud if it fell on the forest floor or a splash followed by a gurgling if it fell in the brook. Both the billiard scenario and the stone dropping task are paradigms which involve cross-modal events and spatial judgements, so could be drawn upon to further investigate cross-modal effects in depth perception.

Rebillat et al. (2010) have been working on a project entitled SMART-I^2 which stands for "Spatial Multi-user Audio-visual Real-Time Interactive Interface". The system outlined in this paper uses Multi-Actuator Panels to create a WFS system. Multi-Actuator Panels are stiff light-weight panels with multiple electro-mechanical exciters attached to the back (Boone 2004). They can be used as multi-channel speakers, though they are rarely more than 1m^2 in size. For SMART-I^2 a novel 5m^2 Multi-Actuator Panel was created and used as a projection screen for the user-tracked S3D display. The result is a high-quality spatial audio and S3D video system that can be used in a wide range of virtual reality applications.

The study by Kuhlen et al. (2007) focuses upon content delivery systems for broadcasting S3D immersive environments, including spatial audio and multi-view S3D images. A system called DIOMEDES (Distribution Of Multi-view Entertainment using content aware Delivery Systems) has been created with *Digital Video Broadcasting-Terrestrial* and *Peer to Peer* technologies. Once again, WFS is used to display spatial audio.

These studies all address the design of potential application scenarios for the work presented in this thesis. They provide a long term direction for work assessing cross-modal interactions. By considering the challenges faced in these studies, we may also bring to light situations where cross-modal effects could be useful. For instance, the ventriloquist effect could play a role in reducing the spatial resolution required from a WFS system. People may perceive the auditory position to be the visual position within a certain degree of audio-visual separation. The cross-modal use-case scenarios used to test these systems also provide useful application scenarios for the work presented in this thesis.

3.2 Introducing the auditory-visual interaction

There is a significant amount of literature reporting examples of interaction between auditory and visual perception. Here we introduce this cross-modal interaction through a brief overview (Section 3.2.1) and describe two of the most significant examples found in the literature: the McGurk effect (Section 3.2.2) and the ventriloquist effect (Section 3.2.3).

3.2.1 A broad overview

There has been gathering interest in the potential benefits to both task performance and presence in HCI that may be given by integrating auditory information carefully with visual information. Examples of this in commercial computing includes the click made by an Apple iPhone, iPad or iPod Touch when pressing a keyboard button, or the noise made by Microsoft Windows when an error occurs. In visualisation, sound has been usefully paired with the following (Minghim and Forrest 1995):

- data representation
- perceptual issues
- interaction processes
- adding a time dimension
- validation of graphical processes
- memory of data and properties

It is clear that auditory displays can effectively convey information and cope with complex structures, complementing visual information. However, this thesis is primarily concerned with the ability of auditory information to alter our visual *perception*. The auditory-visual interaction is being explored for the purposes of extending the depth budget and enhancing S3D media, as discussed in Chapter 1. Specifically we are interested in whether the auditory-visual interaction can change our perception of depth in a S3D image. In this chapter we seek to give a broad-based review of audio-visual interactions so that we might draw upon studies in related fields to find new routes through our own field of audio-visual depth perception in S3D media.

The auditory-visual interaction is a broad field, which is difficult to break down into a clear and simple taxonomy. For the purposes of this review we consider

cross-modal influences upon auditory perception in sections 3.2.2 and 3.2.3, then investigate cross-modal influences upon visual perception in Section 3.3 while leaving any discussions of cross-modal influences upon depth perception until Section 3.3.

Although discovery of the interaction between visual and auditory perception is widely attributed to McGurk and MacDonald in 1976 (discussed in Section 3.2.2) scientists had been comparing the two senses for some time. Colavita (1974) published an interesting paper entitled *Human Sensory Dominance* in which he undertook four experiments to investigate the reaction times of humans to auditory and visual stimuli, and to see whether either of the two senses dominated the other attentionally. The experiments were broken into a series of tests, in which either a light would be shown or a tone would be played. The participants were asked to press a button if the light was shown and a different button if the tone was played, as soon as they recognised a stimulus. The light from the bulb and the tone from the loudspeaker were subjectively matched for intensity prior to the experiment. On the whole it was found that reaction to light was slower than the reaction to sound. In some of the tests both the light and sound were shown, which were initially dubbed as "mistakes". Interestingly, despite the slower reaction time to light, the vast majority of participants would respond with light in these cases. In fact, in some cases, participants would not comment on the fact that both stimuli had been played and a "mistake" had occurred in the system, suggesting they did not notice that the sound had been played. Colavita's work therefore demonstrated that, despite the auditory system's quicker response time, light is attentionally dominant.

Other cross-modal interactions have also been explored, such as the haptic-visual interaction. It has been shown that haptic augmentation of visual display can improve perception and understanding of the display content and can provide a two-fold improvement in task performance (Brooks et al. 1990). Also, haptic stimulation of a finger placed at the end of a static line on a display is capable of inducing the perception of the line unfolding from the point of stimulation (Shimojo and Hikosaka 1997). Given that this review is primarily focused upon the use of sound to influence visual perception we will not go into further detail regarding other cross-modal interactions.

3.2.2 The McGurk effect

The earliest example of the auditory-visual interaction in the scientific literature is the paper *Hearing Lips and Seeing Voices*, published by McGurk and MacDonald (1976). It was ground-breaking because it was the first suggestion that speech recog-

Stimulus			Response (% of subjects)				
Aud.	Vis.	Subjects	Aud.	Vis.	Fused	Comb.	Other
ba-ba	ga-ga	3-5 yr (n=21)	19	0	81	0	0
		7-8 yr (n=28)	36	0	64	0	0
		18-40 yr (n=54)	2	0	98	0	0
ga-ga	ba-ba	3-5 yr (n=21)	57	10	0	19	14
		7-8 yr (n=28)	36	21	11	32	0
		18-40 yr (n=54)	11	31	0	54	4
pa-pa	ka-ka	3-5 yr (n=21)	24	0	52	0	24
		7-8 yr (n=28)	50	0	50	0	0
		18-40 yr (n=54)	6	7	81	0	6
ka-ka	pa-pa	3-5 yr (n=21)	62	9	0	5	24
		7-8 yr (n=28)	68	0	0	32	0
		18-40 yr (n=54)	13	37	0	44	6

Table 3.1: Showing the percentage breakdown of responses given in the *Hearing Lips Seeing Voices* experiment (McGurk and MacDonald 1976) between subject groups as they depend on the auditory and visual stimuli presented. The participants have the option of responding either correctly with the auditory stimulus (aud.), correctly with the visual stimulus (vis.), with a fused syllable, with a combination of syllables (comb.) or with some other response.

nition may be more complicated than a stand-alone auditory process. The effect presented in the paper has been dubbed *the McGurk effect* and is a relatively well known illusion outside the scientific community, simply due to its illusive strength. McGurk and Macdonald write about the strength of the effect in their paper: "We ourselves have experienced the effect on many hundreds of trials; they do not habituate over time, despite objective knowledge of the illusion involved".

A film of a young woman's head speaking the syllables [ba],[ga],[pa] and [ka] was mismatched with a sound track of her speaking [ga][ba][ka][pa] respectively. Large percentages of people reported an illusion where the mismatched visuals altered what they heard. In the case of a visual [ga] mismatched with an auditory [ba], a fused syllable such as [da] or some combination of the constituent syllables such as [gabga] or [bagba] were heard. The full results for this experiment are shown in table 3.1.

It should be noted though, that in the context of this field, the McGurk effect marks no more than the founding of the research into auditory-visual interactions. It is an example of visual perception influencing auditory perception, whereas we

are investigating whether auditory perception can influence visual perception.

3.2.3 The ventriloquist effect

The term "ventriloquist effect" refers to the perception of a sound emanating from a spatially disparate visual source instead of its true source. This illusion became popular as a form of entertainment from puppeteers during the Victorian era. The effect is also used when viewing video with a sound track played through stereo or surround sound systems. Although the sounds come from speakers in significantly different locations to the visual sources on the screen, is still perceived as the sound emanating from the screen.

There have been various different attempts at predicting bias (the distance between the perceived stimulus and the true stimulus) caused by the ventriloquist effect. The earliest attempt modelled the effect as a winner-takes-all competition between the the two senses, in which the most reliable sense "captured" the other. Typically, in the vast majority of situations considered, this would be the visual sense and so the model was called *Visual Capture*.

Later, a model proposing that the perception of an object's location is based upon a blend of information from both senses using maximum likelihood estimation (MLE) was proposed (Clark and Yuille 2001). The perception of an object's location is predicted to be the weighted average location of the constituent sensory sources, where the weights are dependent upon the reliability of that source (shown in Figure 3.2). Reliability of a sensory source is measured as the inverse of the variance of the distribution of inferences based upon that source. The general statistical concept of MLE (Rice 2007a) is used to find the most likely variance value from the set of spatial single-sense inferences. Visual Capture can then be described as an extreme case of MLE, in which the visual information has a reliability of 1 and the auditory information as a reliability of 0.

Mathematically, MLE can be constructed in the following way (Battaglia et al. 2003). Let L be the the best possible location estimate and a and v indicate audition and vision respectively, such that L_a and L_v indicate the best location estimate of the auditory and visual sources respectively. Also let σ^2 be a variance such that ${\sigma_a}^2$ and ${\sigma_v}^2$ are the variances in judgements of the auditory source's location and the visual source's location separately. Then using this terminology MLE states that:

$$L = w_a L_a + w_v L_v.$$

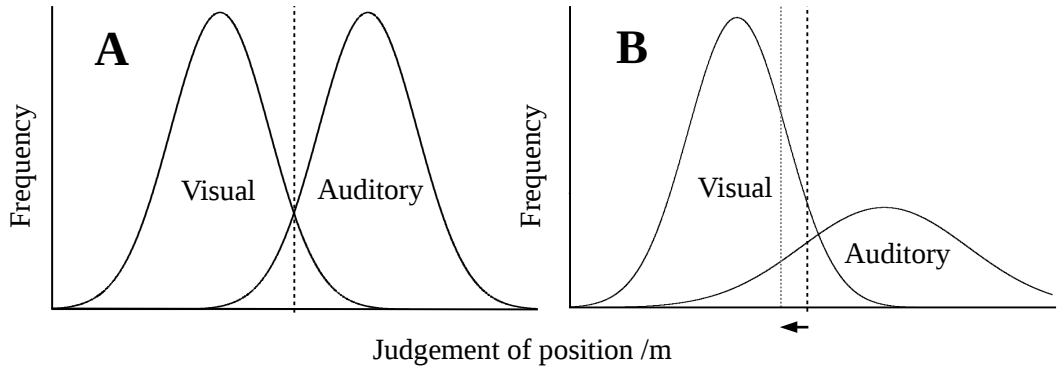

Figure 3.2: The predicted location of a cross-modal spatial perception depends upon the variance of the constituent single-sense perceptions in MLE. **A:** Auditory and visual sensory information is equally reliable, with equal variance in judgements of each source's position, causing the cross-modal perceived location to be predicted as exactly half way between the visual and auditory source. **B:** The visual sensory information is more reliable (smaller variance) than the auditory information, so the cross-modal perceived location is nearer the visual source than the auditory source.

Where:

$$w_a = \frac{{\sigma_a}^{-2}}{{\sigma_a}^{-2} + {\sigma_v}^{-2}} \qquad w_v = \frac{{\sigma_v}^{-2}}{{\sigma_v}^{-2} + {\sigma_a}^{-2}}$$

Figure 3.2B shows a worked example of these equations. In a one dimensional environment where the visual stimulus is placed at 0.35 m, the auditory stimulus is placed at 0.65 m and they have single-sense judgement variances of 0.1 m and 0.17 m respectively, the perceived cross-modal location predicted by MLE is 0.427 m. Whereas in Figure 3.2A the positions are the same but the variances are both 0.1 m, so the MLE predicted position is simply half way between the two sources at 0.5 m.

Battaglia et al. (2003) suggest that both of these models can be improved upon by combining them in a Bayesian integration. They undertook an experiment in which the participants were asked to judge relative spatial differences in auditory, visual and auditory-visual stimuli. The visual signal could be degraded by using five levels of noise. The first experimental phase looked at single-sense responses and the second phase looked at cross-modal responses. Responses to the multi modal case showed a tendency for the use of the visual signal to be used less, as the signal quality degraded, but there remained a bias towards the visual signal over the auditory signal. Therefore, a Bayesian integration (Rice 2007b) was proposed that is identical to MLE, except that a prior probability distribution is used that leads the model to make greater use of the visual system.

Other models have been proposed, and of particular note is the normative model

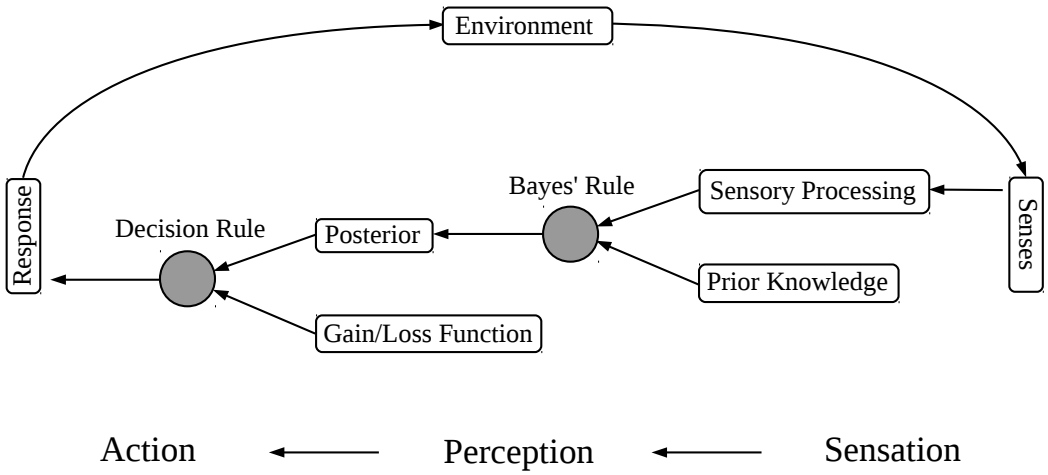

Figure 3.3: The information flow in the Bayesian model of perception (Ernst and Bulthoff 2004). The difference between MLE theory and Bayesian theory is the use of prior knowledge combined with sensory information to form the final percept (the posterior).

developed by Shams et al. (2005) that does not assume a single cause for all sensory signals. As the separation between an auditory and visual stimulus increases so does the likelihood of them not converging as one percept. Behavioural studies have shown that for large auditory-visual conflict the two stimuli are treated independently by the nervous system, and for moderate conflict the two stimuli may be partially integrated (shifted towards each other) but not fused as one object (Shams and Kim 2010). This model is also constructed using Bayesian statistics, employing Bayesian inference to infer signal causes and prior knowledge of events.

Hairston et al. (2003b) further investigated how cross-modal bias and perceived spatial unity depends upon the separation between auditory and visual components of the stimulus. They undertook two experiments. These showed that the audio-visual bias is correlated with perceived spatial unity and inversely correlated with localisation variability. The audio-visual bias was maximised when the visual stimulus appeared in centre of vision, but the degree of variability observed between participants was substantial.

The vast majority of research addressing the ventriloquist effect has focused upon perception of static objects. Static objects are relatively rare in commercial S3D media. Despite this an investigation into cross-modal motion perception in depth is beyond the scope of this project, so just a brief discussion of literature in this field will be undertaken.

It is important to identify the many different forms of illusory effects in motion perception that arise due the auditory-visual interaction. Firstly, the literature has

examples of a static stimulus in one modality affecting aspects of motion processing in another modality such as trajectory (Spelke et al. 1983; Sekuler et al. 1997), speed (Manabe and Riquimaroux 2000) and threshold for apparent motion (Staal and Donderi 1983; Ohmura 1987). There are also reports of a moving stimulus in one modality influencing perception of a static stimulus in another modality (Ehrenstein and Reinhardt-Rutland 1996). But perhaps most relevant and interesting to this project is the question of whether motion in one modality can influence the perceived motion in another modality (Mateeff et al. 1985; Soto-Faraco et al. 2002).

Soto-Faraco et al. (2002) specifically searched for the ventriloquist effect in motion perception. They used two light emitting diodes on either side of the participant's mid-line of sight to create visual apparent motion and two loudspeakers positioned either side of the participant's median auditory plane (Figure 2.4) to create the auditory motion with amplitude panning. Their results demonstrate a strong cross-modal interaction in the domain of motion perception. Vision is shown to cause an illusory reversal of auditory motion (which is non-existent when auditory motion is solely concerned). The strength of this effect depends upon spatial coincidence in the trajectory of lights and sounds and and upon the type of apparent motion being experienced.

Just as with the McGurk effect, the ventriloquist effect is an example of visual perception influencing auditory perception. In this research we are seeking to invert the ventriloquist effect, which is something that Recanzone (2009) says is possible but technically challenging.

3.3 Cross-modal influences upon visual perception

We now focus on the literature reporting examples of auditory perception influencing visual perception. The matter has been reviewed extensively by Shams and Kim (2010). The ability for auditory information to influence visual perception is dependent on the strength of each sensory percept, as discussed in Section 3.2.3. In the majority of situations visual perception is undoubtedly stronger, which then gives rise to the ventriloquist effect and the McGurk effect. For this reason, most examples of cross-modal influences upon visual perception are found when the visual cue is bi-stable or degraded in some way. Bi-stable stimuli examples are discussed in Section 3.3.1, and degraded visual stimuli are addressed in Section 3.3.2. Finally, Section 3.3.3 looks at the few examples where the visual stimulus appears to be

Figure 3.4: Bi-stable stimuli have a characteristic that can be viewed in one of two different ways. **A:** This image can be viewed either as a vase or two faces. Typically, only one of the two views can be seen at any point in time. **B:** These images are frames from the *Spinning Dancer* illusion (Kayahara 2003) showing a dancer whose spinning direction is bi-stable. All frames in the animation can be viewed in one of two ways: either the left leg points outwards, or the right leg points outwards.

neither bi-stable nor degraded.

3.3.1 Bi-stable visual stimuli

Bi-stable visual stimuli have a characteristic that can be interpreted in one of two ways and are usually classed as a form of visual illusion. Some examples are shown in Figure 3.4.

One of the first examples of a cross-modal influence upon visual perception was reported by Sekuler et al. (1997), in which the visual perception of the path of two moving discs was shown to be altered by a simple auditory click. Two discs were shown to stream diagonally across each other, one travelling from the top-right to the bottom-left and the other from the top-left to the bottom-right. The discs can either be perceived to stream through each other, or collide and bounce apart. A click, played as the discs coincide, caused the majority of people to perceive a collision where previously they had perceived the discs to stream through each other in the silent case. Watanabe and Shimojo (2001) extended this work by investigating the effect of further sounds with similar acoustic characteristics supplied before and after the collision. They discovered that the collision perception was attenuated by these further sounds, suggesting that there is an aspect of auditory-grouping that is context sensitive and utilized by the visual system for solving ambiguity.

Two super-imposed horizontal gratings moving in opposite vertical directions form a bi-stable pattern that can be perceived to have either an upwards motion or a downwards motion. A participant's perception of the resulting pattern's motion

can be influenced by a tone of changing pitch. This shows that sound without spatial information is capable of influencing visual motion perception (Maeda et al. 2004). Human speech, including words that lead the direction of motion (such as "up" and "down"), were not found to have the same effect. Several experiments were undertaken to ensure that this was a perceptual effect and not a response bias (results determined by a post-perceptual effect, rather than the desired perceptual effect during the task completion). The second experiment used eye-trackers to exclude possible confounding of motion perception, due to sound-triggered eye movement being used as a cue. The third experiment varied the stimulus-onset asynchrony between gratings and sound to exclude the possibility that the effect is due to any other top down influences.

Binocular rivalry is a phenomenon where two different images are shown to the two eyes, resulting in a randomly alternating perception of each image. It can also be classed as a bi-stable visual stimulus. Experiments have been undertaken using binocular rivalry that have demonstrated the importance of congruent auditory and visual stimuli in creating a fused percept (Conrad et al. 2010). Other experiments exploring the mechanisms of our control over awareness in response to sensory input (van Ee et al. 2009) have shown that matching temporal sound greatly enhances control over holding a temporal visual stimulus dominant in binocular rivalry. This was also the case when the sound was temporally delayed, because of the constant phase difference. Temporal auditory perception is much stronger than spatial auditory perception.

3.3.2 Degraded visual stimuli

The Bayesian and MLE models say that the final cross-modal perception is dependent upon the reliability of the constituent stimuli (Section 3.2.3). Therefore, to achieve cross-modal influences upon visual perception, the auditory signal needs to be as strong as possible relative to the visual signal. This can be obtained either by strengthening the auditory stimulus, or degrading the visual stimulus. In this section we discuss examples of a degraded visual stimulus causing auditory information to influence visual perception.

It has been shown that spatial auditory cues are primarily used in cross-modal search when the visual cues are degraded or not immediately available. Grohn et al. (2003) surrounded participants with visual stimuli and sound stimuli in order to investigate the process of search for a cross-modal stimulus. Participants used a wand-like device with a magnetic 3D tracker to find as many stimuli, or "gates", as

possible in three minutes. By observing the navigational paths of the wand, they found that auditory cues are used to locate the general area that a stimulus exists in, and then visual cues are used to pinpoint the stimulus' precise location. In the cases where visual perception was degraded because the stimulus appears out of sight or in the far periphery, the auditory cues played the dominant role. It was shown that cross-modal search is quicker than search using a single sense.

A similar experiment undertaken by Hairston et al. (2003a), shows that sound can significantly speed up stimulus location when the visual stimulus is degraded by induced myopia. Myopia is the condition of near-sightedness where light fails to focus upon the eye's retina but rather just in front of it, causing distant objects to be out of focus. The participant was asked to locate visual stimuli using a laser on a yoke which tracked their movements.

Placing the visual stimulus in the periphery of sight appears to yield a particularly malleable visual perception. Shams et al. (2002) present a cross-modal modification of visual perception which involves a phenomenological change in quality. They showed that a singe flash of a white disc in the periphery is viewed as multiple flashes when accompanied with multiple bleeps of sound. In later work they also show that a brief binaural tone creating the sensation of a laterally moving source is capable of inducing perceived visual motion of the static white disc (Shams et al. 2001). In these cases, the placement of the disc in the periphery of sight may be seen as a degradation of visual spatial and temporal acuity. Both judgements have a temporal aspect which also explains the strength of the results, as we know that the auditory system is significantly better at temporal judgements than spatial judgements (Recanzone 2009).

Burr and Alais (2006) questioned whether the ventriloquist effect can be inverted, and investigated the effect of a degraded visual stimulus upon bias. Their experiment required participants to judge the relative change in location of visual blobs and sound clicks. The smallest visual stimulus was 4° across, and was then blurred to create two other stimuli that were 32° across and 64° across. The larger stimuli can then be said to be spatially degraded versions of the smaller stimulus. The auditory stimulus was spatially placed using just a binaural cue conveyed through headphones. They found that in conflict conditions where the visual and auditory stimuli were spatially disparate, people responded to the change in a way that was consistent with the ventriloquist effect (audition biased towards vision) for the small stimulus, but the opposite (vision biased towards audition) for the large stimulus. This demonstrates that spatial auditory information does have the potential to capture spatial visual perception to a certain extent. It may be possible to improve the

effect by using a more detailed spatial audio system and congruent auditory and visual stimuli. In similar work undertaken by Battaglia et al. (2003), five different levels of noise were used to degrade the visual stimulus and thus alter the resulting bias.

The ventriloquist effect has been inverted for a brain damaged individual with Balint's syndrome (Phan et al. 2000). An individual with Balint's syndrome may still have 20/20 vision, but the acuity of their visual system is significantly reduced because they are restricted to only seeing one object at a time. Relative auditory spatial judgements are therefore more reliable than relative visual judgements. Visual spatial perception for this individual was shown to be altered by auditory information in an inversion of the ventriloquist effect.

3.3.3 Other examples

There are examples in the literature that are difficult to fit into either of the above categories. In Hidaka et al.'s paper entitled *Alternation of Sound Location Induces Visual Motion Perception of a Static Object* (2009) the visual position of a static flashing visual stimulus was perceived to alter when accompanied by synchronised auditory information with an alternating spatial location. The visual stimulus was a white bar on a black background and it was accompanied by bursts of white noise. The amount of perceived movement increased as the retinal eccentricity was also increased – a conclusion in agreement with the work on auditory-visual interactions in the visual periphery, discussed in Section 3.3.2 (Shams et al. 2002, 2001). Results from the alternating-sound case were compared with results from the no-sound case and from the static-sound case. The results for the no-sound and static-sound cases were statistically indistinguishable from each other but were statistically distinct from the alternating-sound case. Further analysis confirmed that the effect is unattributable to eye movements, response biases or attentional modulations.

3.4 Cross-modal interactions in depth perception

There is very little literature that explores cross-modal interactions in depth perception, and even less that explores it in a S3D environment. The literature that has been discovered by the author is discussed here. The field is split into two sections: those examples which use a 2D displays (Section 3.4.1), those examples which use

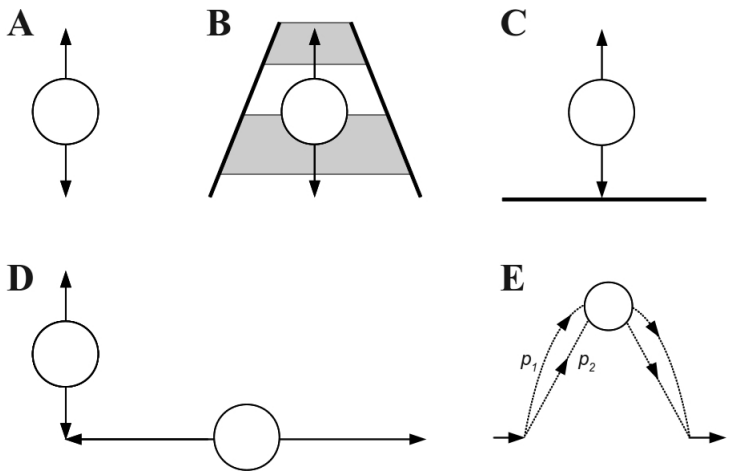

Figure 3.5: Vertical movement in 2D images may be interpreted either as a change in height (bouncing), or as a change in depth (rolling), due to the height-in-the-plane depth cue. Such a bi-stable stimulus was used by Ecker and Heller (2005). **A:** The ball's motion without any references. **B:** The ball's motion with references suggesting a change in depth (rolling). **C:** The ball's motion with references suggesting a change in height (bouncing). **D:** A side view of both paths together. **E:** Showing how a change in the curvature of the balls path is able to suggest a bounce (p_1) or a roll (p_2).

S3D displays (Section 3.4.2), and finally examples which use other environments (Section 3.4.3) including augmented reality and real world environments.

3.4.1 Using 2D displays

Displaying content for visual depth judgements on a 2D screen can be considered a cue degradation because the binocular and physiological cues are not used. Ecker and Heller (2005) showed that sound can significantly alter perception of a moving ball's path on a 2D screen. Consider a ball bouncing upon a surface, and a ball rolling directly away from the viewer. Figure 3.5 shows how these two paths can appear spatially identical when all reference stimuli are removed from the scene. Such a stimulus is therefore bistable (Section 3.3.1). Ecker and Heller used this stimulus to undertake a series of experiments.

In the first experiment the visual stimulus was accompanied by a non-spatialised rolling sound, a non-spatialised bouncing sound or silence. Participants were asked whether the ball appeared to jump or roll. The ball's trajectory was varied using different degrees of curvature in the ball's deviation from a horizontal roll (Figure 3.5E). A curved parabola deviation suggests a bounce as this is the trajectory formed by gravity acting upon the ball as it bounces, whereas a triangular deviation with a sharp angle at the top of its path suggests a roll as this trajectory could

never be achieved in a natural bounce. The results showed that sound was effective in cueing perception of the ball's path even when the visual cue strongly favoured the alternative perception. The fact that the participants' responses didn't unanimously agree with sound shows that observers must have been using both stimuli to make their judgement.

The second experiment aimed to confirm that participants responses were perceptual and not post-perceptual. Instead of indicating the ball's path, participants were asked to indicate the speed of the ball's movement. As shown in Figure 3.5D, the jumping ball covers less distance than the rolling ball so it must be perceived to move slower. Many additional trials were undertaken to ensure that the result is not caused by response bias. Ecker concludes the paper noting that, "Depth perception [...] is accomplished through the interaction of multiple cues."

Meru (1995) looked at whether non-spatialised sound as a depth cue is capable of aiding depth perception in a task-based environment. The participants were asked to pick a target in 3D space located on the surface of a smooth blobby object. The 3D space was presented to the participant using a 2D screen and picking was undertaken using a mouse to control x-y dimension and keys to control the z dimension. While the left mouse button was unclicked a sound indicated the cursor's depth, and while the left mouse button was clicked a sound indicated the targets depth. so by clicking and unclicking the participant could compare the cursor and target sound cues to depth.

Four different types of sound cue were used, labelled by Meru as: "tonal", "musical", "orchestral" and "silence". In the tonal cue possible sound dimensions included: volume, balance, vibrato and pitch. In the music cue tempo and key could also be altered. The orchestral cue used different instrumental sections and their placement within an orchestra to navigate by (e.g. violin was front left, whereas timpani was back right). Not unsurprisingly, users found the tonal cue most effective in cueing depth, but music also performed well. Meru concluded that whilst sound is weaker than vision, performance is best when they are used together.

Motion perception has repeatedly been shown to be subject to strong interactions. Valjamae and Soto-Faraco (2008) looked at sound induced visual flashes (discovered by Shams et al. (2002)) when perceiving time-sampled object motion in depth. They showed that a combination of a slow train of flashes with a rapid train of bleeps leads to sound induced illusory flashes which help fill in visual object motion. The visual stimulus was a white disc changing in size but including no binocular cue for depth. The sound was not 3D – information was instead encoded using pitch. This study suggests that in some cross-modal media, a slower frame

rate may be used.

3.4.2 Using S3D displays

The ventriloquist effect in depth has been investigated by Bowen et al. (2011) using a S3D display to show the visual stimulus and a crosstalk-cancelling headphone-based spatial sound system to play the auditory stimulus. They discovered that bias of the auditory signal towards the visual signal did still occur in depth localisation, but it was significantly less than the bias measured for lateral localisation. Participants were asked to report the perceived depth of a cross-modal stimulus by using a joystick to position a small white square at the perceived depth. However, the mismatch between the *cross-modal* perception and the *visual* response task, may have unfairly biased the cross-modal cue combination towards visual cues. Furthermore, for simple stimuli, such as the small white textureless squares used in this study, visual depth perception is degraded by the lack of cues; in this case just the size and binocular cues were available to use.

Cullen et al. (2012, 2013) explored the effect of sound upon action tasks in depth perception. They found that sound could significantly alter depth perception accuracy, though neither of their studies accurately reproduced the sensation of auditory depth. Their studies required subjects to make judgements concerning the position of a shape floating towards the viewer between a sky plane and a ground plane. This shape could be accompanied by a complex auditory stimulus with some distance effect employed. In their first study a sense of depth was created using frequency and amplitude fall-offs. The subjects were required to determine when the floating object had reached a particular depth indicated by a marker arrow pointing upwards from the ground plane. The second study created the sensation of depth by panning the sound between front and rear speakers in a surround sound system. As the object floated towards the viewer it became invisible at a certain point, though the sound depth kept panning for a while. Subjects were asked to determine on scale of 1-5 how far away the object was when it became invisible. Even though this always happened at the same visual depth, responses were significantly different if the audio was panned from front to rear, rather than just played in front. These studies support the work of the preliminary experiment reported in Chapter 4.

Corrigan et al. (2013) undertook work to determine the allowed differences in depth between audio and visual stimuli in S3D environments. Using pink noise and female speech in two different environments, they undertook a series of tests in which subjects were asked whether the audio depth was either nearer than, further than or

the same distance as a visual stimulus. The study was undertaken using a binaural sound system and a S3D display. It was found that perceived spatial congruence held for significant depth differences between the audio and visual components. They also concluded that this increased with the distance of the visual stimulus. In light of this, it may be appropriate to express the congruence range as a percentage of the reference distance in a similar manner to the MAD. No statistical significance was found with environment or stimulus difference. The large congruence ranges may indicate that high resolutions in audio depth are not needed for audio-visual media.

3.4.3 Other examples

Zhou et al. (2004) investigated the benefit of spatial sound in an augmented reality environment. Their work had four main aims: to assess the impact of spatial sound upon depth perception in a monoscopic augmented reality environment; to study the impact of spatial sound upon task performance and the feeling of "human presence and collaboration"; to better understand the role of spatial sound in human-computer and human-human interactions; and to investigate whether gender can affect the impact of spatial sound in augmented reality environments. Note that the environment used was monoscopic so there were no binocular depth cues incorporated into the experimental stimuli.

The work was broken into two experiments. The first experiment compared depth perception in the vision-only condition with depth perception in the vision-with-spatial-sound condition. Participants were asked to judge the relative depth difference between two telephones placed on a table. These telephones were positioned such that there was no height-in-the-plane depth cue (Section 2.1.3). The second experiment investigated human co-operation to achieve a joint task in a game-based augmented reality environment. It was found that spatial sound both significantly improved and quickened the participants judgement of the relative depth between the two telephones. The visual stimuli in these experiments can be classed as "degraded" due to the lack of the binocular, physiological and height-in-the-plane cues.

Zahorik's paper *Estimating Sound Source Distance With and Without Vision* (2001) concludes that visual capture is not as general for depth as previous literature has suggested. This investigation aimed to re-run work undertaken by Gardner (1968) and extended by Mershon et al. (1980), in which the "Proximity-Image Effect" was investigated. The proximity-image effect refers to the phenomenon of auditory distance being determined by the nearest plausible visual source. In other words, the

term refers to the visual capture (see Section 3.2.3) specifically of auditory distance. Zahorik noted that Gardner's work was undertaken in an anechoic chamber, which we now know reduces the acuity of our auditory depth perception as reverberation is a key cue to auditory depth. He therefore re-ran Gardner's work in a semi-reverberant environment.

Speakers were set in a one-behind-the-other row at eye level facing the listener such that the listener only saw the nearest speaker. Sounds were played out of a speaker and the listener was asked to judge the depth. Gardner found that the vast majority of people said the sound appeared to have the depth of the nearest speaker, which was also the nearest visually rational source for the sound. Zahorik showed that this was not the case in a semi-reverberant environment, though he noted that relative cues may have occurred contributing to the lack of visual capture.

3.5 Conclusions from the literature

Our ability to see visual depth is dependent upon twelve different cues outlined in Table 2.1 (page 13). Conventional 2D displays only make use of the cues in this table that are classed as "pictorial", but S3D displays are also able to convey the stereopsis cue to depth. The human visual system is capable of perceiving very fine differences in stereo depth. Various factors can cause this level of acuity to deteriorate, such as increasing blur or decreasing contrast. Due to both human and technological factors, S3D displays are widely considered to have associated depth budgets. Content designers are often required to suppress their desired range of depth in a S3D image, so as to avoid exceeding the depth budget. If content breaks the limits defined by this depth budget, discomfort and diplopia can occur.

The primary cues available in RAD perception are loudness, reverberation, frequency spectrum and the inter-aural time and level differences. Reverberation and inter-aural differences can also serve as effective absolute cues to depth. The PDH suggests that RAD perception can be reduced to a loudness discrimination task. By doing so, a MAD of 5% of the reference distance is predicted. Despite other cues being available, empirical studies have struggled to match this value, particularly for smaller reference distances.

The literature concludes that audition and vision are capable of influencing each other. Furthermore, auditory visual interactions have been shown to occur in S3D environments. The behaviour of a cross-modal effect depends upon the accuracy of each separate sense's judgements. In the natural world, acuity in visual depth perception is typically much better than acuity in auditory depth perception, giving

rise to the cross-modal model of "visual capture". This model suggests that when vision and audition conflict, vision will generally override audition. Despite this, research has revealed many scenarios where "visual capture" does not apply. In these studies sound has induced some detectable influence upon visual perception. By controlling which depth cues are present in a scenario, cross-modal effects have been found to occur in both conventional 2D displays and S3D displays.

The variability between participants' performance (Hairston et al. 2003b) may be a significant problem when assessing the commercial viability of auditory-visual effects. Many of the papers discussed show that people respond very differently to the auditory-visual interaction, and the preliminary experiment reported in Chapter 4 confirms this. S3D displays are also subject to variability between viewers' perception of content. Different people perceive the binocular cue in different ways, and some cannot perceive the binocular cues at all (this is referred to as "stereo-blindness"). The use of cross-modal effects could therefore draw on similar techniques used in S3D media, in order to account for the variation in how the effects are perceived.

A common difficulty with experiments in this field is distinguishing between perceptual and post-perceptual effects (or response biases). In other words, do the results actually show an auditory-visual interaction occurring, or do the results show an unwanted bias caused by the experimental conditions? The techniques that can be employed to address this problem depend upon the experimental design. Several of the cross-modal studies presented here directly address this threat to the validity of their results, though they do so in different ways (Maeda et al. 2004; Hidaka et al. 2009; Ecker and Heller 2005)

Other possible areas of research can be drawn from this review. In particular, relatively little is known about the acuity of RAD perception in virtual and real environments. This therefore supports our work in answering our third subsidiary research question from Section 1.2: *What is the MAD in our experimental setup?*. Also, the development of a spatialised hearing test seems sensible as the literature points to a high degree of variability between participants' acuity in RAD perception. Several tests are already in place for stereo vision such as the Titmus test (Ohlsson et al. 2001), but no such tests exist for hearing in depth.

Recanzone (2009) states in his review with reference to MLE and Bayes theory, "If these conceptual ideas are true, then it should be the case that auditory stimuli would capture visual stimuli if the visual stimulus was less salient," and goes on to conclude, "This is a technically challenging experiment for normal human subjects, but the available evidence suggests that this could be the case." Although this is

not automatically distinguishable in the real world, the literature suggests that bias of visual stimuli position could occur under certain conditions.

CHAPTER 4

Reviewing a preliminary trial

This chapter reviews an experiment previously conducted by the author that marks the start of the trail of research presented in this thesis. The purpose of the experiment was to observe an audio-visual interaction that could offer an approach to answering our over-arching research question, described in Section 1.2. Due to limited resources and time it is best to view this experiment as a *preliminary* study. However, the results of the experiment provide useful pointers for further work which should seek a more robust analysis of the effect.

The experiment was completed and published (Turner et al. 2011) prior to the author beginning work on this thesis. However, the results play such a pivotal role in understanding the narrative of this thesis, that we have decided to report the experiment in detail as a separate chapter in the literature review. It builds upon our understanding of depth perception in spatial sound and S3D displays presented in Chapter 2, and our understanding of cross-modal interactions presented in Chapter 3.

The design of this experiment was motivated by a desire to extend the depth budget associated with S3D display, which is shown in Figure 1.2. Sound has already been found to impact depth perception in 2D images. We know that the ventriloquist effect is a complicated interplay between audio and visual localisation that can result in audition influencing vision under certain circumstances. Whilst being an improvement on 2D images, depth perception in S3D images still offers a degraded form of depth perception when compared with the real world. We therefore decided to investigate whether there was any possibility of audio being used to influence visual depth perception, and thus provide a means of extending the limited S3D depth budget. The aim of this preliminary experiment was very precise: to observe

auditory depth cues influencing perception of depth in an S3D image. Any attempt to explore the nature and usefulness of the effect was left to further research.

We begin by detailing the experimental method in Section 4.1, including the procedure, equipment, participants, and qualitative data capture. The results are then presented in Section 4.2, and discussed in Section 4.3. The chapter finishes by drawing together conclusions and directions for further work in Section 4.4.

4.1 Method

The design of this experiment began with an experimental hypothesis, from which we also draw the appropriate null hypothesis:

> **Experimental Hypothesis:** A participant will perceive a visual stimulus as nearer if they hear an accompanying auditory stimulus at a nearer depth.
>
> **Null Hypothesis:** Perception of a visual stimulus' depth will not be effected by an accompanying auditory stimulus at a nearer depth.

This section outlines the method used to test these hypotheses. We begin by detailing the experimental procedure in Section 4.1.1, before outlining in Section 4.1.2 the equipment and software used in the experimental setup. In Section 4.1.3 we provide details of the set of participant's who took part in the experiment, before finally outlining in Section 4.1.4 how we also collected qualitative data to support the quantitative data through a post-experiment questionnaire.

4.1.1 Experimental procedure

A 2AFC experimental design was used, in which each participant was asked to make a particular judgement concerning a given scenario. This judgement forced the participants to respond with one of two alternatives. In the null case, where no aspect of the scenario implied a particular response, we could expect the participant to make randomised "guesses". The probability of giving a particular response, when responses are randomised, can be quantified because we have forced the number of alternatives. If a participant were truly guessing in a 2AFC experiment there would be a one-in-two, or 50%, chance that they give each response. So if the participant were to give a significantly different distribution of responses, we could infer that some aspect of the scenario was affecting their judgement.

Participants were asked to sit through a series of short 2AFC tests in a darkened and quiet room. In each test they were sequentially presented with two S3D images of a visual stimulus accompanied by an auditory stimulus. The visual stimulus was a life-sized mobile telephone, with an accompanying telephone ring as the auditory stimulus. For each test the participant was asked to "tell us verbally which image displays the telephone nearest to you." The responses had to be either "the first" or "the second" – "I don't know" or "they were the same" were invalid responses.

The visual depth never changed during the experiment, whilst the auditory depth changed between images in each test. The auditory stimulus could be presented from one of two depths, giving two possible types of tests from the two possible permutations: 'near-then-far', or 'far-then-near'. Participants were not informed of the static visual depth at any point in the experimental procedure and the speaker arrangement was hidden under a thin black cloth. Assuming that the auditory stimulus provided the only cue to a depth change in the images, we could associate a chance performance with the null hypothesis and a better than chance performance with the experimental hypothesis.

Each participant participated in a randomised order of 24 tests where the first four tests were treated as "warm up" tests and discarded. The remaining 20 tests were selected randomly by the software, resulting in eleven 'far-then-near' tests and nine 'near-then-far' tests. Every participant experienced the same order of tests. We discuss the potential threat to the validity of our results posed by the manner in which the tests were randomised in Section 4.3.3.

Upon finishing the experiment, participants were asked to fill out a post-experiment questionnaire. The design of this questionnaire is detailed in Section 4.1.4 and the implications of its results are discussed in Section 4.3.1.

4.1.2 Experimental setup

The decision to use a mobile telephone as the stimulus was an idea taken from the paper by Zhou et al. (2004), who also used telephones as stimuli. It seemed a good choice for a variety of reasons:

- People naturally use auditory information to locate mobile phones.
- The "traditional" telephone ring is a complex multi-frequency sound that will offer more cues to depth than a pure tone (Warren et al. 1958).
- Mobile phones are easy to model graphically.
- They are recognisable and ordinary objects

Figure 4.1: A 2D image of the 3D mobile telephone model, used as the visual stimulus for all cross-modal experimentation reported in this thesis. The image was shown on a white background with a 1:1 scale.

A 2D version of the stimulus is shown in Figure 4.1. A plot of the auditory stimulus' frequency spectrum is shown in Figure 4.2. The auditory stimulus lasted for 3.01 s, and consisted of 1.57 s of rapidly repeated metallic rings followed by 1.44 s of the final ring, dying away to near 0 dB amplitude. Because we felt 3 s was too short a time to present each phone, we played the sound twice for each presentation of the visual stimulus. Each image was therefore displayed for 6 s, and a 1 s silent interval and black screen separated the images in each pair.

The stimulus was modelled using the freely available "Wings3D" software, and textured with the freely available "UVMapper" software. It was then imported into the test software as a Wavefront file (.obj) and displayed using the quad buffering technique on a "Hyundai 46" LCD Monitor Xpol Virtual 3D" TV. The binocular depth of the visual stimulus was controlled using the algorithms outlined by Jones et al. (2001) to create an orthoscopic image (created with a one-to-one mapping between *real space* and *image space*). The software, written in the C programming language, handled the presenting of stimuli and the recording of responses.

To play the auditory stimulus from two different depths, a Logitech Z-5500 Digital stereo loudspeaker system was used with two different stereo sound files. In one file the stimulus was panned completely to the left speaker, whilst in the other the stimulus was panned completely to the right speaker. The left and right speaker were then placed on the participant's median plane (see Figure 2.4), one at each of the required depths. This was because the optimum viewing position for a TV screen is typically taken to be a point on the plane perpendicular and centred to the

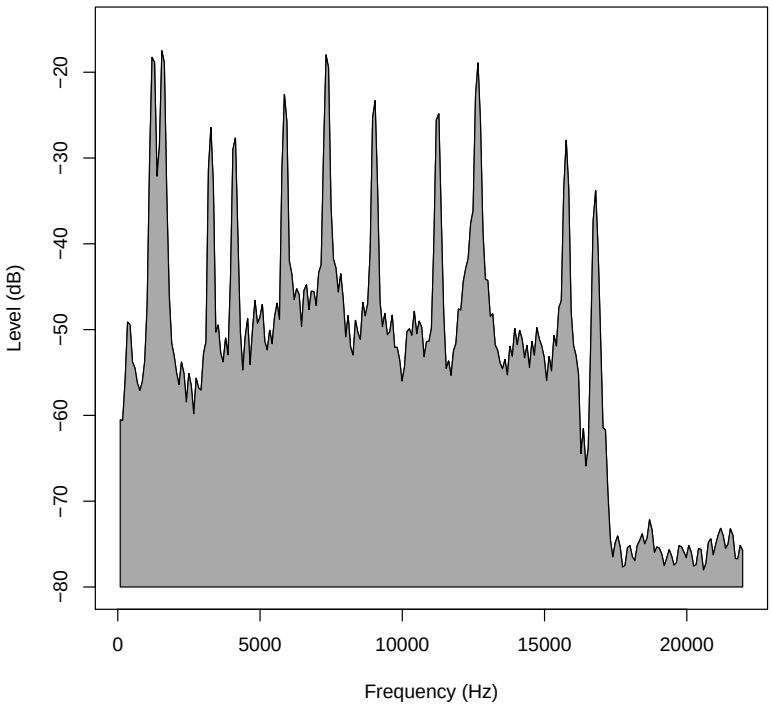

Figure 4.2: The frequency spectrum of the traditional telephone ring we used as our auditory stimulus. The spectrum was calculated in the open source software package "Audacity", using the fast Fourier transform algorithm with a Hanning window of 512 audio samples (the stimulus' sample rate was 44.1 kHz) (Audacity 2015).

screen (THX 2013).

The experimental setup is shown in Figure 4.3. The speakers were offset to avoid the near speaker occluding the far speaker, and thus reduce any interference of the near speaker's body with the far speaker's sound. An offset in height was chosen because humans are worse at distinguishing height differences than lateral differences (Perrott and Saberi 1990). Both loudspeakers were placed under a thin black cloth in order to disguise the purpose of the sound system; participants were not aware of the different loudspeaker depths. The height of the participant's chair was adjusted prior to undertaking the experiment to roughly place their eye level at the same height of as the visual stimulus which was approximately 25cm above the centre of the near loudspeaker.

The distances of 1 m and 25 cm were taken from a previous study undertaken in the laboratory and briefly outlined with this experiment in the paper by Turner et al. (2011). This study showed that in the significant majority of cases, participants could correctly distinguish between two auditory sources 25 cm apart from a distance of a meter.

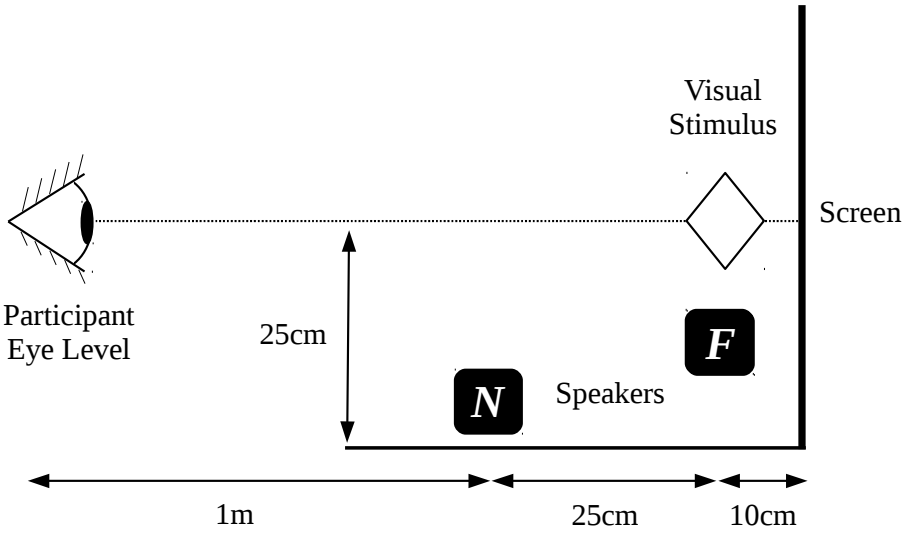

Figure 4.3: The arrangement of equipment in the preliminary trial. The loudspeakers were positioned on the participant's median plane, with the back loudspeaker raised to avoid occlusion of its sound. The participant's eye level was roughly matched to the height of the visual stimulus.

4.1.3 Participants

Fifteen undergraduate students sourced through St John's College, Durham took part in this experiment. All participants were screened for their hearing, visual and stereo acuity. Hearing was checked using the British Society of Hearing Aid Audiologists (BSHAA) online hearing test which confirmed that they could hear tones of 500hz, 1000hz, 2000hz and 4000hz. The participants were required to have at least 20/30 vision, which was tested with a Snellen eye chart. Their stereo acuity was tested using the Titmus test; we required all participants to identify a binocular horizontal disparity of 40 arc-seconds (Ohlsson et al. 2001). We did not collect any further information about the participants, such as their age or their gender. The sample size of 15 was based upon a recommendation from (Moore 1995).

4.1.4 Post-experiment questionnaire

A post-experimental questionnaire was designed to offer some qualitative insight into each participant's results. After asking the participant for their name, it collected responses to the following questions:

- Did you understand the task required of you?
- Did you feel that your answers were a correct representation of what you saw?
- If not, why?

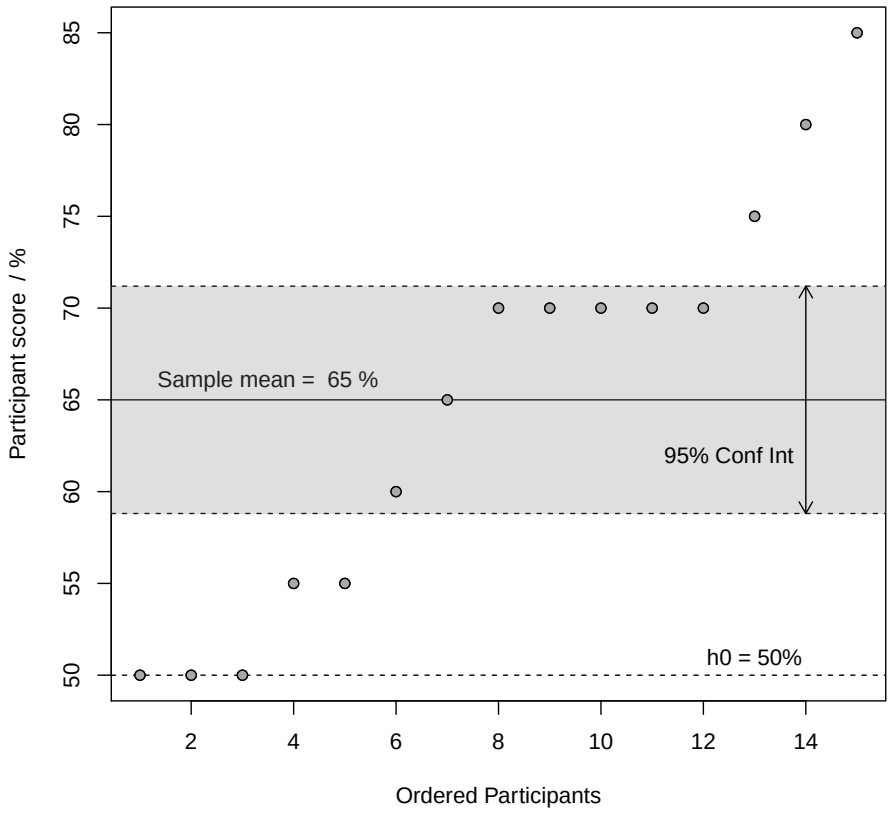

Figure 4.4: A graph plotting each participant's score as a percentage in increasing order. The scores represent the percentage of times they believed the stimulus accompanied by the nearer auditory stimulus was nearer. A Student's T-test gives more than 95% confidence that the data is significantly different from the null hypothesis value of 50%.

- Do you have any other significant comments that may be worth recording, regarding the execution of the test?

The first two questions required the participant to tick a box labelled "yes" or a box labelled "no". For the last two questions the participant was offered a box in which to respond with prose. The responses given, in particular to the second and third question were then used in the analysis of the results to give greater credibility to our conclusions.

4.2 Results

For the purposes of analysis, we define a "correct" response as one in which the participant believes the nearer stimulus is the one accompanied by the nearer auditory stimulus. So in a 'near-then-far' test the correct response would be "the first", and in a 'far-then-near' test the correct response would be "the second". A participant's

score was the percentage of their responses that were correct.

Figure 4.4 shows the distribution of participant scores. The Shapiro-Wilk test for normality tells us that we cannot reject the null hypothesis that the sample is normally distributed. We were therefore able to use a one-sample Student's t-test to determine whether the mean score across participants is different to chance performance – a score of 50%. Figure 4.4 therefore also shows the sample's mean score with a 95% confidence interval. The mean score across all participants was 65% with a standard deviation of 11%.

A two-tailed single sample Student's T-test tells us whether we can reject the null hypothesis that the sample mean and a given value, in this case a chance performance of 50%, are the same. The test yields a p-value of 0.0001, which is significantly smaller than the chosen statistical significance value of 0.05. We could therefore reject the null hypothesis in favour of the experimental hypothesis with a 99.99% level of confidence. Our results do not reflect chance performance.

4.3 Discussion

The results of this experiment appear to be strong, giving more than the 95% confidence required to reject the null hypothesis. It is important to acknowledge the limitations of these results as well as their strengths. We begin in Section 4.3.1 by discussing the results in the context of the post-experiment questionnaire. We then evaluate the experimental procedure and equipment in Section 4.3.2 before finally identifying threats to the validity of our results that should be considered in further work.

4.3.1 Results from the post-experiment questionnaire

Whilst many individuals simply had nothing to comment (often those with higher scores), there were also several who felt that in some of the tests the phone's depth did not change, but in others it did change. Some individuals said that they saw significant depth changes, and one particular participant followed this by saying he, "focused on the edges of the phone." It's important to note that we would expect performance to vary between participants, as the literature revealed that depth perception acuity, and other aspects of the human visual and auditory system, vary significantly. Furthermore, we cannot assume that cross-modal cue combination is consistent across humans.

A few candidates said that their responses were not a correct representation

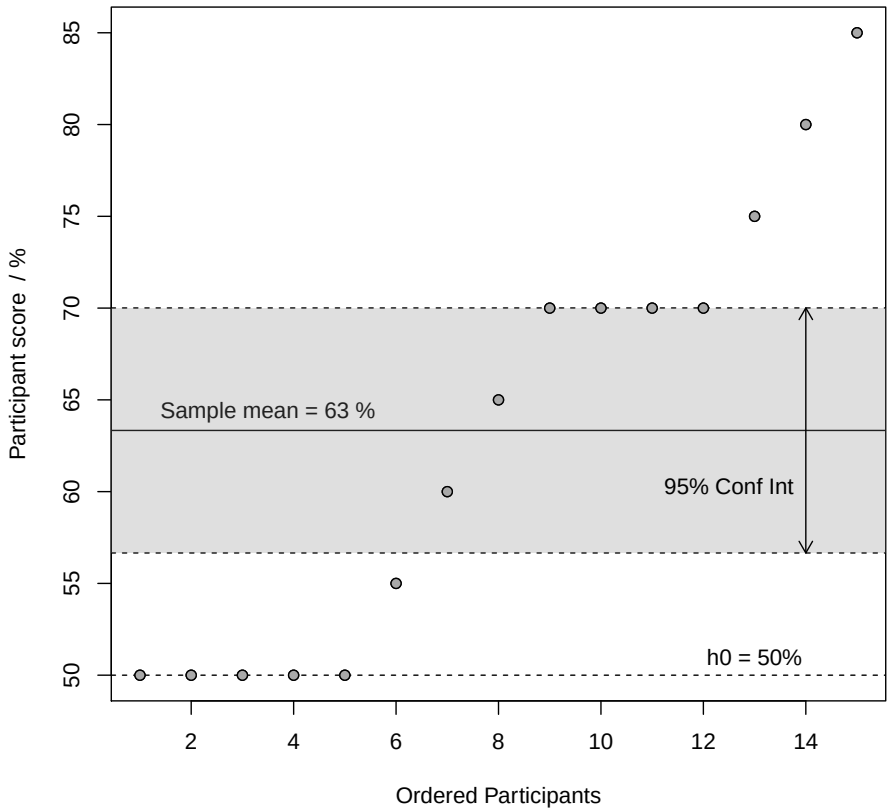

Figure 4.5: If a participant claimed (in their questionnaire) to have not seen any depth changes, then any deviation from 50% in their results is likely to be a response bias. In this graph we have adjusted such participant results to 50%. A Student's T-test still gives more than 95% confidence that the data is significantly different from the null hypothesis value of 50%.

of what they saw. In each case this was qualified with a comment saying that their answers were complete guesses or that they thought the phone's depth never changed. Any deviation in these participant's scores are therefore likely to be biased responses. By this we mean that their results are not indicative of a perceptual effect occurring, but rather a post-processing bias in their response. In other words, having viewed the stimuli and decided that neither phone was nearer than the other, the participants forced response was biased by the auditory depth change that they heard. This threat to the validity of the results is discussed further in Section 4.3.3.

As an attempt to account for this possibility, we decided to re-analyse the data using the information gathered in the questionnaire. We identified participants who said that their responses were not a correct representation of what they saw. The scores for these participants were adjusted to 50% and the Student's T-test re-run to see if the resulting distribution was different from chance. The mean of the adjusted distribution, shown in Figure 4.5, was 62% with a standard deviation of

11%. A T-test yielded a p-value of 0.0008, which was significantly smaller than the chosen statistical significance value of 0.05. We could therefore still reject the null hypothesis in favour of the experimental hypothesis with a 99.92% level of confidence. This suggests that the result still holds even after allowing for some of the results to be response biases.

4.3.2 Evaluation

These results agree with the MLE and Bayesian understanding of the ventriloquist's effect (see Section 3.2.3). The task required participants to identify a depth difference correctly between images, where visually no depth difference occurred. In such a scenario the visual sense is highly unreliable, unlike the auditory sense which does perceive a depth difference. Both MLE and Bayesian understandings of the ventriloquist's effect conclude that, when inferences based upon the visual sense are unreliable to the extent of being random, the visual sense can be easily overridden by some conflicting auditory cue.

The simple experimental design yield statistically significant results, resulting a comparatively simple analysis and interpretation. The simplicity of the experimental design was a strength that could be drawn upon in further work. It does, however, also limit the scope of the effect significantly. We have measured just one data point on the effect's psychometric function (Wichmann and Hill 2001), and we know very little about how external factors influence the effect. Further work should therefore seek to understand more about the effect's scope.

The qualitative data capture should have been more thorough. The post-experimental questionnaire design did not consistently yield insight into how the participant completed the task. In some cases it was clear that they were using vision, or using just audio, or consciously using both. However, some participants provided no prose in the questionnaire, making it difficult to compare their results with those who offered a lot of detail in the questionnaire. An interview, instead of a questionnaire, could have allowed the researcher to probe further and so extract a roughly consistent amount of detail from each participant.

Very little was known about the equipment that was used, or how the placing of the equipment within the experimental setup effect its performance. As we report in Section 6.1.1, finding small commercially available loudspeakers with matched frequency response curves was hard, and we know very little about the frequency response of the loudspeakers used in this experiment. Furthermore, it was assumed that placing both speakers on the median plane and raising the back one above

the near one was the most sensible way of arranging the loudspeakers. We note in Section 6.1.1 that this was probably a bad assumption to make, as it breaks the symmetry of the arrangement causing reverberation (from the desk) of the far loudspeaker's sound to be different to the sound from the near loudspeaker.

In Section 2.2.4 we reviewed the literature concerning the acuity of RAD perception. We discovered that there are only a handful of studies that measure the MAD, and that the value appears to vary with environment, participants, stimuli and experimental method used. The auditory depth difference used in this study was based upon another preliminary study also reported in the paper by Turner et al. (2011). However, that study was undertaken in a different experimental setup, with different participants and using different stimuli. Since the cross-modal effect we are seeking to observe will depend upon the participant's acuity in RAD perception (see Section 3.2.3), the reliability of the results could have been improved significantly by screening participants for RAD perception.

4.3.3 Threats to validity

From our procedure, results and evaluation we can identify a number of threats to the validity of our results that should be addressed in further work. It is possible that participants gave biased responses instead of responses indicative of a fused cross-modal perception. If a participant felt that neither of the allowed responses were correct they may have (consciously or sub-consciously) let their responses be biased by the auditory depth difference. In this case the participants should have felt that their responses were not a correct representation of what they saw. Participants were therefore asked whether this was the case in the post experiment questionnaire. If it was, they were also asked to give details. In Section 4.3.1 we therefore report a re-analysis of the results, after selecting participant's who felt their responses were not a correct representation of what they saw, and correcting their score to a chance 50%. The result is still strongly significant. However, this threat should be more carefully handled in future work. This could be done by improving the qualitative data capture as suggested in the previous section.

As we note in the previous section, our equipment was not calibrated. We know very little about the actual performance of the loudspeakers and display within their experimental setup, or how humans perceive the stimuli they present. There could have been audible differences in the sounds that were due to factors other than the depth difference – namely, the loudspeakers frequency response or the height difference. If these audible differences exaggerated or reduced the sensation depth,

then we would expect the validity of our results' to be compromised. However, the results would still suggest that an effect exists – we just wouldn't be able to draw conclusions concerning how the effect's magnitude depends upon the auditory depth difference between images.

The author acknowledges some mistakes in the execution and design of the experimental procedure. Firstly each participant received the same random order of tests. Ideally a new random order would have been selected for each participant. This ensures that results are not skewed by the order and timing of tests – i.e. participants may be more likely to give a particular response after sitting through a certain sequence of tests. This software error was caused by the random number generator being given the same seed each time the software was run.

Secondly, the number of near-then-far tests did not equal the number of far-then-near tests. This oversight may have caused the results to be skewed by some bias rooted in the type of test being run – i.e. participants are more likely to give a particular response in a near-then-far test than in a far-then-near test.

None of these procedural mistakes undermine the experiment's core result, given its preliminary nature. There is no intuitive reason to suppose that a participant's response to a test was biased by the type of the test, or by the type of tests that preceded it. Given the large effect size and high level of statistical confidence that has been observed, it still seems appropriate that this experiment's result guide and influence future research, to seek a more reliable confirmation of the effect's existence.

4.4 Conclusions

Participants undertook a series of tests. In each test they were asked which of two mobile telephones, viewed consecutively, appeared nearer to them. The visual depth of the telephones were the same, but the auditory depth of the accompanying telephone ring varied. This auditory component of the cross-modal stimulus could either appear at the same depth as the visual component, or at 25cm in front of the visual component. The 2AFC paradigm required participants to respond with either "the first" or "the second" – "I don't know" was not a valid response.

The results show that across all participants a mean 65% of responses said that the phone accompanied by the nearer auditory stimulus was nearer. A one-sample two-tailed Student's T-Test gives us more than 95% confidence that this response is different to 50% chance. If neither the visual or auditory depth of the cross-modal stimulus changed between the two viewings, then there would be no cue to an answer

so we would expect chance performance. We therefore conclude that auditory depth is capable of altering perception of depth in S3D displays, and could possible have the potential to extend the range of depth that is comfortable to view on a S3D display, without requiring more S3D depth budget.

We have acknowledged a number of threats to the validity of these results that have arisen due to the preliminary nature of the experiment. None of these threats stop us from drawing the conclusion that a cross-modal effect exists that is worth further exploring. However, they do provide pointers for how further work should proceed. In particular, the simple experimental design should be developed to provide an insight into the effect's scope and external influencing factors. Further experimentation should use calibrated equipment, a better means of capturing qualitative data and a screening test for RAD perception.

This result should therefore be approached with some caution; it is important to acknowledge the preliminary nature of this experiment. However, it does suggest that a deeper and more thorough study would be valuable, in order to explore the nature and commercial viability of the effect. This experiment provides the starting point for the trail of research presented in this Thesis.

CHAPTER 5

Evaluating subjective responses to quality-controlled S3D depth

We begin our experimental work with a study assessing the subjective value of quality-controlling the binocular depth cue in S3D media. More specifically, we evaluate the subjective responses of audiences to viewing high-quality S3D media created using the quality-control algorithms outlined by Jones et al. (2001) and Holliman (2004). These algorithms specify how to map scene depth to a given display's depth budget. As discussed in Section 1.2, much of the research presented in this thesis is motivated by the desire to extend the depth budget. It therefore seems wise to assess the subjective value of restricting binocular cues to a given depth budget, before further work ensues. More generally, by showing there is subjective value in the quality-control of the binocular cue, we motivate research concerning the quality-control of other depth cues, including auditory cues.

We have used a pre-test post-test quasi-experimental design to measure changes in the audiences' subjective impressions of S3D media. In our experience, films create using quality-control algorithms, such as those detailed by Jones et al. (2001) and Holliman (2004), typically elicit positive responses on technical quality from both expert and non-expert audiences alike. This chapter seeks to answer the thesis' second research question, reported in Section 1.2, and further explore the scope of our results through replications of the original experiment. In this chapter we therefore address the following questions:

1. Does viewing S3D content with quality-controlled binocular cues create measurable positive changes in the audience's subjective attitudes towards S3D

media? (The second Thesis research question)

2. Are the measured changes repeatable on displays with different sizes?
3. Can we replicate these results outside our laboratory?

We have addressed these research questions through an audience-centred study that gathers self-report responses and written comments from all audience members. Furthermore, this study incorporates an original experiment and a number of differentiated replications. As Lindsay and Ehrenberg (1993) write, replication is a crucial aspect of the scientific method that is perhaps often overlooked when evaluating subjective impressions. The differentiated replications we report here, in which we vary the film, display and site used, offer insight into how generalisable our results are.

Hassenzahl and Tractinsky (2006) tells us that the study of *user-experience* (and likewise *audience-experience*) is concerned with technologies that fulfil more than just instrumental needs. It is important to recognise the subjective, situated, complex and dynamic encounter that occurs between the user and the technology. As such, the user experience arises from characteristics of their internal state, the designed system and the context of interaction. Creating a good S3D film viewing experience must therefore bring together the right film, display, audience and viewing environment.

For the film content, we used two short 3D films entitled *Cosmic Cookery* and *Cosmic Origins*. These were developed by a collaboration between Physicists and Computer Scientists at Durham University, and produced using algorithms that quality-control the binocular depth (Holliman et al. 2006; Holliman 2010). Both films illustrate how theories of dark matter have influenced the formation and movement of stars and galaxies. They were initially created to be shown at the annual Royal Society's Summer Science Exhibition in London in 2005 and 2009 respectively, and have consistently received positive informal feedback from large, non-expert audiences. *Cosmic Cookery* won first prize in the national VizNet Visualisation Showcase 2006, whilst *Cosmic Origins* was winner of the "Best Computer Graphics Film Award" at the Stereoscopic Displays and Applications Conference 2010, San Jose, California.

For the display technology, we began by using the large 160" projected display that the films were designed to be viewed upon. Once we had used this display to establish that high quality films can have a measurable effect on audiences, we then investigated whether our results were repeatable on a 50" TV sized screen. Our displays were carefully selected for their low cross-talk and high resolution.

For each round of experimentation, the participants were recruited from the local academic staff and student communities. All participants were screened for stereo acuity prior to their involvement in the study. The first rounds of experimentation were undertaken in a laboratory at Durham University (UK) and then, once we had established a suitable 60" TV sized platform, we investigated whether our results were repeatable at other sites. First, we took the study to another UK site, York, and then we moved to an international location in Twente, The Netherlands. We sought to keep the environment, specifically brightness, sound volume and viewing angle, as similar as possible across all experimentation.

In the above ways, we designed an experiment that met the requirements specified by Hassenzahl and Tractinsky (2006) for content, display, audience and environment. Our report of this experiment continues with a summary of the methodology adopted (Section 5.1), before detailing the specific setup and results of the experiment (Section 5.2) and replications (Sections 5.3, 5.4 and 5.5). We discuss the results in Section 5.6 and draw together conclusions and further avenues for research in Section 5.7.

5.1 Method

In this section, we outline the general method used to answer our research questions. This begins with the experimental design in Section 5.1.1, followed by the questionnaire design in Section 5.1.2. We then give details of the participants recruited for our experiment in Section 5.1.3 and consider the statistical design of the experiment in Section 5.1.4. This section finishes with a summary of the final general experimental procedure in Section 5.1.5. Further details of our methodology, such as the the display and location of each replication, are discussed in later sections.

5.1.1 The Experimental Design

As this study is concerned with identifying a change in attitude to 3D films before and after viewing a high quality 3D film, we adopted a one group pre-test post-test quasi-experimental design (Shadish et al. 2002). This design is simple, effective for identifying change, and widely used by researchers. Participants are tested before and after an intervention in order to identify any change in test responses. These response changes are then assumed to be caused by the intervention. In this study the intervention is a 3D film and the tests are questionnaires seeking insight into the participant's attitude towards 3D and awareness of the film's content.

ID	**Location**	**Display**	**Film**	**Coding**
D-LP-CC	Durham	160" Projection	Cosmic Cookery	Original SD Resolution
D-LP-CO	Durham	160" Projection	Cosmic Origins	Original HD Resolution
D-TV-CC	Durham	50" TV	Cosmic Cookery	Blu-Ray SD resolution
D-TV-CO-HFR	Durham	50" TV	Cosmic Origins	Blu-Ray Higher frame rate
D-TV-CO-HR	Durham	50" TV	Cosmic Origins	Blu-Ray Higher resolution
D-SP-CC	Durham	50" Projection	Cosmic Cookery	Blu-Ray SD resolution
D-SP-CO-HFR	Durham	50" Projection	Cosmic Origins	Blu-Ray Higher frame rate
Y-SP-CC	York	50" Projection	Cosmic Cookery	Blu-Ray SD resolution
Y-SP-CO-HFR	York	50" Projection	Cosmic Origins	Blu-Ray Higher frame rate
T-SP-CC	Twente	50" Projection	Cosmic Cookery	Blu-Ray SD resolution
T-SP-CO-HFR	Twente	50" Projection	Cosmic Origins	Blu-Ray Higher frame rate

Table 5.1: All the interventions evaluated are shown here. The IDs are of the form Location-Display-Film-Coding. Where for location: D = Durham, Y = York, T = Twente, for display type: LP = 160" projection, TV = 50" TV, SP = 50" projection, for film name: CO = *Cosmic Origins*, CC = *Cosmic Cookery* and for coding HR = high resolution, HFR = high frame rate. The first group of two interventions were our first evaluations on the large screen, the second group of three interventions were our evaluations of the 50" TV and the different possible BlueRay codings for CO, the final group of six interventions were those we settled on as suitable for evaluations at all three geographic locations using the 50" projection display.

In order to protect the validity of the results the design needs to minimise the effect of any external variables that might impact upon the results. For example, boredom and tiredness, or loss of concentration may occur if the duration of the intervention is too long. The films we presented did not last more than eight minutes, keeping the intervention short. In addition, we minimised the effect of other possible external variables by running interventions in a blacked out room and monitoring image brightness and audio volume levels. The test questionnaires run before and after the intervention were kept simple and easy to complete. The study was approved by the ethics committee of the School of Engineering and Computing Sciences, Durham University.

We used differentiated replications to investigate how varying key aspects of the intervention affected the audience's responses. Details of each intervention are given in Table 5.1 and are discussed below.

We tested responses to two films, *Cosmic Cookery* and *Cosmic Origins*, in order to determine whether the measures we used were stable across similar but different films. Both films were created at Durham University using similar depth budget controls and similar content, but the music, narration and images make them distinctly different films. Details of the original experiment, in which both these films were shown on the 160" large screen projected display, are given in Section 5.2.

We also sought insight into the potential effect from response variance caused by the display technology. In particular we compared results from the large screen projected display (160") with those from using a TV and a small screen projected display (both 50"). Again, we were interested in exploring whether audience responses changed across different viewing platforms. The differentiated replications that used small screens are detailed in Sections 5.3 and 5.4.

Finally, we investigated whether audience responses would vary at locations outside our laboratory in Durham. To do this we ran experiments at the University of York (UK) and overseas at the University of Twente (NL). The display technology used at these locations was the best performing TV sized display from the experiments run in Durham. The details of these differentiated replications are given in Section 5.5.

5.1.2 Questionnaires

The preliminary and post-intervention tests were performed using paper questionnaires that began with the same five questions:

1. Please rate your impression of the viewing experience 3D films can provide.
2. Please rate your impression of how well 3D films can convey complex information.
3. Please rate your impression of how comfortable you think viewing 3D films can be.
4. Please rate your impression of how natural the sensation produced by viewing 3D films can be.
5. Please rate your knowledge of how galaxies are made.

Questions 1 and 4 are included with reference to the study by Seuntiëns et al. (2005) and Question 3 with reference to the literature concerning visual discomfort in S3D media (Lambooij et al. 2009; Ukai and Howarth 2008; Nojiri et al. 2004). Questions 2 and 5 were added to gather evidence about whether S3D media is a good way of presenting complex, cosmological data. Another question was included in each test, in the preliminary questionnaire this was a closed multiple choice question:

- How would you rate your experience of 3D films? None/Limited/Good/Expert

Whereas in the post-intervention questionnaire it was an open question that included a request for comments:

- Please write any comments or observations you have about 3D films below.

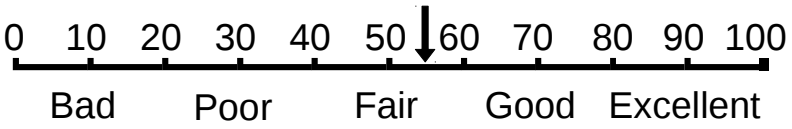

Figure 5.1: The response scale used by subjects to answer the first 5 questions in each questionnaire. Subjects were asked to indicate their response with an arrow as shown. The print size of this scale was 10 cm long to meet the specifications outlined in the ITU-R Recommendation BT.500-12 (2009)

Responses to the first 5 questions were provided by asking participants to draw an arrow on a Likert scale as shown in Figure 5.1. These scales were designed to meet the recommendations described by the ITU (ITU-R Recommendation BT.500-12 2009). The indicated values were read off the scales by human eye and recorded in data sheets as integers. The small random error incurred in doing this can be estimated as ± 1.

5.1.3 Participants

The participants were recruited from the academic communities where each round of experimentation was performed. The majority of participants were undergraduate or postgraduate students, though some members of staff also took part. In total, 176 people took part in the study of which 67% were male and 33% female. The ages ranged from 18 to 57, with a median age of 23 and an inter-quartile range from 20 to 26.

As in the preliminary experiment, all participants were required to give a complete set of responses to the Stereo Titmus Test before their participation (discussed in Section 4.1.3). Participants who failed to score 100% correct in this test were informed that their results "may not contribute towards the project conclusions" and were invited to choose whether or not to continue their participation, in case their results become of use at a later time. All 56 participants in this situation chose to continue their participation. The study took approximately 30 minutes, for which participants were each paid an honorariam of £5, or €5 in the case of our overseas experiments.

We gathered data until we had at least 15 participants who had passed the screening test in each sample. This sample size of at least 15 is a recommendation from Moore (1995) based upon a series of large computational studies (Pearson and Please 1975; Posten 1979). The number of participants who could simultane-

ously take part in each viewing was dependent upon the screen size of the display technology used.

5.1.4 Statistical design

Paired Student's t-tests were used to identify whether there was any significant difference between preliminary and post-viewing questionnaire scores across each sample. Student t-tests assume normally distributed samples, so the Shapiro-Wilk test for normality was used to check this. In the case of a sample failing the normality test, the Wilcoxon signed rank test was used instead of the t-test, with the median and inter-quartile range used in place of the mean and standard deviation. In the case of no response difference being identified, two one sided t-tests were used to check for equivalence against the null value of zero. All significance testing used an alpha criterion of 0.05 to indicate a "strongly significant result" and 0.10 to indicate a "weakly significant result".

Analysis of variance (ANOVA) was used to assess the differences between the experiment and replications. Although ANOVA also assumes a normal distribution, it is reputedly insensitive to data normality (Glass et al. 1972; Lix et al. 1996). We therefore use ANOVA to test between all samples, even where some samples fail the Shapiro-Wilk test for normality.

5.1.5 Procedure

The procedure required participants to fill out four forms on a clip board. It was decided that the participants should not be allowed to refer to their preliminary responses whilst giving their post-viewing responses. This is because we were seeking a change in attitude towards S3D films, not a self-referenced consideration of the specific film they had viewed. The preliminary questionnaires were therefore collected prior to watching the film and completing the post-viewing questionnaires. The final procedure for each viewing involved the following distinct stages:

1. Welcome participants and outline the procedure to them.
2. Ask them to read and fill out the instructions and consent form.
3. Ask participants to complete the stereo Titmus test by reading and filling out a second form in conjunction with viewing the appropriate images.
4. Ask participants to fill out the preliminary questionnaire and then collect all forms in.

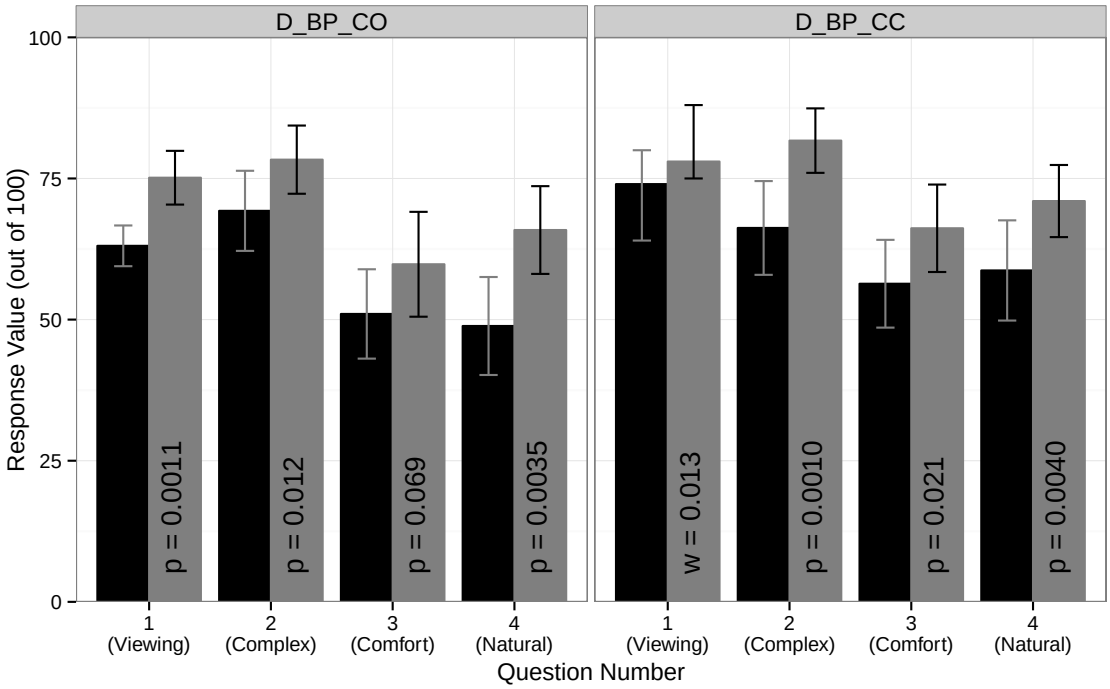

Figure 5.2: The results of the original experiment using the big screen projection. Depending upon the result of the Shapiro-Wilk test for normality, the black bars indicate the mean or median preliminary response, whilst the grey bars indicate the mean or median post-viewing response and the errors bars denote the standard deviation or inter-quartile range across the sample. The result of a paired Student's t-test, or Wilcoxon signed rank test is also shown for each question. In the cases where the Shapiro-Wilk test failed and ranked statistics are used, the statistical test result is labelled with a w instead of a p.

5. Hand out appropriate glasses and show participants a random dot stereogram to ensure that their glasses are working.
6. Switch lights off and show them the film.
7. Switch the lights on, hand out the post viewing questionnaire and ask them to fill it out.
8. Pay them for their time.

5.2 Experiment: big screen projection

This original experiment used the display technology that we hypothesised was most likely to give positive results — our big screen, low crosstalk, active shutter glasses display system. If an effect was found here for both *Cosmic Origins* and *Cosmic Cookery* we would then have the motivation to consider the other factors of interest.

5.2.1 Experimental setup

The setup for this experiment consisted of:

- Christe Mirage 3D 1080 HD digital light processing (DLP) projector
- Rear projection screen 3.50 m wide and 1.97 m high
- Virtalis Activeworks 3D Glasses
- JBL EON1500 stereo speaker system

Participants sat in a row centred on the centre of the screen and at a distance such that the central viewer received a 40° viewing angle as recommended by THX (2013). Five participants completed the experiment at a time. In total 19 participants took part in the *Cosmic Origins* viewings, of which 4 failed the screening test, and 21 participants took part in the *Cosmic Cookery* viewings, of which 4 failed the screening test. These participants were recruited primarily through the first year undergraduate engineering course, resulting in an age distribution of 18-32, with a median of 19 and an inter-quartile range of 18-19.

Brightness was measured using a Sekonic L-758 Cine light meter. The receptor was placed behind a "lens" of the active S3D glasses and positioned at approximately the viewing position, with the room darkened as for viewing. A stereo black image pair was shown and the luminance reading through the glasses was found to be too small to detect, meaning that it was less than 0.63 lux. The luminance of a stereo white image pair was found to be 1.3 lux through the glasses.

The maximum volume during the opening few seconds of the narration was measured so that it could be matched in the other experiments. This was done using a decibel meter on a tripod positioned at approximately the central viewer's listening position. The maximum volume for the opening phrase of narration was set at 73.9 dB.

The content was shown at full original-edit quality: *Cosmic Origins* in frame packed 1920x1080 HD with a frame rate of 30 fps and *Cosmic Cookery* in frame packed 1024x768 with a frame rate of 25 fps.

Question 5 (knowledge) was not included in the questionnaires used in this original phase of the study, though we have no reason to believe that this would affect the results in any significant manner.

5.2.2 Results

Figure 5.2 shows summarised results for this experiment including both the *Cosmic Origins* and *Cosmic Cookery* films. For the normally distributed data, a mean pre-

liminary response is indicated by the black bar, whilst a mean post-viewing response is indicated by the grey bar, and the error bars denote the standard deviation.

Responses to each question for each film passed the Shapiro-Wilk test for normality with a significance criterion of 0.05 in all but one of the eight cases. The post-viewing responses to Question 1 (viewing experience) in the *Cosmic Cookery* data yielded a p-value of 0.0107 for the Shapiro-Wilk test for normality. This is less than our significance criterion, meaning that we need to reject the null hypothesis that the data is normally distributed. We therefore display ranked statistics (median and inter-quartile range) for this question in Figure 5.2, and used a Wilcoxon Signed Rank Test instead of a Student's t-test to compare preliminary and post-viewing responses. The result of this test is labelled with a w in Figure 5.2 and is smaller than our alpha significance criterion, allowing us to conclude that the response difference is significantly different from zero.

In all cases, except Question 3 (comfort) for *Cosmic Origins*, we concluded that the difference between preliminary and post-viewing responses is strongly significant - the Student's paired t-test or Wilcoxon signed rank test yields a p-value less than our chosen significance criterion of 0.05. The t-test p-value for Question 3 (comfort) is 0.069, which is less than 0.1 so we still conclude that it is weakly significant.

The results from this experiment suggest that viewing both *Cosmic Origins* and *Cosmic Cookery* can have a significant effect upon a viewer attitude towards S3D films.

5.3 Replication 1: television display

The effect observed in the original experiment provided motivation for further study seeking significance in other displays. This differentiated replication investigated whether a similar effect is found in Television (TV) displays, which are smaller and make use of very different S3D technologies.

5.3.1 Experimental setup

The following equipment was used:

- Panasonic TXP50ST50B Plasma Active shutter Glasses 3D TV.
- Glasses
- Sony BDP-5780 Blu-ray disc player

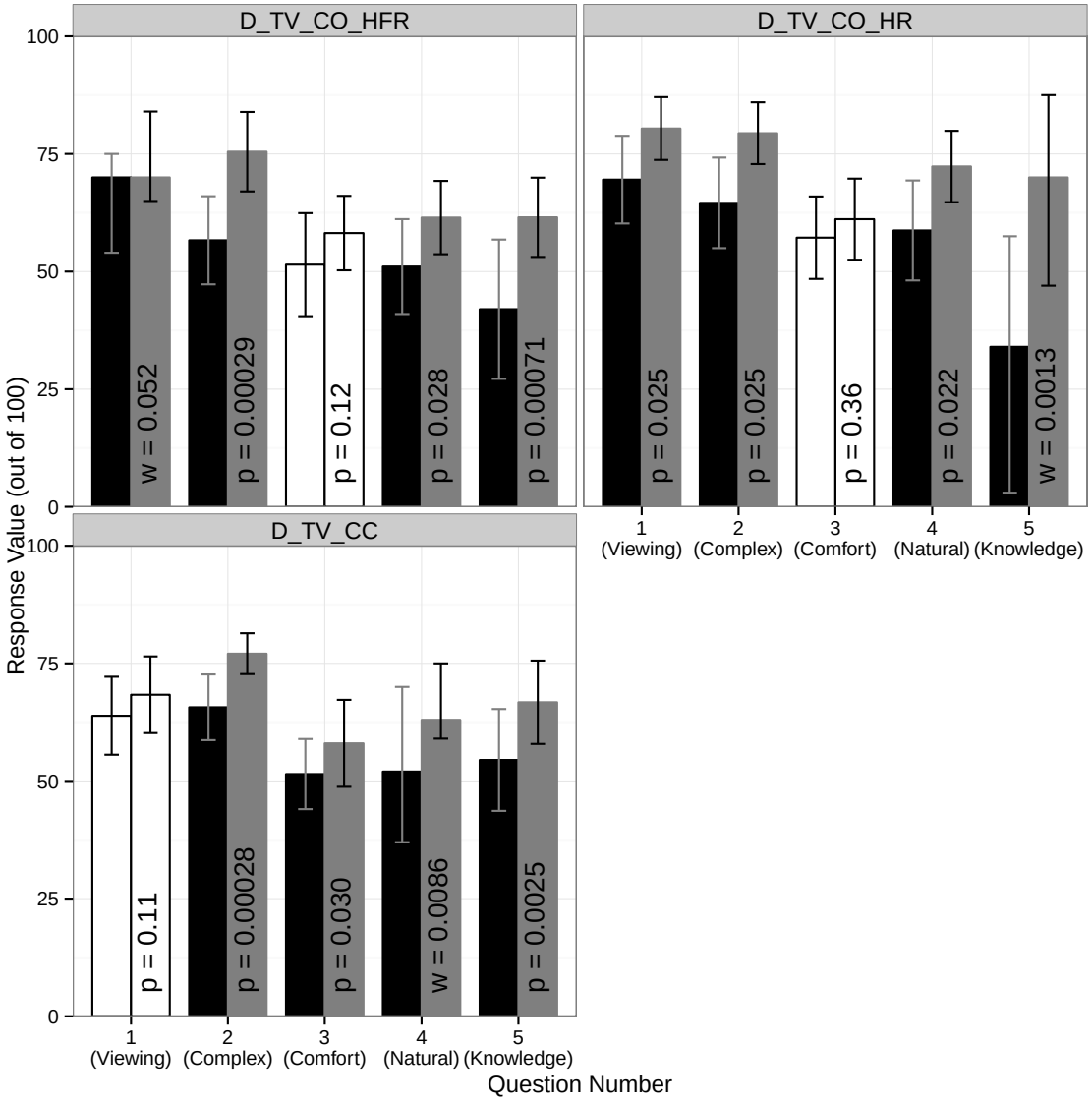

Figure 5.3: The results of the differentiated replication using the TV display. Depending upon the result of the Shapiro-Wilk test for normality, the black bars indicate the mean or median preliminary response, whilst the grey bars indicate the mean or median post-viewing response and the errors bars denote the standard deviation or inter-quartile range across the sample. The result of a paired Student's t-test, or Wilcoxon signed rank test is also shown for each question. In the cases where the Shapiro-Wilk test failed and ranked statistics are used, the statistical test result is labelled with a w instead of a p. The white bars indicate questions where the statistical test failed to find a significant difference between preliminary and post viewing responses (the result did not meet our alpha significance critereon of 0.1).

The films were played using a 3D Blu-ray disc and player, in order to keep the equipment portable for later use at external sites. As a consequence the films could not be shown in original-edit quality, so we experimented with several encodings to determine the best approach. Using the Sony Vegas software package, we re-encoded the video to the Multiple View Coding format, which limited us to a frame rate of 27 fps with full 1080p HD or 60i fps with 720p HD. The conversion from 30 fps to 27 fps was not smooth and caused noticeable jerkiness when viewing. The conversion from 30 fps to 60i fps was smooth, but the loss in resolution was noticeable. We were unsure which encoding would be preferred, so we ran separate viewings for each of 3 different films: 720p HD *Cosmic Origins* with a Higher Frame Rate (HFR) of 60i fps, 27 fps *Cosmic Origins* with a Higher Resolution (HR) of 1080p HD and 50i fps *Cosmic Cookery* with a resolution of 1280x720 pixels. *Cosmic Cookery* suffered a small loss in resolution as the 1024x768 image was mapped onto a 1280x720 image. The original aspect ratio was maintained, resulting in black space down the left and the right hand sides.

As in the original experiment, participants sat in a row centred on the centre of the screen and at a distance such that the central viewer received a 40° viewing angle. This time, due to the smaller screen size, only three participants could be accomodated in each viewing. The TV was set upon a desk in front of the participants. Twenty participants took part in the *Cosmic Origins* HFR viewings, of which 3 failed the screening test, whilst 17 participants took part in the *Cosmic Origins* HR viewings, of which 2 failed the screening test. Sixteen participants took part in the *Cosmic Cookery* viewings of which 1 failed the screening test. These participants were primarily recruited from the Chemistry, Engineering and Mathematics postgraduate groups, resulting in an age distribution of 19-37, with a median of 24 and an inter-quartile range of 22-26. The gender balance was 53% male to 47% female.

Brightness was measured using the same technique as in Section 5.2.1. The black screen luminance was again less than 0.63 lux whilst the white screen luminance was 1.6 lux. The volume level at the viewer's listening position was matched to the original experiment using a decibel meter.

5.3.2 Results

The results of this replication are shown in Figure 5.3. Three cases failed the Shapiro-Wilk test for normality, and a Wilcoxon signed rank test was used in place of a Student's t-test to account for this. The preliminary responses to Question 1 (view-

ing experience) in the *Cosmic Origins* HFR data yielded a Shapiro-Wilk p-value of 0.0359, whilst the post-viewing responses to Question 5 (knowledge) in the *Cosmic Origins* HR data yielded a Shapiro-Wilk p-value of 0.0291. The *Cosmic Cookery* post-viewing responses to Question 4 (naturalness) yielded a Shapiro-Wilk p-value of 0.0129.

The three cases that failed the response difference significance tests are coloured white in Figure 5.3: Question 3 (comfort) for both *Cosmic Origins* films and Question 1 (viewing experience) for *Cosmic Cookery*. None of these cases can be considered weakly significant. It is important to note that a failed significance test does not allow us to conclude that no effect exists; instead, it tells us whether we can reject the possibility that no effect exists. However, equivalence tests do allow us to conclude that the mean response difference was equal to zero implying no effect occurred. The significance criterion was taken as 0.05 and a conservative region of equivalence of ± 5 points was chosen, giving an interval width of 10 corresponding to the minor interval on the response scale in Figure 5.1. No significant result was found. These three cases are therefore null results - they neither support nor oppose the hypothesis that a measurable change in response occurred whilst watching the film. Further discussion is presented in Section 5.6.

The experiments undertaken with a TV display have yielded a number of significant results suggesting positive changes in response occurred when viewing the films. However, due to the three null results, the effects do not appear to be as strong as those from the big screen projected display. In Section 5.6 we discuss what might have caused these failed significance tests and how they sit alongside the results from the original experiment.

5.4 Replication 2: small screen projection

The TV display gave results with a weaker set of effects than the original experiment. We noticed that our TV display had significantly higher crosstalk than the original projection display – a result of the different imaging technology being used in the display (plasma screen vs DLP projection). This differentiated replication extends the work outlined in the previous section by matching the TV display size using the same DLP projection technology from the original experiment.

5.4.1 Experimental setup

This experiment used the following equipment:

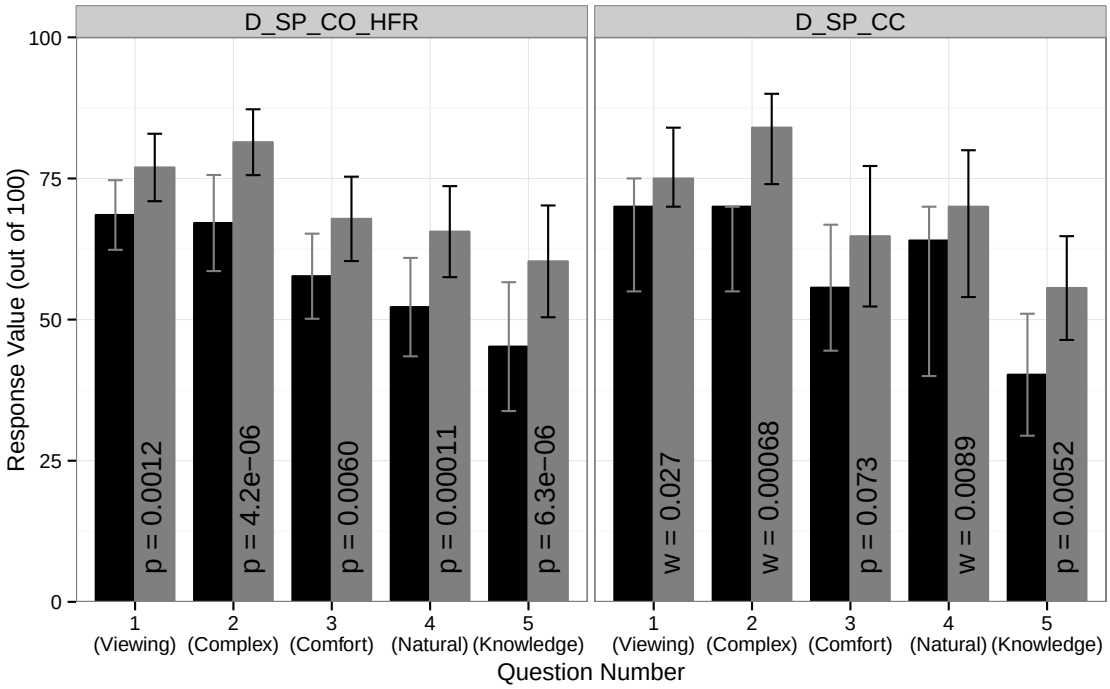

Figure 5.4: The results of the differentiated replication using the small screen projection. Depending upon the result of the Shapiro-Wilk test for normality, the black bars indicate the mean or median preliminary response, whilst the grey bars indicate the mean or median post-viewing response and the errors bars denote the standard deviation or inter-quartile range across the sample. The result of a paired Student's t-test, or Wilcoxon signed rank test is also shown for each question. In the cases where the Shapiro-Wilk test failed and ranked statistics are used, the statistical test result is labelled with a w instead of a p.

- Optoma HD33-B DLP portable 3D projector
- Optoma ZF2100 glasses and emitter
- Polk-audio Silicon Graphics stereo loudspeaker pair
- Sony BDP-5780 Blu-ray disc player

The films were played using the 3D Blu-ray disc and player, but this time the HR version of *Cosmic Origins* was not shown because in the TV viewings. This is because it consistently yielded response differences the were less significant than the HFR version of *Cosmic Origins* and attracted negative comments from the audience in written feedback.

As in the previous replication, three participants at a time sat in a row centred on the centre of the screen and at a distance such that the central viewer received a 40° viewing angle. Twenty-two participants took part in the *Cosmic Origins* viewings, of which three failed the screening test, and 21 participants took part in the *Cosmic*

Cookery viewings of which four failed the screening test. These participants were primarily recruited through the second year undergraduate engineering course and a Durham college's postgraduate group, resulting in an age distribution of 19-35, with a median of 21 and an inter-quartile range of 20-23. The gender balance was 61% male to 39% female.

Brightness was measured using the same technique as in Section 5.2.1. The black screen luminance was again less than 0.63 lux whilst the white screen luminance for this screen was notably brighter at 9.3 lux. The volume level at the viewer's listening position was matched to the previous experimentation using a decibel meter.

5.4.2 Results

Figure 5.4 shows the results from the small screen projection viewings. All of the data sets taken using the *Cosmic Origins* film passed the Shapiro-Wilk tests for normality, whilst three questions from the *Cosmic Cookery* data failed the test. Both preliminary and post-viewing responses in Question 1 (viewing experience) and Question 2 (complex information) failed with respective p-values of 0.0211 and 0.00695 in Question 1 and 0.00509 and 0.00242 in Question 2. The Shapiro-Wilk test also failed in Question 4 (naturalness) with preliminary responses yielding a p-value of 0.023.

The only significance test to yield a result that was not strongly significant is Question 3 (comfort) for the *Cosmic Cookery* data. The Student's t-test gives a p-value of 0.0727, which indicates a weakly significant effect. These results are therefore similar to the big screen results, despite the significant amount of compression applied to the films so that they could be played from a Blu-ray disc. The data also shows that a more significant effect occurred than when watching the films on the TV display. As a result we chose the small screen projected display to evaluate response differences outside our laboratory at Durham.

5.5 Replications 3 & 4: York and Twente

We next sought to demonstrate that our results are repeatable beyond our own laboratory and the academic community where the films were created. This was done by taking the best performing portable display - the small screen projection - first to another site in the UK, and then further afield to an international site in the Netherlands.

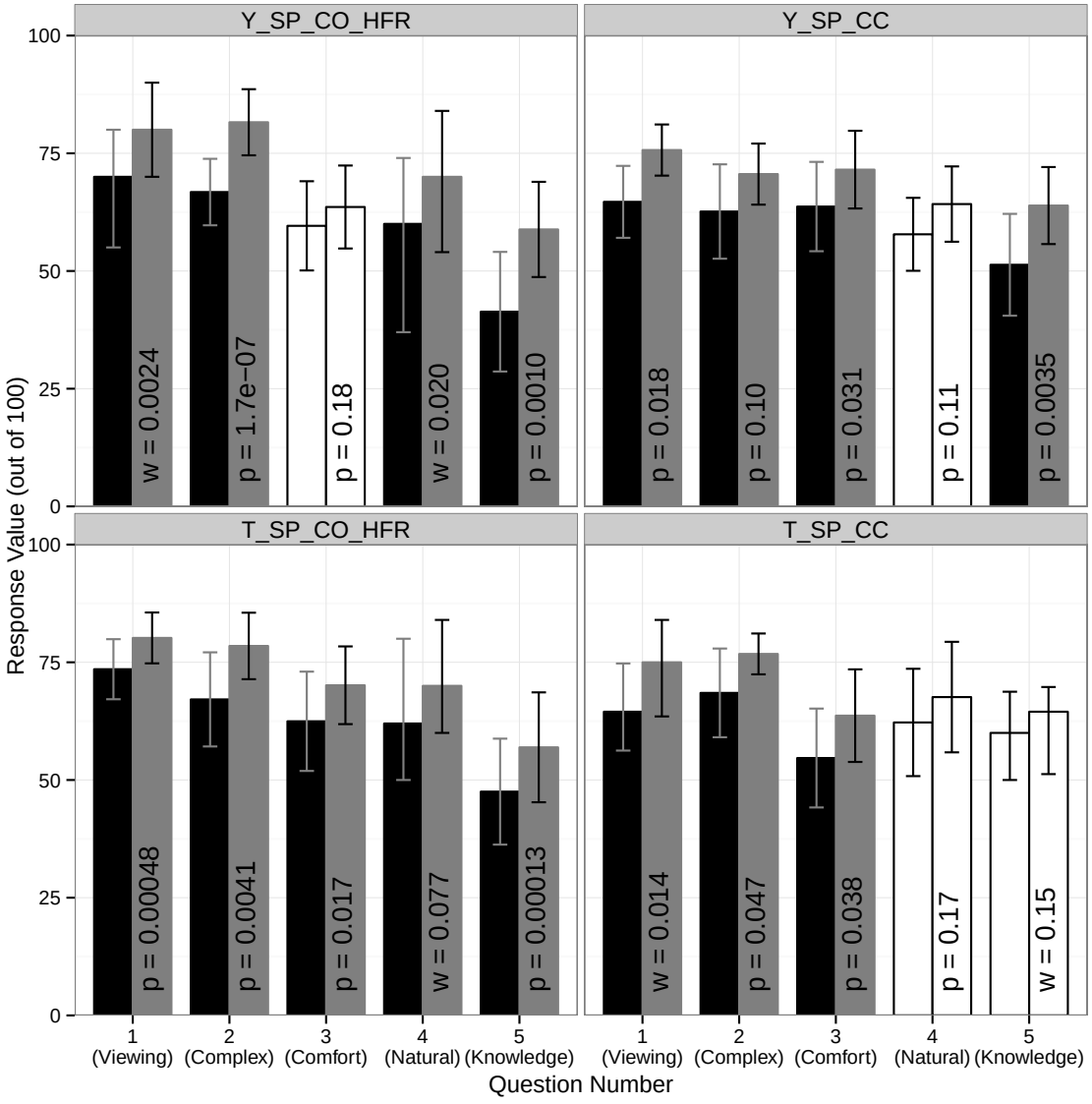

Figure 5.5: The results of experiment 4 using the small screen projection at sites in York (UK) and Twente (The Netherlands). Depending upon the result of the Shapiro-Wilk test for normality, the black bars indicate the mean or median preliminary response, whilst the grey bars indicate the mean or median post-viewing response and the errors bars denote the standard deviation or inter-quartile range across the sample. The result of a paired Student's t-test, or Wilcoxon signed rank test is also shown for each question. In the cases where the Shapiro-Wilk test failed and ranked statistics are used, the statistical test result is labelled with a w instead of a p. The white bars indicate questions where the statistical test failed to find a significant difference between preliminary and post viewing responses (the result did not break our alpha significance criterion of 0.1).

5.5.1 Experimental setup

This differentiated replication used the equipment outlined in Section 5.4.1. The equipment was taken to rooms in York University and Twente University and set up in the same way. Using the technique outlined in Section 5.2.1, the black screen and white screen luminance at both sites were measured to be less than 0.93 lux and 9.3 lux respectively. The volume level at the viewer's listening position was matched to the previous experimentation using a decibel meter.

At York University participants were recruited from the undergraduate and postgraduate courses run in the Department of Theatre, Film and Television and the Department of Computer Science. Some members of staff also took part. Eighteen participants undertook *Cosmic Origins* HFR viewings, of which 1 failed the screening test, and 24 participants took part in the *Cosmic Cookery* viewings of which 5 failed the screening test. Ages were distributed between 18-57, with an interquartile range of 19-27 and a median of 21. The gender balance was 71% male to 29% female.

The experimentation at Twente was run during the summer holidays, so participants could not be recruited from the undergraduate body. Instead they were sourced primarily using postgraduate and staff mailing lists. Twenty-one participants took part in the *Cosmic Origins* HFR viewings, of which 4 failed the screening test, and 22 participants took part in the *Cosmic Cookery* viewings, of which 5 failed the screening test. Ages were distributed between 22-38, with an interquartile range of 24-28 and a median of 26. The gender balance was 80% male to 20% female.

5.5.2 Results

The results for the experimentation undertaken in York are shown in the top graphs of Figure 5.5. Only the post-viewing responses to Question 1 (viewing experience) and the preliminary responses to Question 4 (naturalness) failed the Shapiro-Wilk test for normality with p-values 0.0308 and 0.0266 respectively. The response differences failed to prove statistically significant for Question 3 (comfort) in the *Cosmic Origins* data, and Questions 2 (complex information) and 4 (naturalness) in the *Cosmic Cookery* data. Equivalence tests show that these mean response differences are not equal to zero, so we conclude that they are null results (like those discussed in Section 5.3.2).

The Twente results are shown in the lower two graphs of Figure 5.5. The post-viewing responses to Question 4 (naturalness) was the only data set in the *Cosmic Origins* data to fail the Shapiro-Wilk test for normality with a p-value of 0.0296. The post-viewing responses to Question 1 (viewing experience) and the preliminary

ID	Question 1		Question 2		Question 3		Question 4		Question 5	
	Test	p-Value	Test	p-Value	Test	p-Value	Test	p-Value	Test	p-Value
D-BP-CO	t	0.0011	t	0.012	t	0.069	t	0.0035	-	-
D-BP-CC	w	0.013	t	0.0010	t	0.021	t	0.0040	-	-
D-TV-CO-HFR	w	0.053	t	2.9E-4	**t**	**0.12**	t	0.028	t	7.1E-4
D-TV-CO-HR	t	0.025	t	0.025	**t**	**0.36**	t	0.022	w	0.0013
D-TV-CC	**t**	**0.11**	t	2.8E-4	t	0.030	w	0.0086	t	0.0026
D-SP-CO-HFR	t	0.0012	t	4.2E-6	t	0.0060	t	1.1E-4	t	6.3E-6
D-SP-CC	w	0.027	w	6.8E-4	t	0.073	w	0.0089	t	0.0052
Y-SP-CO-HFR	w	0.0024	t	1.7E-7	**t**	**0.18**	w	0.020	t	0.0010
Y-SP-CC	t	0.018	t	0.10	t	0.031	**t**	**0.11**	t	0.0035
T-SP-CO-HFR	t	0.00048	t	0.0041	t	0.017	w	0.077	t	0.00013
T-SP-CC	w	0.014	t	0.047	t	0.039	**t**	**0.17**	**w**	**0.15**

Table 5.2: The p-values from all the significance tests used to determine whether we can reject the null hypothesis that there is no change between preliminary and post-viewing responses. The ID symbol is broken into three parts. The first letter indicates the site: D for Durham, Y for York and T for Twente. The second two letters indicate the display: BP for Big Projector, TV for Television and SP for Small Projector. The final set of letters indicate the film: CO for *Cosmic Origins* and CC for *Cosmic Cookery*. As multiple versions of *Cosmic Origins* have been used a further identifier code is used: HR corresponds to the Higher Resolution version and HFR corresponds to the Higher Frame Rate version.

responses to Question 5 (knowledge) in the *Cosmic Cookery* data failed the Shapiro-Wilk test for normality with p-values of 0.0152 and 0.0399 respectively. Questions 4 (naturalness) and 5 (knowledge) from the *Cosmic Cookery* data failed to pass the significance tests.

5.6 Discussion

We begin by reviewing the individual cases where our significance testing was successful (Section 5.6.1), before turning to speculate on those cases where it was not (Section 5.6.2). We then use ANOVA to identify differences within the data (Section 5.6.3), which is followed by an analysis of the combined data taken from all our experimentation (Section 5.6.4). This section concludes by discussing threats to the validity of our results (Section 5.6.5).

5.6.1 Significance test successes

The p-values from all significance tests are shown in Table 5.2. They show that the results are overwhelmingly positive, with the majority (79%) of significance tests yielding a "strongly significant" result. questions 1 (viewing experience), 2 (complex information) and 5 (knowledge) performed particularly well, with only

one significance failure for each.

Question 2 was the strongest performing question in this study, with a mean response difference of 15.17 and all cases proving at least weakly significant. Furthermore, there was only one experiment in which the significance test for Question 2 did not prove strongly significant. These strong results are supported by the comments of 23 participants that suggest S3D is particularly suitable for conveying complex spatial information. For instance, *"Watching 3D films may improve and enhance understanding, particularly on complex topics which need 3D graphics to emphasise a point."* It seems that the binocular cue can greatly improve the processing of complex visual information.

The results from Question 5 (knowledge) also performed well, with only one significance failure and an average response difference of 15.09. A small number of comments contrast with these strong numeric results by arguing that the visuals distracted them from the film's narration. One such comment said, *"Sometimes the 3D effects can distract from the narration as I found I was too focused on the visuals."* It would be interesting to undertake further study assessing the impact of 3D visuals upon processing audio-visual information.

Twelve participants stated in their comments that the purpose of S3D in films needs further consideration. One such individual said S3D effects *"have tended to be seen as a gimmick rather than a form of visual expression. If we can move away from the sensationalist "theme ride" nature of current 3D viewing [it] could be very effective."* This suggests that, for many, the S3D effect comes at a cost, which they feel should clearly be re-paid through added value in the content. Such added value may be found in complex visual information, of which the content in *Cosmic Origins* and *Cosmic Cookery* is an example.

5.6.2 Significance test failures

The seven results that failed to prove even weakly significant are shown in bold in Table 5.2. In this section we speculate on why these cases failed to show significance.

Three of the null results occurred when viewing the films on the TV display. When analysing the comments we found that 19% of participants who took part in the TV viewings actively complained about crosstalk (see Section 2.3.2). Whereas only one comment from the rest of the experimentation could potentially be connected to crosstalk: *"Images are still split into two when they come further away from the screen"*. Crosstalk is a negative factor associated with the S3D displays that may possibly explain these three failed significance tests (Pala et al. 2007).

Question 3 (comfort) yielded the weakest set of results (3 out of 11 cases failed to prove even weakly significant). It seems that discomfort can still be a problem even when when viewing S3D films with quality-controlled depth. Analysing the comments can perhaps offer some further insight into this matter. Whilst 23 participants did complain about discomfort/ache/tiredness specifically in the eyes, almost the same number (22) complained about discomfort due to wearing glasses — a factor that cannot be influenced by high quality content. There were a number of comments concerning comfort that were very favourable, such as, *"The film seen today was noticeably more comfortable to watch than normal 3D films."* A few people acknowledged improved comfort whilst questioning whether this would hold for longer time periods, such as *"Obviously, I have just watched a brilliant 3D film and feel comfortable. I just wonder whether the technique of the short film can be successfully applied to other long films."* The short length of each film is a limitation of this study since visual comfort can degrade over viewing time (Lambooij et al. 2009; Nojiri et al. 2004).

All significance failures, except those in Question 3 (comfort), occur in *Cosmic Cookery* viewings. It is hard to see why *Cosmic Cookery* performs so erratically, with failures in every question except number 3 (comfort). It seems most likely these failed significance tests are the result of the small sample size limiting the statistical power. When designing this experiment we sought to achieve the commonly accepted value for statistical power of 80%. For a sample size of 15, with standard deviation and effect size set at 10 scale units, the statistical power is actually found to be 85%. However, this still suggests that we should fail to reject correctly the null hypothesis in 15% of the Student's t-tests. In actual fact our t-tests have failed in 6 of 43 cases, which is equivalent to 14% of the tests. If we were to repeat the experiment, we would consider using samples of approximately double the size, to attain 98.5% power. Whilst the statistical power may explain our failed t-tests, it does not threaten the validity of conclusions drawn from successful tests.

5.6.3 Looking for differences with ANOVA

Although there is some perplexing variation in the results of the individual significance tests as noted above, ANOVA performed across all 11 studies for questions 1-4 yielded no significant differences between studies. Table 5.3 shows the F-values and the probabilities associated with these ANOVA. The only question with any significant difference between studies is Question 5 (knowledge). The participants for this study have been recruited from selected academic communities. One could expect

	ANOVA		Combined Data (n=186)		
Question	**F-Value**	**Pr(F)**	**Mean**	**Std. Dev.**	**p-value**
1 Viewing Experience	0.53	0.87	8.715	12.96	9.0e-17
2 Complex Information	0.85	0.58	12.82	14.75	1.8e-24
3 Comfort	0.36	0.96	7.672	14.99	5.1e-11
4 Naturalness	0.86	0.57	10.79	16.34	2.6e-16
5 Knowledge	3.6	8.2E-4	-	-	-

Table 5.3: The results of ANOVA seeking any differences between the original experiment and differentiated replications for each question. Where the ANOVA failed to find any differences, details of t-tests using the combined data across all experimentation are given. These t-tests again use the null hypothesis that the mean response difference is zero. The alpha significance criterion was 0.05, so only Question 5 (knowledge) yielded a significant ANOVA result, whilst all of the combined data t-tests proved significant.

differences to occur in the learning of content information, and thus response differences to Question 5, based upon the academic discipline (i.e. Maths students may be more interested in, and better prepared to learn about, galaxy formation than Anthropology students). As the recruiting of participants often involved targeting specific groups of academics, each sample of participants did not represent a random selection across academic disciplines. This could explain the variance observed in Question 5.

The failed ANOVA tells us that there is not enough evidence to conclude that the contributing samples are taken from different distributions. Therefore, analysis of the combined data (from all rounds of the experimentation) may be of interest. For each question that failed the ANOVA, Table 5.3 also includes the details of Student's t-tests that have been performed using combined data. Every test passes, including the erratic Question 3 (comfort). We can also conclude from these ANOVA that the results are repeatable for different films, sites and display technologies.

5.6.4 Analysing combined data

Figure 5.6 shows the results of combining data from all rounds of experimentation. In total, 186 participants contributed to this combined data set. Student t-tests were run on each film's combined data to establish whether there were significant differences between preliminary and post-viewing responses. All tests yielded strongly significant results.

For each of the first four questions the combined data was split by gender and the means and standard deviations of each gender's responses to each question

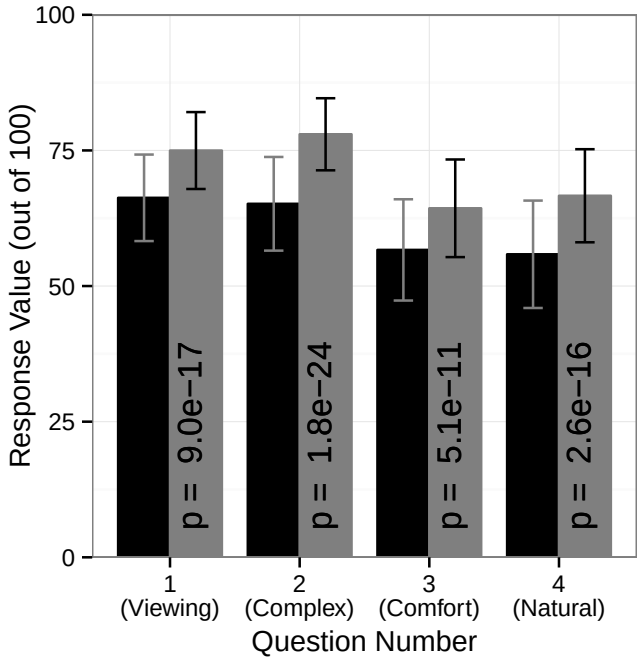

Figure 5.6: Showing the results of combining all our data from 186 participants who passed the screening test. The black bars indicate the mean preliminary response, whilst the grey bars indicate the mean post-viewing response and the error bars denote the standard deviation in responses. The p-value (labelled p) of a paired Student's t-test is also shown for each question.

calculated. Independent two sample t-tests for samples with unequal sizes and variance were then used to determine if the mean responses differed significantly with gender. No significance was found, suggesting that gender is not an influencing factor upon the observed change in attitude towards S3D films.

5.6.5 Threats to the validity of our results

The steps we have taken to minimise threats to the *construct validity* of our results have already been discussed in Section 5.1.1. By using short films, simple questionnaires and controlling certain aspects of the environment, we have removed a number of factors that literature suggests may threaten the existence of a causal relationship between our intervention (the S3D film viewing) and the differences in the test results (the questionnaire response differences).

Unfortunately the presence of significant threats to the *internal validity* of our results cannot be ruled out, because we were unable to find a suitable intervention for a control study. There is no accepted definition of a "normal" S3D film for us to test our "high-quality" S3D films against. A pre-test post-test quasi-experimental

design is often used when no control is available, as the preliminary responses act in a similar manner to a control study for the post-intervention results to be compared against. The preliminary responses rule out any bias caused by prior experience of S3D film quality. Consequently, if we can trust that participants answered our questions honestly and appropriately, and were not led to do otherwise by some aspect of the experiment's execution other than the intervention, then we can trust the validity of our results.

This study is made up of differentiated replications of the same experiment, using different participants, films, displays and sites to gain a wider understanding of the scope of our results. Despite this, it is important for us to acknowledge that there are bounds to the scope, which pose threats to the *external validity* of our results. We can conclude very little concerning the bounds of the scope, so researchers should be careful about assuming that our results hold in scenarios with notably different characteristics. For instance, our participant samples were not truly random, as they were sourced from academic communities of students and researchers, so would typically be dominated by a particular academic discipline and a particular age group. Therefore, our results may not hold for audiences with a significantly different demographic, such as those made up of children or the elderly.

5.7 Conclusions

In this study we have shown S3D films with quality-controlled binocular depth to groups of participants. Before and after watching the film we asked the participants to fill out a questionnaire. Both questionnaires asked the same questions concerning their attitude towards S3D films. Responses were given on a 0-100 point scale, where a greater number indicated a more positive response. This paper reports an original experiment and four differentiated replications, across which we varied the display, film, and site used. The original experiment investigated reactions to a large screen projected display in our Durham based laboratory. This was followed by replications using a TV display and a small (TV-sized) projected display. The small projected display was then taken off-site to the University of York (UK) and the University of Twente (The Netherlands). The films that we used were created by a collaboration of physicists and computer scientists at Durham University and were entitled *Cosmic Origins* and *Cosmic Cookery*. Between 15 and 19 participants who had been successfully screened for stereo vision took part in each viewing. The difference between their preliminary and post-viewing questionnaires were tested against the null hypothesis that they would be equal to zero. Paired Student t-tests

or Wilcoxon signed rank tests were used as appropriate to determine the confidence with which we could reject this null hypothesis, and say that a response change had occurred across the audience. ANOVA were used to look for differences in mean values between the original experiment and replications. The statistical results were discussed alongside comments left by participants at the end of the post-viewing questionnaire.

In answer to this chapter's first research question, we have seen that high quality S3D films using quality-controlled binocular cues can create a measurable positive change in an audience's attitude towards S3D films. This change was observed in response to all of the following questions:

1. Please rate your impression of the viewing experience 3D films can provide.
2. Please rate your impression of how well 3D films can convey complex information.
3. Please rate your impression of how comfortable you think viewing 3D films can be.
4. Please rate your impression of how natural the sensation produced by viewing 3D films can be.
5. Please rate your knowledge of how galaxies are made.

Use of ANOVA failed to find any differences between the experiment and replications in response changes to each of the first four questions. It is possible, then, that each data sample comes from the same distribution. Paired Student's t-tests between preliminary and post-viewing responses across the combined data gave strongly significant results for the first four questions. This therefore indicates that the positive changes in attitude towards S3D films that have been observed in Questions 1-4 are repeatable at national and international sites, as well as for different display technologies and quality-controlled film content, which answers this chapter's second and third research questions. Significant differences in response changes were found between the experiment and replications for Question 5 (knowledge). We have speculated on whether this is due to participants being recruited through specific academic disciplines.

This study motivates research concerning high quality S3D content creation by showing that such content elicits measurable, repeatable changes in audience attitude towards S3D. Furthermore, these attitude changes remain significant for different displays, sites and high quality content. Our research therefore concludes that the current popular attitude towards S3D may be significantly improved by

the wider distribution of high quality content, created with algorithms such as those outlined by Jones et al. (2001) and Holliman (2004).

CHAPTER 6

Minimum audible depth for 3D audio-visual displays

We continued to consider whether sound can influence viewers' perception of quality-controlled S3D visual depth. In order to do this, we needed to build and calibrate a display system capable of conveying both auditory and visual depth cues to the viewer. There were several steps in this process, most of which were concerned with the audio component of the display system. These steps included: the selection of suitable loudspeakers; the sourcing of equipment to position the loudspeakers; the design of software to control the presentation of audio and visual stimuli; the measurement of background noise levels and luminance; the calibration of stimuli volume and luminance; and, crucially, the measurement of the MAD associated with the audio component of the system. This chapter therefore directly addresses the third subsidiary research question presented in Section 1.2: *What is the MAD in our experimental setup?*.

It is clear from the previous studies outlined in Section 2.2.4, and the work undertaken by Durham University students (Turner 2010; Berry 2011; Wills 2012), that the MAD is sensitive to the sound system, environment, and participants used. It is therefore important to measure the MAD for the experimental setup that will be used in future experimentation. This chapter primarily focuses on the work undertaken to do this, though it also reports various other aspects of the calibration process. Our measurement of the MAD, which is environmentally valid for TV viewing scenarios, forms a novel contribution of this thesis. It adds a data-set to a small group of pre-existing studies that measure the MAD, but also uses a unique setup which was designed to give the result environmental validity for TV viewing scenarios. This unique setup includes the use of a semi-reverberant environment, the positioning of a TV screen to reflect sound, and the selection of listening distances

from recommended TV viewing distances. The data and experience acquired in this study were also used to propose a participant screening test for RAD perception that can be used in future experimentation. This is another a novel contribution, and one that is important because of the variability in RAD performance that we observed between participants.

The MAD is a sensory threshold, meaning the research question addressed in this chapter is a threshold measurement problem. A perfect threshold for a given sensory task would allow us to plot a step function on the graph of task performance against the dependent variable. Chance performance would be followed by a step-up to perfect performance at the threshold. In practice we see a logistic function as performance improves from chance to perfect. The point on this curve which is chosen as the threshold is largely arbitrary. The studies by Turner (2010), Berry (2011) and Wills (2012) use the point at which performance is significantly different from chance, whilst one could also argue for the point at which performance becomes significantly different from perfect. The popular approach is to aim for a point half way between chance and perfect performance (Palmer 1999).

We begin by discussing the experimental method for this study in Section 6.1. The results are outlined in Section 6.2 and followed by a discussion in Section 6.3. We draw relevant conclusions and discuss their implications for the rest of this thesis in Section 6.4.

6.1 Method

Here we outline the steps taken to design the calibrated display and the experimental method used to measure the MAD. We begin in Section 6.1.1 by outlining the preliminary work that contributed to the final experimental design, described in Section 6.1.2. We then give details of the equipment and environment used in Section 6.1.3, before discussing the design of a post-experiment questionnaire in Section 6.1.4. We finish in Section 6.1.5 by giving details of the participant samples used.

6.1.1 Preliminary trials

Substantial preliminary experimental work was undertaken before settling upon a final method. Although the results from this preliminary work proved unsatisfactory, they offered an important contribution towards the design of the final experiments. Here, details of this work are reported briefly for completeness.

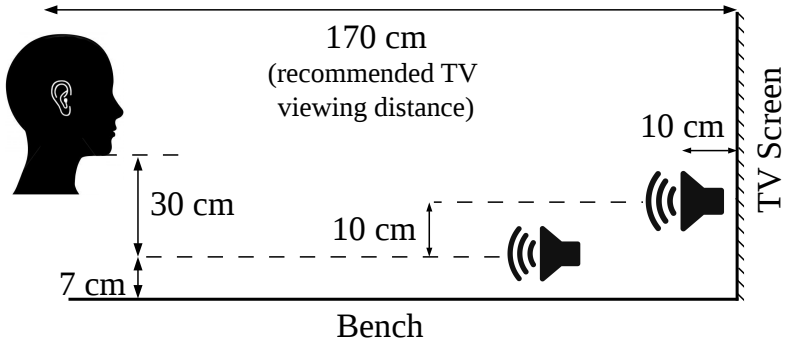

Figure 6.1: The layout of equipment in the preliminary trial, with respect to the participant in the experiment. Diagram is not to scale.

As already mentioned, this study continues previous work undertaken by undergraduates at Durham University (Turner 2010; Berry 2011; Wills 2012). In these studies, the threshold was defined as the depth difference at which the mean sample performance is significantly different from chance. The trouble with this definition is that sample performance can be significantly better than chance whilst the majority of individuals are still unable to perform better than chance. The widely accepted definition of a sensory threshold is the halfway point between chance and perfect performance, assuming a logistic model for the data (Palmer 1999). This requires a change in the experimental method and data analysis.

A 2AFC paradigm was used to assess task performance. Participants were played pairs of telephone rings and asked to respond with the ring they judged to be nearest to them. The percentage of correct responses across a sample gives a measure of task performance for the given auditory depth difference between the rings. A plot of task performance against increasing depth difference should follow a logistic curve between chance (50% for 2AFC) and perfect (100%) performance (Palmer 1999). For a data set covering a range of depth differences, logistic regression can then be used to interpolate the data. The MAD is then taken as the depth difference from this model, corresponding to a task performance of 75%.

One of the two loudspeakers was mounted on a motorised platform that could be positioned by a computer at any point on a 91 cm rail. The other loudspeaker was statically mounted so that the mobile loudspeaker could just slide underneath it. In the null case, when the depth difference was zero, the loudspeakers were therefore positioned with one directly above the other. This arrangement was chosen because the minimum audible vertical angle is larger than the minimum audible horizontal angle (Perrott and Saberi 1990). A TV screen was placed behind the static speaker, just as in the cross-modal experiments reported later in this thesis. A chin rest was

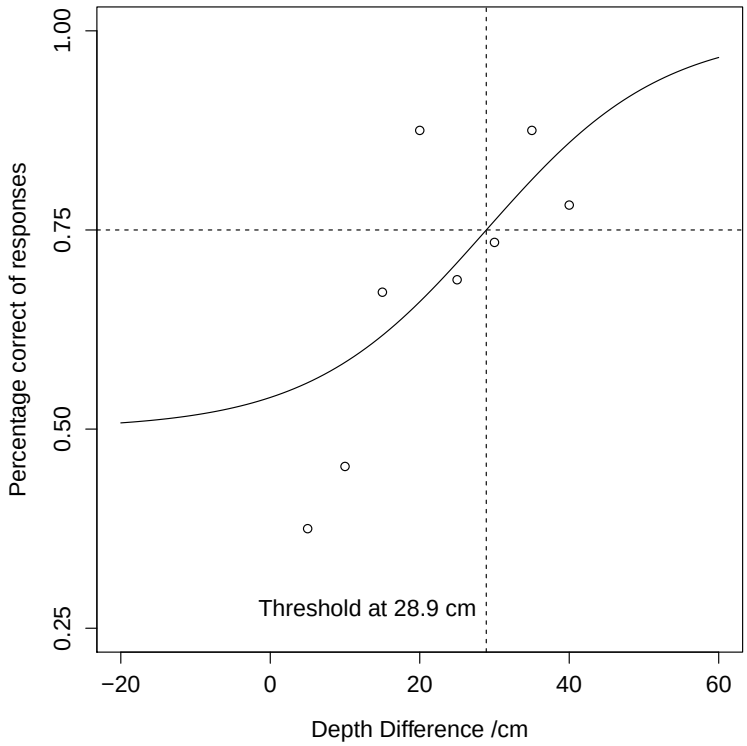

Figure 6.2: The results of a preliminary experiment designed upon the principle of logistic regression. The logistic curve fitted to the data is shown, with the depth difference corresponding to 75% correct also marked. The threshold calculated using this method was 28.9 cm, which is 18% of the distance between the listener and the far speaker.

used to position the head at the recommended viewing distance supplied by the TV manufacturer. The experimental set up is shown in Figure 6.1.

The depth differences ranged from 0-40 cm inclusive in 5 cm intervals. Sixteen participants completed four tests at each depth difference, making 36 tests per participant. Participants were selected from the student population at Durham University. The experiment was abandoned after 16 participants had contributed results, because the results that we had collected were not what we expected. The percentage of correct responses for each non-zero depth difference is plotted in Figure 6.2, together with the logistic curve of best fit and the 75% threshold.

A logistic function, forced to pass between chance and perfect performance, does not appear to be a good fit for these results. For small distances, the scores were notably less than chance, and in the null case responses were split 77% to 23% between the two loudspeakers (where we would expect a 50% to 50% split, indicating chance performance). Further investigation found that participants could consistently distinguish between the two loudspeakers in the null case; over 20 tests they

could consistently identify the same loudspeaker's noise from the randomised pair of sounds. Clearly there were confounding auditory cues caused by the experimental setup, as two sounds played from the same depth should be indistinguishable.

We first considered whether the loudspeakers were sufficiently well matched. Upon analysis of the original loudspeakers' frequency response curves for the telephone ring we found differences in excess of 10 dB in some frequency bands. After some searching for small but well matched loudspeakers, we settled upon a pair of K-array KT20s. These loudspeakers were 6.4 cm in diameter and 8.3 cm deep. They were matched by K-Array so that their frequency curves remained within 1.6 dB of each other.

Whilst using the matched pair provided a substantial improvement upon previous equipment, they did not solve the problem of audible differences between loudspeakers in the null case. By positioning a microphone at the listening position, we discovered that the pairwise matching was being broken by some aspect of the experimental set-up. In the null case, the only significant non-symetrical aspect of the setup was the loudspeaker positioning: one above the other. The difference in distance between the loudspeaker and bench surface could cause reverberation and interference patterns capable of breaking the matching.

Placing the loudspeakers side-by-side, instead of above-below, did remove the audible frequency differences between loudspeakers in the null case, though it introduced a new problem: the inter-aural differences arising because of the azimuthal offset were just audible. This new left-right cue therefore had to be randomised to stop it leading participants' responses. Another motorised platform was used to randomise which speaker was assigned to the near or far position.

As well as re-designing our experimental setup, we decided to revise our chosen experimental method and statistical design. The logistic regression method outlined here offers little insight into each individual's performance and is not widely used by others for measuring sensory thresholds. The final design used a method that yielded threshold estimates for each individual.

6.1.2 Final design

Blindfolded participants were asked to undertake a series of 2AFC tests. In each test they were presented with an auditory stimulus played sequentially from each of two static loudspeakers, placed about the median plane at different depths. The depth difference varied between tests. For each test, participants were asked, "Which sound appears nearest to you?" They were required to chose from two possible answers:

"The first," or, "The second." The correct answer to this question was randomised for each test and each participant.

A two-down/one-up transformed adaptive procedure (Levitt 1971), with PEST for the step size adaption (Taylor and Creelman 1967), determined the depth differences in each test and the calculation of the MAD for each participant. All participants began with a depth difference of 40 cm. The depth difference would decrease following two correct answers, and increase following a single incorrect answer. Each depth difference change is called a "step" and multiple steps in the same direction (either increasing or decreasing) are called a "run". When a step in one direction is followed by a step in the opposite direction, a "reversal" is said to have taken place. The size of each step is specified by the rules of PEST (Taylor and Creelman 1967):

- On every reversal of step direction, halve the step size.
- The second step in a given direction, if called for, should be the same size as the first.
- The fourth and subsequent steps in a given direction are each double the previous step.
- The third successive step in a given direction is double the second if the step immediately proceeding the most recent reversal was a result of a doubling. Otherwise, the third step is the same as the second step.

Taylor and Creelman (1967) developed these rules using a mix of intuition and computer simulation. The first two rules of PEST create something similar to a binary search (Cormen et al. 2009), since each reversal indicates that the target value may have been passed. The inconsistent nature of human perception means a reversal does not always imply that the target value has been passed, causing the search to occur in the wrong area. This is likely to be the case when multiple steps are taken in the same direction. The third rule therefore dictates that when multiple steps occur in the same direction, the step size should be increased to efficiently find the right search area. The final rule improves efficiency by breaking continuously repeating patterns that will occur if the third step is either always or never doubled.

The initial step size was set to 20 cm, whilst the final step size was set to 0.625 cm. The experimental procedure therefore ended after the step size had been halved five times. The depth difference between the loudspeakers at this point was taken as the participant's MAD. The mean MAD across the sample of participants can then be compared with other samples and with the PDH using appropriate statistical tests.

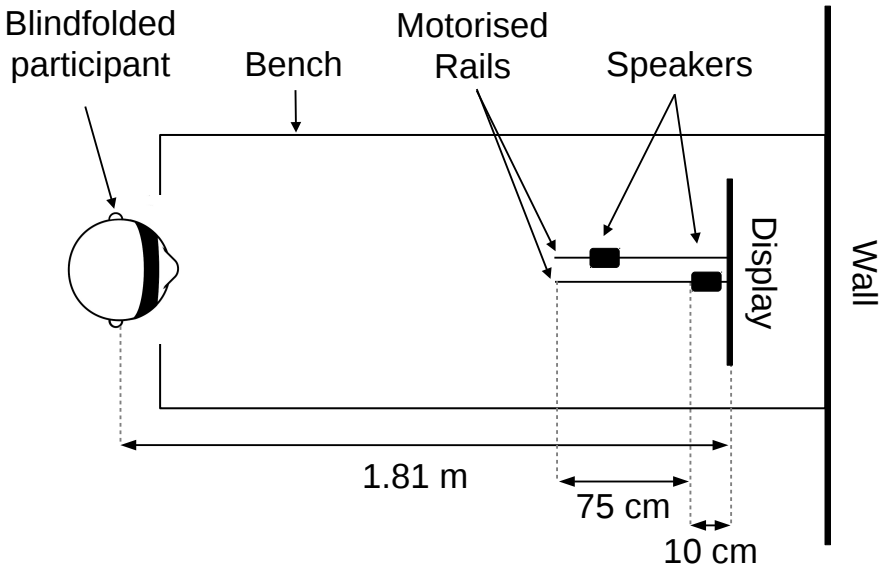

Figure 6.3: The experimental setup, as seen from above. Motorised rails, controlled by a computer, were used to change the depths of two K-Array KT-20 loudspeakers in front of a 47 inch display. The loudspeakers were positioned 8 cm apart and 14 cm above the desk, whilst the chin rest stood 37 cm above the desk. Participants were blindfolded to avoid vision influencing their responses.

Prior to the recorded tests, participants were given nine training tests. This sequence of tests began with a 40 cm depth difference, which decreased to 0 cm and then back to 40 cm in 10 cm intervals. They were told whether their response was correct for each test, though their response was not recorded. The aim of this training period was to reduce the size of any learning effect that might occurduring the recorded experiment.

6.1.3 Experimental setup

The experimental setup is shown in Figure 6.3. We have chosen to physically position loudspeakers at the desired depths, rather than employ a virtual 3D sound system to create auditory depth. This removes any dependency of our results upon the validity of a virtual sound system's design. As mentioned in Section 6.1.1, two K-Array KT20 loudspeakers were chosen because of their small size (to minimise occlusion and interference) and their availability in frequency-response matched pairs. These loudspeakers were placed side-by-side, each mounted on a motorised platform that could slide along a rail as controlled by a computer. The side-by-side arrangement did introduce small audible inter-aural differences, so the near speaker was randomly selected in each test to remove the left-right cue.

The motorised rails protruded out underneath an LG BM-LDS302 47 inch 3DTV screen, so that the motors were behind the screen whilst the platforms were in front of the screen. This arrangement allowed the loudspeakers to be positioned as close as possible to the screen. The TV display is included in the setup to give environmental validity to our results. The blindfolded participant was positioned, using a chin rest, such that the two loudspeakers were symmetrically distributed around the median plane.

As the speaker is 8.3 cm deep with a cable plugged into its rear, its front face could not be positioned nearer than 10 cm in front of the screen. The distance between the listener and the far loudspeaker's front face is called the reference distance, and is taken as 10 cm less than the viewing distance. Three different reference distances were tested, based upon distances at which the TV screen fills the $40\,^\circ$, $30\,^\circ$ and $20\,^\circ$ viewing angles. The $40\,^\circ$ viewing angle corresponds to the smallest viewing distance and is recommended by THX (2013), whilst the $30\,^\circ$ viewing angle is widely quoted as the Society of Motion Picture and Television Engineers (SMPTE) recommendation (Rushing 2004). This corresponds to reference distances of 1.33 m, 1.81 m and 2.88 m.

The experiment was performed in a semi-reverberant laboratory with a background noise of approximately 41.5 ± 0.3 dB (the mean and standard deviation of 18 measurements separated by 10 second intervals). The loudspeakers were driven by a Cambridge Audio Topaz AM1 Amplifier. The volume was set to be a maximum of 70.0 dB at the approximate listening position with the loudspeakers positioned at the reference distance. All equipment was placed upon a desk that stretched across the gap between the participant and the TV screen. The loudspeakers were positioned so that the centre of their front faces were 8 cm apart and 14 cm above the desk, whilst the chin rest stood 37 cm above the desk. A photograph of the experimental setup is shown in Figure 6.4.

The sounds and positions of the loudspeakers were controlled by a computer program that automated the experimental method in a traceable manner, leaving the experimenter to input the participant's responses (which stimulus appeared nearest - the "first" or the "second") and to choose when to run each test. At the end of the experiment, the program returned a measurement of the participant's MAD and a text file recording the participant's responses to each test with the corresponding test details.

The depth differences that the system could present were limited to between 0 cm and 60 cm due to the length of the motorised platform rail. This posed a problem when the experimental procedure dictated that these limits be exceeded.

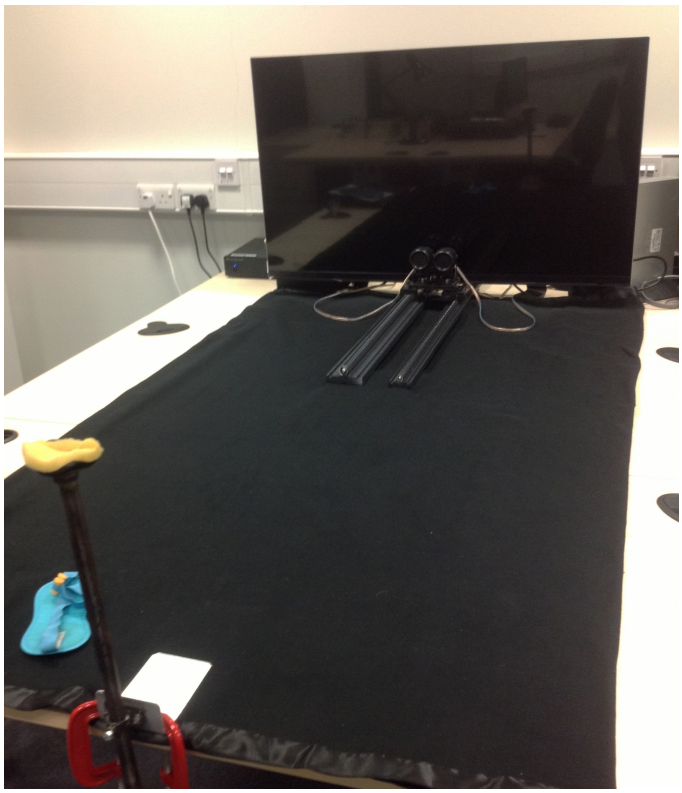

Figure 6.4: A photo of the experimental setup and laboratory environment used.

Our solution was to replace the desired depth with the limit and allow the rules of PEST to continue as normal. If the participant's MAD did not fall within these limits, the above solution would result in the procedure never converging upon a final result. Instead, the depth difference would become "stuck" on the limit, only changing in size according to chance performance. As the procedure converged upon a final value for all participants, this does not threaten the validity of the experiment's results.

We chose to use the telephone ring from the preliminary experiment as the auditory stimulus for this study, as a mobile telephone seemed to be an excellent cross-modal stimulus for the reasons outlined in Section 4.1.2. This decision also reflects our aim as display systems engineers to obtain environmental validity by avoiding abstract laboratory conditions. The stimulus lasted for 3.01 s, which consisted of 1.57 s of rapidly repeated metalic rings followed by 1.44 s of the final ring dying away to near 0 dB amplitude. In Figure 4.2 (page 60) we plot the frequency spectrum of the stimulus, which shows its complex nature.

6.1.4 Qualitative data capture

Each participant was required to fill out a post-experiment questionnaire concerning their involvement in the experiment. The purpose of this questionnaire was to seek qualitative evidence related to the quantitative data and to identify any threats to the validity of the participant's results. Using a questionnaire ensures that each participant's data is captured in a consistent and repeatable manner. The questionnaire recorded responses to the following questions:

1. What is your age?
2. What is your sex?
3. Did you understand the task required of you? (Yes/No)
4. Do you feel that your answers were a correct representation of what you heard? (Yes/No) If not, why?
5. Please comment briefly on how you determined which ring was nearer.
6. To your knowledge, is there any reason why you may have performed particularly well or particularly badly at the experimental task? Include any reasons you may have for thinking your hearing is different from "normal" hearing.
7. Do you have any other significant comments that may be worth recording regarding the execution of the test?

Responses to Questions 1, 2, 3 and 4 were given by ticking a box labelled "No" or "Yes". The participant was given a box in which to enter their answer for Questions 4, 5, 6 and 7 as prose. In Question 4 they were only asked to enter prose if their answer to the first part of the question was "No". Participants were given as much time as they required to fill out the form to their desired level of detail.

6.1.5 Participants

The participants were sourced from the postgraduate and undergraduate student groups at Durham University and did not include the author. They were recruited through various departmental and college mailing lists. Twenty participants took part for each reference distance, making 60 MAD measurements in total. This sample size was based upon a power analysis, which used results from preliminary trials to test equality and difference between sample means and the PDH. Participants were allowed to contribute even if they had prior knowledge or experience of the experiment, though they were only allowed to contribute one MAD measurement for each reference distance.

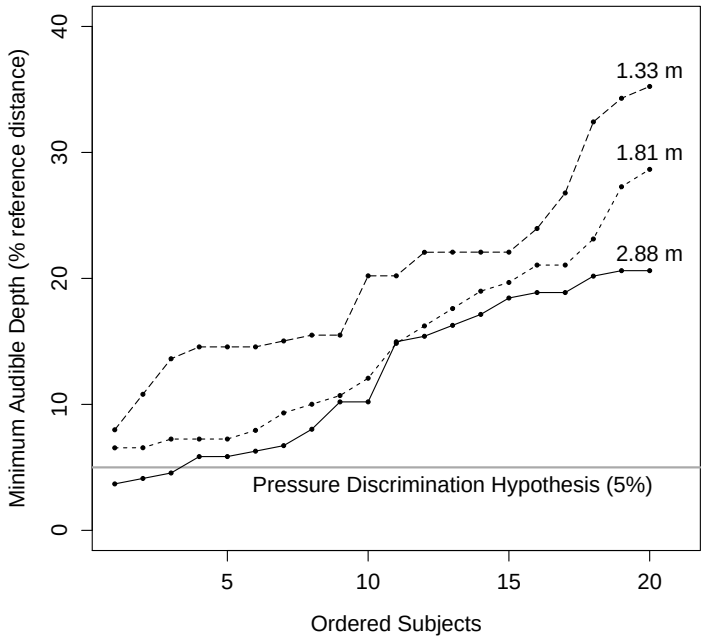

Figure 6.5: The ordered distribution of participants' thresholds. Each experiment's data set is labelled with the reference distance (distance between listener and far speaker). The PDH is marked as a grey line for comparison.

For the 2.88 m reference distance, 68% of the participants were male and 32% female. Their ages ranged from 21-32, with an inter-quartile range of 23.5-24.5 and a median of 24. For the 1.81 m reference distance, 65% of the participants were male and 35% female. Their ages ranged from 20-37, with an inter-quartile range of 23-25.25 and a median of 24. For the 1.33 m reference distance, 45% of the participants were male and 55% female. Their ages ranged from 21-32, with an inter-quartile range of 23.75-26 and a median of 24.5.

All participants were required to pass the BSHAA online hearing test prior to their participation in the experiment. This test requires the participant to demonstrate they can hear four tones of 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz. Doing this allowed us to ensure that all participants met the required standard of hearing.

6.2 Results

For the 1.33 m, 1.81 m and 2.88 m reference distances, we measured MAD samples with respective medians of 20.20%, 13.46% and 12.58% of the reference distance. Figure 6.5 shows how the individual results are distributed for each reference dis-

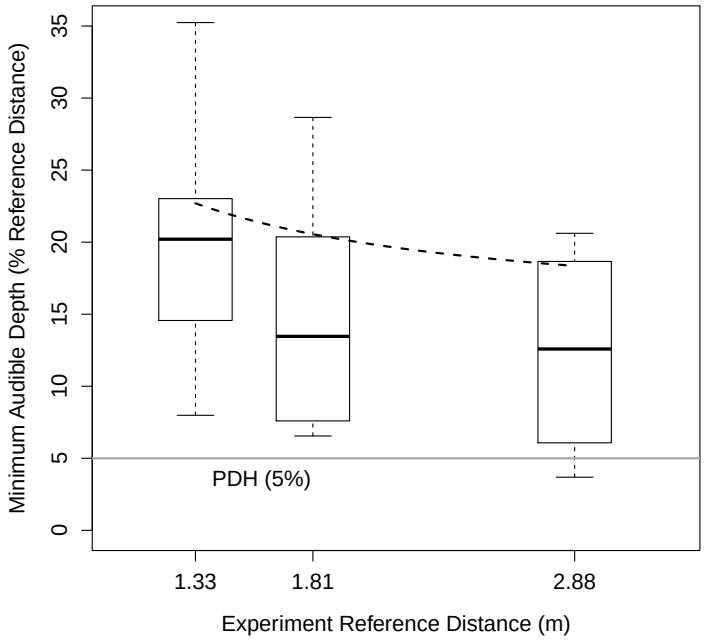

Figure 6.6: Our results presented as box plots. The whiskers denote the total range of the sample, whilst the box shows the inter-quartile range and the central line marks the sample median. From this we observe that the PDH does approximate the smallest levels of acuity observed, but is not a fair estimate of sample performance. We also note that there is substantial variation in the MAD between participants. The dashed curve plots our model for the upper-quartiles, which is discussed in Section 6.3.2: $M = \frac{10.7}{D} + 14.6$, where M is the esitmated upper-quartile MAD value and D is the listening/reference distance.

tance, whilst Figure 6.6 shows the ranked statistics for each reference distance. We have chosen to report ranked statistics, instead of means and standard deviations, as we are interested in how listeners' acuity is distributed around specific threshold values. The PDH seems to approximate the smallest (most accurate) levels of acuity observed, but does not appear to be a fair estimate of sample performance. The large ranges and inter-quartile ranges indicate that there is substantial variation of the MAD between participants. The graph also suggests that an inverse relationship exists between reference distance and participant performance.

The one-sample Wilcoxon Signed Rank test can be used to test whether the sample's median is different to a given value. For the 1.33 m, 1.81 m and 2.88 m reference distances, tests against the null hypothesis that the median equals the PDH value of 5% results in p-values of $9.5e-05$, $1.9e-06$ and $4.8e-04$ respectively. All values are smaller than our alpha significance criterion of 0.05, so we conclude that all three sample medians are significantly different from 5%.

The questionnaire revealed that 100% of participants believed they understood the task required of them, though 8% felt that their answers did not form a correct representation of what they heard. In almost all cases this was due to them being unsure of some of their answers, which we would expect given that the depth differences could decrease to zero. Comments ranged from, *"Very unsure,"* to, *"For some, I am quite sure, but for some tests it is hard to guess which one is near to me."* The only participant who gave a notably different reason appeared to be predicting the experimental design incorrectly, saying they were *"Unsure whether the pitch and volume was kept constant."* Speculating about the cause of the audible differences does not mean they answered incorrectly.

The responses given to Question 5 were categorised according to a selection of popular responses. A single participant could give a response that fell into multiple categories. A 65% majority of the participants reported using loudness as a cue to their responses, whilst 20% reported using a visualisation technique, 17% said they used the tone or pitch of the ring and 10% said they used instinct or feeling. The visualisation techniques included, *"I imagined bells on a line and tried to place each one,"* and, *"I imagined reaching for the phone and determined it on how far I would have to stretch."* A number of people's responses to Question 5 (28%) had some aspect that couldn't be classified as any of the above. This was often due to their response suggesting they didn't really know how to answer, either specifically by saying in one case, *"Don't really know"*, or in other cases by giving answers such as, *"One sounded nearer than the other,"* and, *"I could hear a difference in the quality of the two sounds played, but struggled to connect this with a reference to the distance of the sound."* In some cases an unclassified response did suggest the participant used a cue that was too niche or vague to classify with other responses, such as, *"From the sharpness of the sound."*

6.3 Discussion

In this section we discuss the significance of our results for content creators, system designers and researchers in the field. This begins with a comparison of our results against the PDH value of 5% (Section 6.3.1) and a comparison of our results against those of previous studies (Section 6.3.2). We then argue that, in light of our results, researchers should screen participants for RAD perception acuity (Section 6.3.3). Following this, we report the details of a further minor study undertaken to test the reliability of our experimental design, by investigating the consistency of results when participants repeat the experiment multiple times (Section 6.3.4). Finally, we

identify threats to the validity of our results (Section 6.3.5).

6.3.1 Comparison with the PDH

These results indicate that the PDH value of 5% is not an appropriate estimate of sample performance for our experimental setup. Ashmead et al. (1990) propose that prior studies failed to match the PDH at small reference distances because of their chosen experimental method. They argued that the methods used by previous studies caused participants to adopt a conservative response criterion, thus unfairly biasing the final results for smaller reference distances. In this study we have adopted a similar method to Ashmead et al. (1990), yet our results are consistent with the results they were criticising. This suggests that there are other factors in the experimental apparatus and environment that cause the MAD to increase with smaller reference distances.

Literature reveals a complexity to auditory depth perception that is not acknowledged in the PDH's simplistic approach. As discussed in section 2.2.3, there are other cues to auditory depth that could have been used to inform responses in these experiments, namely the reverberation and frequency spectrum cues. Whilst the inter-aural differences may just be audible, they can still be treated as negligible cues to depth as the loudspeakers are placed very near to the median plane. Ashmead et al. (1990) confirmed that RAD perception is not entirely dependent upon pressure discrimination, although removing the pressure cue does significantly degrade performance. It seems odd, then, that despite the availability of more cues to depth than the PDH acknowledges, performance is significantly worse than the PDH. This suggests that some aspect of the acoustical environment is either confounding the ability to discriminate pressure levels, or the ability to interpret pressure differences.

The size of the MAD is important because it contributes towards determining a benchmark for the level of detail required by a spatial sound system. The level of detail will also depend upon the context within which the spatial sound system is used. Researchers, content-creators and system designers should be aware that the MAD for any individual listener is likely to be larger than the 5% value predicted by theory. However, the PDH does appear to approximate the most accurate levels of acuity in the distributions. Spatial sound systems to be used with TV displays should therefore aim to recreate depth changes of at least 5% of the intended viewing distance. For content designers, different limits are important in different design contexts. If auditory depth is used as a medium to deliver important information,

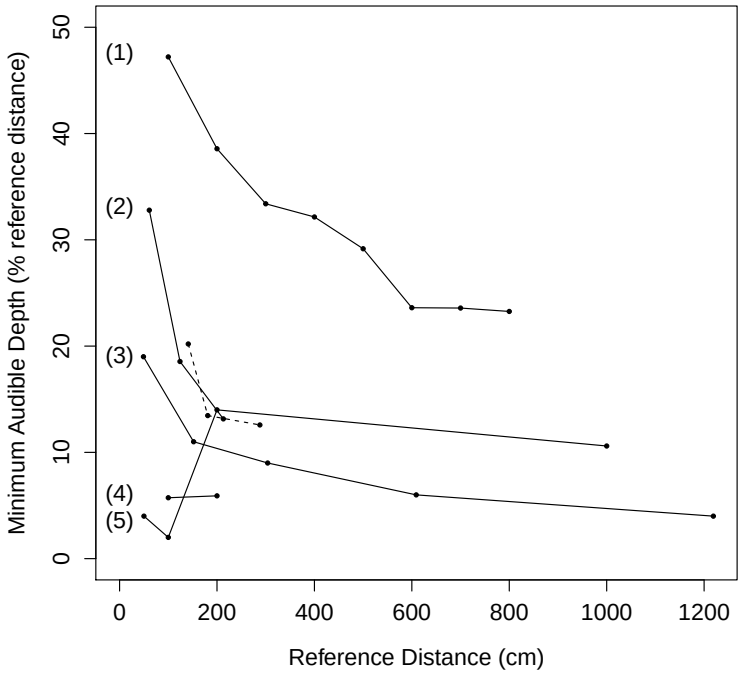

Figure 6.7: Our results, shown by the dashed line, follow a similar trend to the results of several previous studies. The results of previous studies are numerically labelled: (1) Edwards (1955), (2) Simpson and Stanton (1973) (3) Strybel and Perrott (1984), (4) Ashmead et al. (1990), (5) Volk et al. (2012).

then a larger value for the MAD should be considered to ensure the majority of the audience can perceive that delivery. There are many scenarios when this might be the case, such as using auditory depth to create specifc sensation or effect in film, such as multiple gun shots from someone approaching behind the camera, or footsteps approaching the camera in the dark. This might also be important when using audio depth to improve or distract performance in a 3D gaming environment, or even when using audio to improve comprehension of scientific data visualisation. On the other hand, if the purpose of auditory depth is to provide the scene with a degree of fidelity that is valuable for high performing listeners, then one should aim for the smallest MAD values of approximately 5 %. Doing so will give audience members with the best acuity a level of auditory spatial detail they can appreciate.

6.3.2 Comparison with previous studies

Figure 6.7 shows that our medians match the general trend of results from other studies. For instance, the MAD increases for smaller reference distances when ex-

pressed as a percentage of the reference distance – an inverse relationship that is reflected in three of the five other studies. The only study that disagrees significantly with this trend is the work by Volk et al. (2012), where the difference may well be explained by the use of a very different technology to create auditory depth (a Wave Field Synthesis sound system).

Figure 6.7 also shows that there is no real consensus across studies. As mentioned above, Ashmead et al. (1990) argues that the experimental method chosen affects the results. Given that substantial variation exists between our results and those of Ashmead et al. (1990) and Volk et al. (2012), all of which used the same transformed adaptive procedure to measure the MAD, we suggest the environment, stimulus or sound system used also causes substantial variation in the MAD. This implies that researchers, content-creators and system designers should aim to measure the MAD for their own setup wherever possible. If this is not possible, then a rough estimate of the MAD may be taken from data collected using a similar experimental setup.

We have taken several steps to secure the environmental validity of our results, making them a relatively robust data set for a TV viewing scenario. When using auditory depth as medium for deliving information we recommend using the MAD value corresponding to the upper-quartile point in the distributions, rather than the median, to ensure the majority of the audience can distinguish the difference. The MAD will depend upon the intended listening/reference distance, so we recommend using the following model that has been built from our data:

$$M = \frac{10.7}{D} + 14.6 \tag{6.1}$$

Where M is the recommended MAD value expressed as a percentage of the reference distance and D is the intended listening/reference distance in m. Ashmead et al. (1990) observed that all other data sets, excluding the more recent work by Volk et al. (2012), have a reciprocal form. We know the data should tend to infinity at the origin as any percentage of 0 is infinite, so we have selected just two degrees of freedom to give the final form $y = a/x + b$. Estimates of the constants a and b were made using an evolutionary algorithm to minimise the chi-squared statistic. The final fit, which is plotted in figure 6.6, gives a chi-squared statistic of just 0.02, indicating a good fit.

Only the study by Edwards (1955) yielded MAD values larger than ours, and it is the oldest study we are aware of that investigates RAD perception. Very little information is given concerning the environment in which the experiment was performed, leaving us to speculate on how it may have impacted their results. We are

told, however, that the participants sat with their back to the stimulus. Considering our understanding of the role that the pinna and body play in auditory localisation, it seems possible that acuity in front of the head differs from acuity behind the head. Our results are most similar to those acquired by Simpson and Stanton (1973). Their study was performed in a semi-reverberant room using loudspeakers and a complex stimulus in a similar manner to the work we are presenting. They did, however, adopt a significantly different experimental design, using the method of limits rather than a transformed adaptive procedure. The study by Strybel and Perrott (1984) yields results that are a little smaller than ours, though not that dissimilar. Their study was undertaken outdoors, and as such, one would expect significantly less reverberation than in the studies already discussed. This does mean that some background noise would be expected. The only study found to approximate the PDH was that by Ashmead et al. (1990). This study was implemented in an anechoic chamber using a single loudspeaker, which removed the problems associated with frequency response matching. One might speculate whether these studies suggest reverberation increases the size of the MAD. The mixture of different auditory reflections may degrade your ability to detect loudness differences. Further experimentation could explore the role played by reverberation, which aids absolute auditory depth perception (Bronkhorst and Houtgast 1999), in RAD perception.

6.3.3 Screening for RAD perception

Figure 6.6 shows that there is substantial variation of the MAD between participants. This means that the optimal MAD value to be used by content creators, system designers and researchers will depend upon their intended listeners. This does not render auditory depth useless for conveying information, as binocular depth perception in S3D images is popular despite also being subject to variation between viewers. However, it does imply that experimenters should screen participants for RAD perception acuity, in a similar manner to screening them for stereoscopic acuity. Doing so will help ensure their results are not unfairly biased by including participants with very poor RAD perception.

In order to formulate a repeatable screening test, one needs to understand population acuity, for which we require much larger samples than those used. The perception of sound depends upon the listening environment, which further complicates the design of a repeatable screening test. We therefore encourage researchers to screen the participants based upon the specific context and application of their experiment. Participants may be ranked according to their score in a number of

2AFC tests such as those used in our experiment. The number of tests will depend upon the design of the experiment and the time available, as they should use the auditory stimuli and depth differences of interest. The number of poorest performing participants to be selected should also depend upon the design and intended application of the experiment – but as an initial suggestion, researchers may wish to exclude the poorest performing quartile of participants.

6.3.4 Consistency in participant thresholds

It seemed sensible to gain some estimation of participant consistency over repeated tests, which would give an indication of the error in each individual's result. The simple experiment outlined in this section was run as a sanity check to support the reliability of each measured data point.

Two participants repeated the 1.81 m experiment 10 times each, using the procedure in Section 6.1. Both participants had contributed to the main experiment. Each repeat included the training and screening test, though only one questionnaire was completed at the end of all 10 repeats. The repeats were split across five sessions, scheduled for the same time of day, and completed within 8 days. Each participant undertook two repeats in each session, with a short break between them. The participants were female postgraduate students aged 23 and 24.

The results are shown in table 6.1. Participant A's mean threshold was measured to be 3.73% with a standard deviation of 1.19% and a range of 1.04%-5.18%. Participant B's threshold was measured to be 7.18% with a standard deviation of 1.28% and a range of 5.18%-8.63%.

Subject B gave one result that was discarded due to the participant complaining of extreme tiredness. Their participation during the fourth session was halted after the first test in which their MAD was measured to be 46.87 cm (25.87% of the reference distance). The participant had already declared their tiredness during the test and so it was decided to postpone the session and disregard the uncharacteristically large result (4.96 standard deviations larger than the mean). This indicates that performance may depend upon how tired the participant is, which is a hard factor to control.

The standard deviations of both participants are strikingly similar. The values for both standard deviations are larger than the final step size of 0.625 cm in the experimental procedure. This suggests that the minimum step size is not a fair estimate of the error in the measurements.

The post-experiment questionnaires revealed that both participants used loud-

Test Number	Subject A MAD		Subject B MAD	
	cm	% ref. dist.	cm	% ref. dist.
1	1.87	1.04	9.37	5.18
2	6.87	3.80	13.10	7.24
3	9.36	5.17	11.87	6.56
4	8.12	4.49	9.37	5.18
5	9.37	5.18	15.61	8.63
6	6.87	3.80	15.62	8.63
7	6.87	3.80	11.87	6.56
8	5.62	3.10	14.37	7.94
9	5.62	3.10	14.37	7.94
10	6.87	3.80	14.37	7.94
Mean	6.75	3.73	12.99	7.18
St. Dev.	2.16	1.19	2.36	1.28
Minimum	1.87	1.04	9.37	5.18
Maximum	9.37	5.18	15.62	8.63

Table 6.1: The results of ten repeated tests upon two participants. Results are given in both cm and as a percentage of the reference distance which was 1.81 m.

ness to determine which cue was nearer, whilst participant B said they also used pitch. As participant B's result was worse than participant A's result, one might speculate on whether using pitch confounded the judgement rather than improved it. Both said that repeating the test could have improved their performance, with Participant B saying, "The differences seemed to become more pronounced as I repeated experiments." Despite this, neither participants showed any significant learning effect in their results.

6.3.5 Threats to validity

Whilst implementing the experiment, we noticed that a rounding error could cause the size of the depth differences to deviate from the expected values by up to 2 mm. PEST only allows for the halving or doubling of step sizes, meaning all depth differences tested should be a multiple of the final minimum step size. A rounding error in the software caused a small error in the calculation of the new step size. Because these steps are then added and subtracted between tests, this error could add up to a couple of millimetres over the course of an experiment. The decision to halve the step size upon a reversal is a matter of efficiency and not of precision. A slight error in the halving should not damage the precision of the final result, but may increase the number of tests required to reach the final result. As such, this

error should not pose a threat to the validity of the experiment's results.

The motorised platforms used to move the loudspeakers were not silent. This introduced a noise between each test that depended upon the speed and distance that the loudspeaker was being moved. In its natural form this could indicate the nature of the loudspeaker arrangement used in one test relative to that of the previous test. For instance when the loudspeaker arrangement didn't change, there would be no motor noise at all. One participant did comment on this in the post-experiment questionnaire saying, *"Length of motor movement is audible, suggesting similarity to previous test when short?"*. However, it is important to note that *"similarity to the previous test"* is different to the test's correct answer. We designed the experiment so that the motor noise could not give the participant a cue to the correct answer. The correct answer was determined by the order in which the two loudspeakers played the stimuli. This order was randomly decided by the software and remained completely independent of the loudspeaker arrangement and thus the duration of motor noise.

The side-by-side arrangement of the loudspeakers did introduce small inter-aural differences that may have been audible to some participants. Whilst such small inter-aural differences would not have affected auditory depth perception, it did mean that the correct answer to each test had to be randomised across the potentially audible left-right cue. This was done by randomly deciding whether to switch the near loudspeaker before each test. This left-right switch of loudspeakers required motor movement, meaning it was not independent of motor noise – no motor noise indicated that no switch had occurred. By splitting each speaker's movement into two steps we were able to hide the cases where no motor noise occurred. This was done by introducing a movement in one direction followed by a movement back to the original position, when the loudspeaker arrangement did not need to change between tests. We therefore believe that our experimental design removed any threat to the internal validity of our results posed by the motor noise.

6.4 Conclusions

We have implemented a two-down one-up transformed adaptive experiment with PEST to measure the MAD difference in a TV viewing scenario. Acuity in RAD perception, which is the task of correctly distinguishing between two sound sources at different depths, can be measured using the MAD. A pair of frequency matched loudspeakers, placed about the listener's median plane, were physically moved in depth using motorised platforms that were controlled by a computer. In order to

give our results environmental validity, the loudspeakers were positioned in front of a TV screen, in a semi-reverberant environment with some background noise. Three different listening distances were investigated, each corresponding to a recommended viewing distance for the TV screen: a 20°, 30° or 40° viewing angle. At each listening distance, we measured the MAD for twenty participants who had previously been screened using the BSHAA online hearing test.

The results of this study have implications for researchers, content designers and system designers who are interested in using auditory depth to convey information. For the three listening (or reference) distances of 1.33 m, 1.81 m and 2.88 m we found the median MAD values to be 20.2%, 13.5%, 12.6% respectively, when expressed as a percentage of the listening distance. These results are plotted in Figure 6.6, which shows that the PDH value of 5% is a reasonable approximation of the most accurate levels of acuity. For each listening distance, Wilcoxon Signed Rank tests give p-values rejecting the null hypothesis that the median is equal to 5%. From this, we conclude that the PDH is not a good predictor of sample performance. Whilst it may be considered the gold-standard for system design, content creators should use a more conservative MAD value in order to include a greater proportion of the population. We therefore recommend using the upper-quartile MAD value in order to include the majority of the audience. Using our data set, we have built a model (plotted in Figure 6.6) that estimates this upper-quartile MAD value M for a given listening/reference distance D in a TV viewing scenario:

$$M = \frac{10.7}{D} + 14.6$$

Substantial variation of the MAD occurred between participants. In this respect it is similar to stereoscopic depth perception acuity. We conclude that those experimenting with RAD perception should consider screening participants for RAD perception acuity by measuring their MAD values. The form of this screening procedure will depend upon the context within which auditory depth is used. Further studies could use much larger samples to understand population acuity, which would inform the design of repeatable screening tests. In the meantime we propose ranking participants according to the size of their MAD value and removing those with the largest MAD (indicating poorest acuity). The proportion of participants to be removed will depend upon the application of the research. The easiest means of ranking participants is to use the score from a number of 2AFC tests in which the participant has to select the nearest of two sound sources. These tests should use the auditory stimuli and depth differences of interest.

Further research is required to understand how different factors influence the MAD. Such research would help formulate a means of predicting the MAD for a given technology, acoustical environment and stimulus. In the meantime, we recommend researchers measure the MAD for the setup they are interested in, or alternatively design their setup to match an example from literature as closely as possible. Another avenue of further research may explore whether dividing up continuous space using discrete data sets for the minimum audible height, depth and azimuthal angle would offer a means of improving the compression of 3D sound fields. Such sound fields would thus have a resolution, not unlike visual displays, that approximates continuous space. Our immediate work has used the setup reported in this study to investigate cross-modal effects in depth perception for 3D audio-visual media. This will build upon the preliminary work outlined in Chapter 4.

CHAPTER 7

Evaluating the cross-modal effect

We now return to exploring the cross-modal effect identified by the preliminary experiment. We are particularly interested in whether the cross-modal effect identified in the preliminary experiment could extend the S3D depth budget. To explore this further we need to re-evaluate the effect using a calibrated experimental setup and an improved experimental method. In this Chapter, we report an experiment that re-visits the quality-control of audio-visual depth cues, building upon the results, experiences and equipment gathered from the work outlined in Chapters 5 and 6.

Evidence of a cross-modal depth perception effect has already been reported in Chapter 4, which outlines a preliminary experiment that suggests an auditory stimulus can influence the apparent depth of a visual stimulus in an S3D display. We have already discussed in Section 1.2 the potential for this effect to extend the limited depth budget associated with 3D displays. In this chapter we seek to confirm the results from the preliminary trial, whilst also gaining a more detailed understanding of both the quantitative and qualitative nature of the effect. As such we will answer subsidiary research questions 4 and 5 as reported in Section 1.2.

Wichmann and Hill (2001) define the psychometric function as a function that "relates an observer's performance to an independent variable, usually some physical quantity of a stimulus in a psycho-physical task." Such a function for this cross-modal effect would be valuable when considering its practical application. Measuring this function would therefore seem a natural development of our preliminary trial.

We would expect the cross-modal effect to have both upper and lower limits with respect to audio-visual separation. For small audio-visual separations, the spatial difference may not be perceivable and so no cross-modal effect should be expected. For particularly large audio-visual separations, the perceived spatial unity of the

audio and visual component would likely be broken, causing no further cross-modal bias to be created (Hairston et al. 2003b). These limits should be reflected in the shape of the psychometric function. Obtaining an estimate of these limits would be useful when considering the application of the effect. For instance, they could be used to build a mapping between visual space and audio-visual space.

The work in this chapter therefore has four distinct aims:

- To repeat our preliminary trial using calibrated equipment and auditory depth screening.
- To form some understanding of the effect's psychometric function (its dependency upon the audio-visual separation).
- To take from this function estimations of the lower and upper limits of the effect with respect to audio-visual separation.
- To learn more about the qualitative nature of the effect.

This chapter reports a single experiment in the following manner. Section 7.1 outlines the method used, including the experimental design, setup, participants, and the means by which qualitative data was captured. Both the quantitative and qualitative results are presented and discussed in Section 7.2 before we evaluate our work and compare the validity of our results with those of the preliminary trial in Section 7.3. We finish by summarising our work whilst drawing together conclusions and considering future avenues for research in Section 7.4.

7.1 Method

The method used in this experiment is very similar to the method outlined in Chapter 4, though there have been a few additions and improvements. We begin this section by detailing the experimental design in Section 7.1.1, followed by the equipment and setup in Section 7.1.2 and the design of the qualitative analysis in Section 7.1.3. We then outline the RAD perception screening test in Section 7.1.4 before finishing with details of the participant sample in Section 7.1.5.

7.1.1 Experimental design

The experimental design was based upon our preliminary trial in order to confirm its results. The same 2AFC test design was used, but this time the audio-visual separation was allowed to vary between tests. The test required the participant to

judge the depth of a mobile telephone (shown in Figure 4.1 on page 59) in two consecutively displayed images. Whilst the visual depth of the phone did not change, it was accompanied by an auditory stimulus that did change in depth. The auditory stimulus, which was the same telephone ring reported in Chapter 4 and used in the previous chapter, could be played from either of two loudspeakers, one positioned at the same depth as the visual stimulus and one positioned in front of the visual stimulus. For each test the participant answered the question, "Which phone appears nearest?"

Participants were positioned at the viewing distance corresponding to a 30° viewing angle, which was 1.91 m. This distance was chosen using the measurements of the MAD collected in Chapter 6. When these measurements are expressed as a percentage of the reference distance, they prove significantly worse for the nearer distance that was tested (40° viewing angle) and insignificantly better for the further distance that was tested (20° viewing angle). Four different audio-visual separations were tested: 0 cm, 25 cm, 50 cm and 75 cm (the range of achievable audio-visual separations was limited by the equipment used). The interval between these values was chosen to be 25 cm, because this was approximately the MAD we measured in Chapter 6.

Repeat tests were taken at each audio-visual separation, so that a non-binomial participant score could be calculated by taking the mean score over all the repeats. This score contributed towards a mean and standard deviation across all participants that was used for statistical analysis. The number of repeats, set to 16, was limited by the time and resources available. As in the preliminary experiment, participants began the experiment by undertaking four dummy tests, the results from which were discarded.

7.1.2 Experimental setup

The experimental setup that was used in this study is outlined in Figure 7.1. It was similar to the experimental setup reported in Chapter 6 and outlined in Figure 6.3. Due to the audio-visual nature of this trial we were unable to blindfold our participants, so a thin black cloth was used to conceal the speaker arrangement. By reversing the rails that supported the motorised platforms, we were able to make use of a greater length than in the study for Chapter 6. This enabled us to create the desired maximum audio-visual separation of 75 cm.

An LG BM-LDS302 47 inch 3DTV screen was used to display the visual stimulus, whilst two K-Array KT20 loudspeakers, driven by a Cambridge Audio Topaz

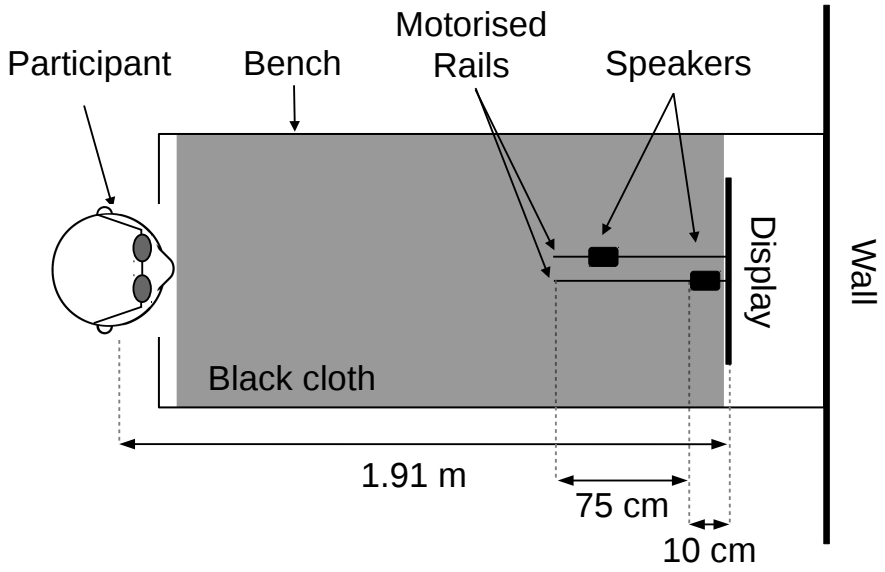

Figure 7.1: The experimental setup was very similar to that outlined in Figure 6.3 with the exception that a thin black cloth kept the speaker arrangement hidden from the participant and the motorised rails were reversed to get some extra length.

AM1 amplifier, were used to play the auditory stimulus. These loudspeakers have been identified by the manufacturer as having well matched frequency response curves. The experiment was performed in the same semi-reverberant laboratory with a background noise level of approximately 41.5 ± 0.3 dB. The volume of the auditory stimulus was set to be a maximum of 70.0 dB at the approximate listening position, when the loudspeakers were positioned at the position furthest from the participant. All equipment was placed upon a desk that extended from the screen to the participant. The centre of each loudspeaker's front face stood 14 cm above the desk, whilst the chin rest stood 37 cm above the desk.

The motorised rails, loudspeakers and display were all controlled by a computer program written in the Java programming language, using the Java open graphics library (JOGL) to render the graphics. The visual stimulus was the same mobile phone used in the preliminary experiment and shown in Figure 4.1 (page 59). The program was controlled by the experimenter who initialised and entered each participant's verbal response for each test. For each participant, the program generated a text file which stored their results and test details.

7.1.3 Qualitative analysis

It was decided to run the qualitative data capture in the form of a semi-structured interview, as it is difficult to get consistently detailed responses on such a complex

topic from a questionnaire. In a semi-structured interview there is a list of questions and themes to be discussed, but the order of the questions is flexible and further probes may be asked as appropriate, depending upon the conversation (Oates 2006). The aim of this interview was to establish the participant's conscious thought patterns whilst taking part in the experiment. We were particularly interested in how the participant believed they combined audio and video to give a final response. Prior to the experiment, the following six questions were prepared:

1. Did you clearly understand the task required of you? (Yes/No answer required)
2. How sure are you that your answers formed a correct interpretation of what you perceived?
3. How did you decide which phone was nearer?
4. How would you describe the nature of the depth changes you perceived when determining which phone was nearer? Consider such factors as clarity, approximate size and variation across tests...
5. Did the sound of a telephone ringing consciously contribute to your decision in any way?
6. Do you have any other significant comments regarding the execution of the test?

Questions 1 and 2 checked that the participants were comfortable with the experiment and felt able to give appropriate responses to the experimental task. Question 3 sought to understand how audio and visual information were combined to make a final response. Question 4 built a qualitative picture of the audio-visual perceptions that participants used to complete the experimental task. Question 5 directly asked how their perception of the sound contributed to the decision making process. This was left until the end of the interview, due to concerns that it might increase the chance of hypothesis guessing, and thus bias their other answers.

7.1.4 Screening participants for RAD perception acuity

The results in Chapter 6 clearly indicate that participant acuity in RAD perception can vary significantly. The literature reviewed in Section 3.2.3 tells us that cross-modal depth perception depends upon auditory depth perception, so it was deemed necessary to design a means of screening participants for RAD perception acuity. This screening test was run at the end of the experiment after the post-experiment interview, in order to minimise hypothesis guessing. To our knowledge, this is the

first time that anyone has screened participants for auditory depth perception acuity. We include in this chapter an analysis of the test's impact upon our results.

The initial screening test required the participant to answer correctly four sound-only tests for the MAD, which in the previous chapter was measured to be 25 cm. That is, they heard one far and one near sound, randomly ordered, before being asked which sound appeared nearest to them. The participant was blindfolded whilst undertaking this test, in order to stop any cross-modal effects impacting their answers. Out of the five participants who undertook this screening test, only one participant successfully passed. One other participant responded correctly to two tests and the other three participants responded correctly to just one of the tests.

The MAD is taken as the depth difference for which participants correctly respond to approximately 75% of the tests. Further to this, our MAD value of 25 cm corresponds to the depth at which we might expect half of our participants to meet this criteria. We therefore decided it was appropriate to relax our screening criteria, as we wanted a larger proportion of subjects to pass the screening test. It seemed sensible to re-design the screening test so that we collected more information about the participant's acuity for different depth differences. The screening criteria could then be decided at the end of the experiment, based upon participant performance.

The new screening test required participants to respond to four sound only tests for each of the audio-visual separations used except zero: 25 cm, 50 cm and 75 cm. The screening test took approximately four minutes to complete. As in the vision screening tests, such as the Snellen eye test or the Titmus stereo test, participants were allowed to experience the stimuli as many times as they wished before giving an answer.

7.1.5 Participants

The participants were undergraduate students of Durham University, recruited largely from the engineering course. Prior to participating in the experiment, all participants were required to pass the Snellen eye test for 20/20 vision, the Stereo Titmus Test and the BSHAA online hearing test.

The first participant's results had to be discarded due to an equipment failure partway through their experiment. Results of the next five participants were collected successfully, but due to a redesign of the post-experiment auditory depth perception screening test (explained in Section 7.1.4), they too were discarded from the analysis below.

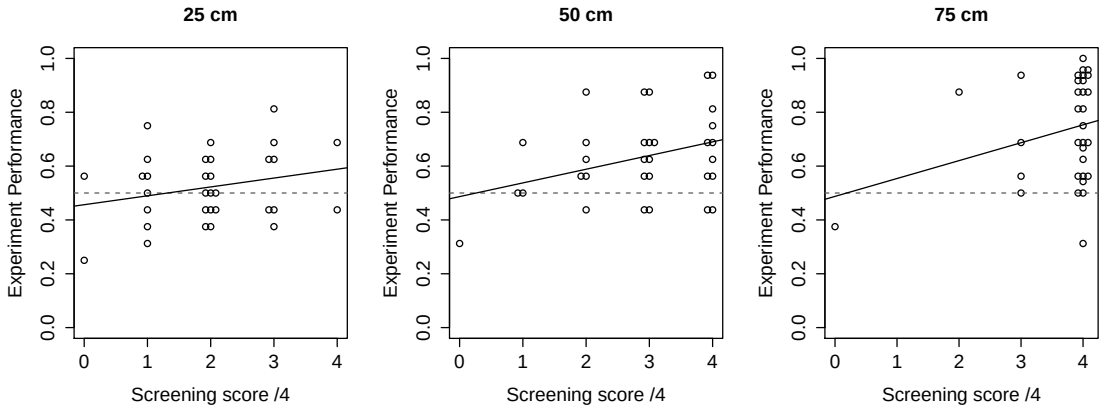

Figure 7.2: The relationship between participants' performance in the experiment and their score in the screening test, for each of the three non-zero depth differences. Audio-visual depth perception should depend upon audio depth perception, so we would expect a positive correlation between screening test and experimental task performance. The data is heavily discretised in both dimensions, resulting in a large number of overlapping data points. To overcome this, we have clustered overlapping data points together – all data points that touch are part of the same cluster.

Depth /cm	Gradient	Std.Err.	p-Value	Pearson's
25	0.03298	0.02325	0.1663	0.2507
50	0.05083	0.02477	0.04896	0.3509
75	0.06643	0.04080	0.1140	0.2849

Table 7.1: The gradients for the linear regression in Figure 7.2 with their standard errors and p-values for finite sample F-tests against the null hypothesis that the gradient equals zero. The Pearson product-moment correlation coefficient is also shown.

Some 38 participants passed the pre-experiment screening tests and were thus paid for their participation. After discarding the results of six participants as stated above, the median age of the remaining 32 participants was 19, with an inter-quartile range of 19-20 and a total range of 18-35. Twenty-seven of the 32 participants were male.

7.2 Results and analysis

We begin by evaluating the RAD perception screening test in Section 7.2.1, specifically with a view to establishing whether the screening tests results offer any means of predicting performance in the cross-modal task. We then report the task performance results in Section 7.2.2 from which we plot the psychometric function. In Section 7.2.3 we fit a logistic model to the data and use it to predict the effect's

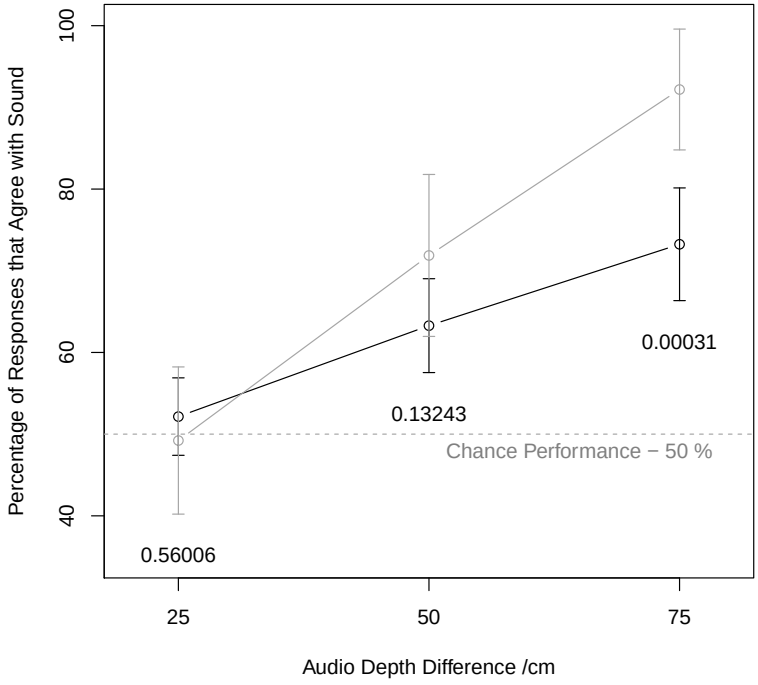

Figure 7.3: Evidence of the cross-modal ventriloquist effect is observed when comparing the audio-only data from the post-experiment screening test with the audio-visual data from the experiment. The audio-only data from the screening test is coloured grey, whilst the audio-visual data from the experiment is coloured black. Each pair of points is labelled with a p-value for a two sample t-test against the null hypothesis that the difference between their means is zero. For the 75 cm depth difference a two sample t-test reveals that there is a significant difference between the audio only case and the audio-visual case. This suggests that vision was impacting the participants' responses.

limits with respect to audio-visual separation. Finally, in Section 7.2.4 we report qualitative responses gathered in the post-experiment interview.

7.2.1 Evaluating the RAD perception screening test

The RAD perception screening test was used to remove all the participants whose score was in the bottom quartile for any of the three depth differences tested. For the depth differences of 25 cm, 50 cm and 75 cm, the participant was therefore required to score a minimum of 1, 2 and 4 out of 4 respectively to pass the screening test. This resulted in a final sample size of 23 participants.

Figure 7.2 shows the relationship between scores in the screening test and performance in the experiment. If the screening test has served its purpose, we would expect better performance in the screening test to correspond with better performance in the experiment. Linear regression can be used to determine whether this is the case. Each graph in Figure 7.2 includes a least squares linear fit to the data, all

of which have a positive gradient. This suggests that our screening test is having an impact, but we do not know whether these positive gradients are statistically significant. Finite sample F-tests against the null hypothesis that the gradient equals zero were used to determine this. The gradients, standard errors and p-values for these tests are shown in Table 7.1, along with the value of the Pearson's product-moment correlation coefficient.

From the Figure 7.2 and Table 7.1 we conclude that screening the participants for the 50 cm audio depth difference had a significant impact upon the experimental results. We are unable to make statistically significant conclusions concerning the impact of our screening test for the 25 cm and 75 cm depth differences. For the 75 cm depth difference, the screening test results are not evenly distributed across the possible scores; only the lower quartile of participants gave an incorrect response. A participant sample with greater variation in screening test performance at this depth may have resulted in the screening test having a significant impact on experimental results at this depth too. Given that the 50 cm audio depth difference yielded a statistically significant positive result, and given that the other two cases yielded positive results, we conclude that the screening test was a valuable addition to our experimental method.

In Figure 7.3 we plot performance in the post experiment screening test against performance in the experiment, allowing us to compare audio-only data with audio-visual data. There is a statistically significant difference in the data for the 75 cm depth difference, where participants perform better in the audio-only case. One can interpret this as: the addition of a constant visual depth reduced the participants' ability to distinguish between the two different audio depths. Or in other words, the participants were experiencing the traditional ventriloquist's effect. We are interested in the inversion of the ventriloquist effect, which as discussed in Section 3.2.3, could occur simultaneously; both audio and visual components could be biased towards each other, with neither appearing at their original location.

7.2.2 Task performance

The distributions were made up of 23 participants' data. Only the data for the 75 cm audio-visual depth separation passed the Shapiro-Wilk test for normality with an alpha significance criterion of 0.05, meaning we may wish to consider the distribution as non-normal. For a sample size of 23, the Shapiro-Wilk test for normality proves very sensitive to non-normality. The skewness and kurtosis values for the distributions remained within the range of -1.2 to 0.2, which we judge to

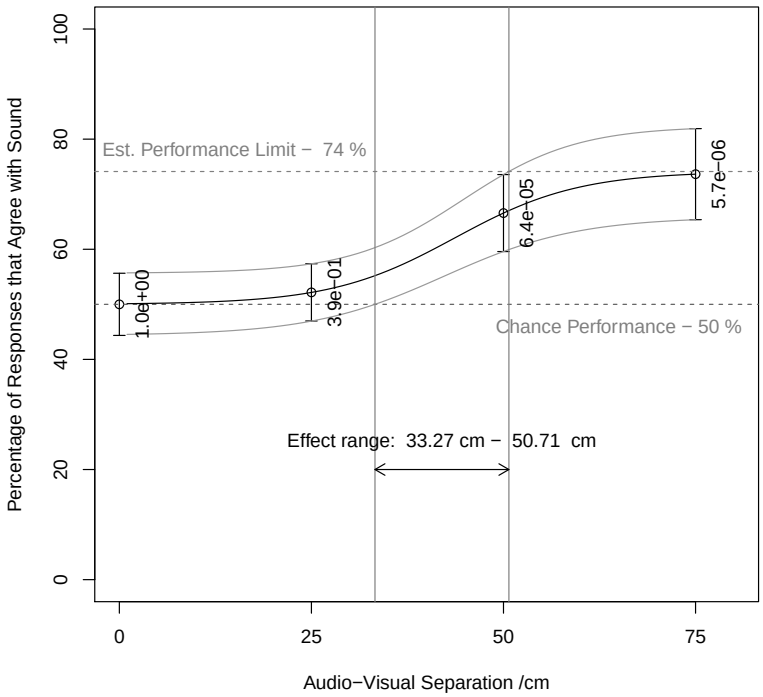

Figure 7.4: Participant performance for each audio-visual separation with 95% confidence intervals. For each mean we also show the p-value for a single sample t-test against the null hypothesis that the mean is equal to 50%. A logistic curve constrained to intercept the y-axis at 50% has been fitted to the data. Further unconstrained logistic curves have been fitted to the 95% confidence interval values. The effect's limits are taken as the range for which we have 95% confidence that performance lies between the limits of 50% and 74%.

be acceptably close to zero in order to use a parametric analysis upon the data – particularly as parametric tests are reputedly insensitive to non-normality (Glass et al. 1972; Lix et al. 1996).

For each audio-visual depth separation 0 cm, 25 cm, 50 cm and 75 cm, we found respectively that a mean of 50%, 52%, 67% and 74% of participants' responses said that the phone accompanied by the nearer telephone ring appeared nearer. For each audio-visual separation, t-tests were used to determine whether we can reject the null hypothesis that the mean score equals 50%. The p-values from these tests are also shown in Figure 7.4. The tests yielded strongly significant results (more than 95% confidence) for 50 cm and 75 cm audio-visual separations, but failed to achieve significance for the 0 cm and 25 cm separations.

A p-value of 1.0 was obtained for the 0 cm audio-visual separation, which is good as we would expect chance performance in this case; and indeed, any deviation from chance would indicate a fault in our experimental setup. The t-test for the audio-visual separation of 25 cm also failed to show significance for both the cross-modal

experiment data and the audio-only screening test data. This suggests that a more conservative MAD value should have been adopted for this experiment; perhaps because 25 cm represents the depth difference for which half of our participants could correctly distinguish between the two sources half of the time.

A one-way repeated measures ANOVA was used to determine whether there are differences between each audio-visual depth separation's data. The test proves statistically significant with greater than 99% confidence. Further analysis using two sample t-tests reveal that the 50 cm and 75 cm data distributions are significantly different from the 0 cm and 25 cm data. This is further evidence that a cross-modal effect is occurring in the task, as it shows that task response depends upon the audio-visual depth separation used.

7.2.3 Estimating the limits of the cross-modal effect

The performance of the screened participants for each audio-visual separation is shown in Figure 7.4. A logistic function, constrained to intercept the y-axis at 50%, was used to model the data and estimate an upper-performance limit of 74%. Further logistic curves have been used to interpolate the 95% confidence region around our model. The effect range was then estimated as the range of audio-visual separations for which the 95% confidence region remained entirely within the performance limits. This was calculated to be between 33.27 cm and 50.71 cm.

The decision to fit a logistic curve to the data was based upon three things: our expectation of the data's shape, the actual shape of the data and literature concerning sensory thresholds. As discussed at the beginning of this chapter, we would expect there to be an upper limit to the cross-modal effect, as there will be a limit to apparent spatial unity of the audio and visual component. The data appears to reflect this expectation, with performance levelling off at a performance limit of approximately 74%. As the visual component does not change between tests, it seems sensible to assume that the cross-modal psychometric function will inherit its form from the sound-only scenario. When vision is removed, the participant's task becomes a signal detection problem - can they correctly detect the auditory depth difference between sounds? The literature tells us that the psychometric functions for such tasks are invariably smooth s-shaped curves, such as the logistic curve (Palmer 1999).

7.2.4 Qualitative analysis

The data gathered from the post-experiment interview can be analysed using a process known as *coding*, whereby quantitative data is extracted from qualitative data (Seaman 1999). This extracted data can then be used with the quantitative experimental results in some appropriate statistical analysis. One implementation of coding requires the grouping of participants according to certain themes in the interview, in order to look for statistically significant differences between the grouped results. The themes chosen for this experiment are:

1. The participant's confidence in their responses.
2. How the participant felt they combined audio and vision when making their responses.
3. The nature of the depth difference that informed the participant's response.

In order to investigate the first theme we used the qualitative data to split participants into three groups. Participants were grouped according to whether they felt confident in their responses, uncertain, or in-between these two extremes. This was largely based upon their answer to Question 2 in the interview. Fourteen participants were found to be confident, whilst eleven were unsure of their answers and the remaining twelve fell in-between the two groups. A one-way ANOVA for each audio-visual separation did not reveal any significant differences between groups, though those who were uncertain did score less on average than those who were confident. Furthermore, no significant relationship was found between confidence in responses and our second theme for study: how the participant felt they combined audio and vision to make responses.

Using their answers to Question 3 and 5 in the interview, participants were again split into three groups depending upon whether they primarily used audio, vision or both senses to make their responses. The number of participants in each group were found to be 14, 10 and 13 respectively, or 10, 6, and 7 after applying the RAD perception screening test. This grouping was not applied when estimating the psychometric function because we are unsure how reliable the self-assessment methodology adopted here is. For instance, half of the participants who felt they primarily used sound still reported perceiving a visual depth difference in Question 4 of the interview.

Figure 7.5 shows the results of those who passed the RAD perception screening test and who didn't primarily use audio to give their responses. Thirteen participants satisfied this criteria and passed the screening test. Their results are still

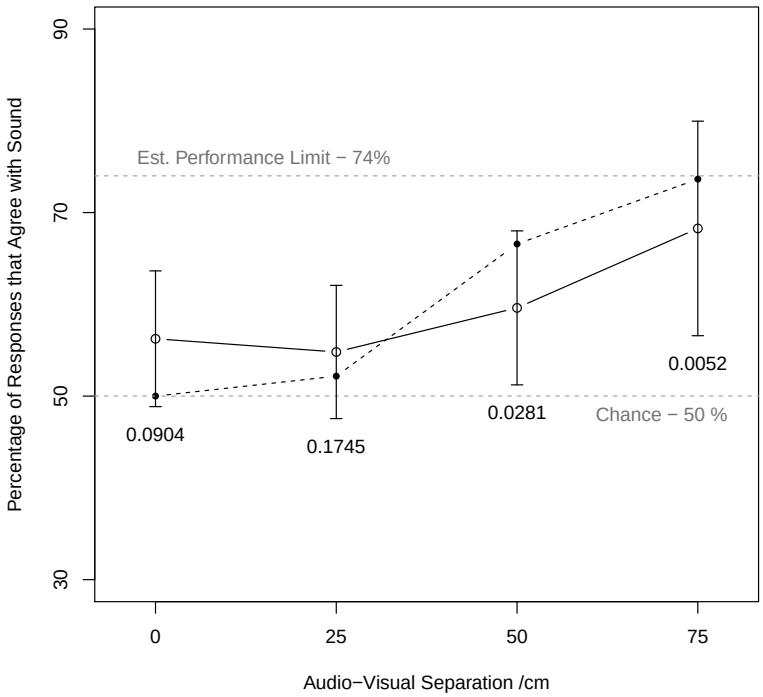

Figure 7.5: The effect of screening out participants who reported using primarily sound to make their responses. The dotted black line indicates the results for the 23 screened participants that passed the screening test and contributed to the analysis in Section 7.2.3. Using each participant's responses in the post-experiment interview, we were able to identify those participants who had primarily used audio information to give their responses. Having removed these participants from our screened sample, the solid black line indicates the results of the remaining 13 participants. The error bars indicate 95% confidence intervals and are accompanied by p-values for t-tests against the null hypothesis that the mean equals 50%.

significantly different from chance for the 50 cm and 75 cm audio-visual separations. By removing participants who consciously answered primarily using audio information, often quoting a lack of visual information as the reason, we can gain confidence that these results are indicative of a cross-modal, perceptual effect rather than a biased response. It is important to note, when considering the effect's potential commercial use, that the number of participants removed is substantial – almost 40%.

Finally, we grouped participants by whether they believed the size of the depth change that informed their responses in each test varied across the whole experiment. This information was gathered in Question 4 of the interview. The number of participants who believed the depth difference varied was 25, whilst just 5 participants believed the depth difference was held at a constant value across the whole experiment. The remaining seven participants could not clearly be placed into either group. Applying our auditory depth perception screening criteria reduces these

numbers to 15, 4 and 4 respectively. No significant differences in participant performance were found between those who believed the depth changes varied and those who believed they didn't. The significant majority of participants felt the perceived depth difference that informed their responses varied in size during the experiment. This supports the logistic nature of the psychometric function, suggesting the effect is not simply "turned on", as if modelled by a step function.

As part of Question 4, participants were probed to give an estimate in units of length of the maximum depth change they perceived during the experiment. It should be noted that not all participants felt comfortable doing this, and for those who did give an estimate, the level of accuracy attached to that measurement varied substantially. Despite this, it is still of speculative interest that 27 of the 37 participants did report perceiving a visual depth difference for which they could estimate the size. All but three of the remaining ten participants were found to have primarily used sound in giving their responses. These estimates range from 0.1 cm to 20 cm, with a median of 1 cm and an inter-quartile range of 0.5 cm to 3.5 cm. This suggests that auditory bias of visual depth in S3D images may be of a size that is genuinely useful in application, particularly in desktop and mobile S3D displays which have very small depth budget. However, we need to acknowledge the limitations of the methodology adopted here, and suggest further research would be needed to provide better evidence.

7.3 Evaluation and comparison with the preliminary trial

The audio-visual separation in our preliminary trial was 20% of the reference distance (from the viewer to the far speaker), for which we found that 65% of participants' responses matched the audio component; that is, that 65% of the responses said the phone accompanied by the closer auditory stimulus appeared nearer. The model of the effect presented in this chapter predicts that for an audio-visual separation of 20% of the reference distance, only 57% of participants' responses would match the audio component. Although our interpolation of the 95% confidence interval still predicts this figure to be significantly different from chance, there is a notable discrepancy between the values of 65% and 57% that calls for an evaluation of the two experimental methods.

Since the preliminary trial, there has been significant investment in developing a calibrated audio-visual display system. The sensitivity of the MAD to environ-

mental factors and the equipment used became apparent during the preliminary sound trials outlined in Section 6.1.1. The K-Array KT20 loudspeakers used in this chapter's experiment were selected for their small size and well matched frequency response curves. After we noticed the frequency matching was broken by the above-below configuration of the loudspeakers, further consideration was given to their positioning. The Logitech loudspeakers that were used in an above-below configuration for our preliminary trial were large and no thought had been given to the pairwise matching of their sounds. Other display system improvements include the precise positioning of loudspeakers using computer-controlled motorised platforms and a significant reduction of aliasing in the visual stimulus.

Further to using calibrated equipment, this study implements a RAD perception screening test to lend credibility to its results. Prior to the analysis in Chapter 6, we had little understanding of the variability in RAD perception acuity across participants for our experimental setup. The chapter reveals that the MAD depends significantly upon the acoustic environment and varies substantially between participants. In this experiment we have removed the results from those participants who appeared to struggle with the task of distinguishing between two different auditory depths.

We might have expected these improvements to yield a stronger result, but there are several possible reasons why this was not found to be the case. For instance, the uncalibrated sound system could have been biasing responses in favour of a correct answer. If there is an audible frequency response mismatch between the two loudspeakers that participants interpret as a cue to auditory depth, it will bias their RAD perception acuity and thus also their cross-modal results. The near/far ordering of the loudspeakers determines whether this appears to strengthen or weaken participants' RAD perception acuity. We saw this happening in the preliminary audio trials reported in Section 6.1.1.

In both this experiment and the preliminary trial, qualitative data was used to gain confidence that the result is not a biased null effect. By this we mean that they are indicative of the cross-modal effect we are looking for, and not a null-result in which responses have been biased by sound for some other higher-level reason. Using a semi-structured interview instead of a questionnaire has given us a more detailed picture of how each participant combined audition and vision to make a final response. This enables us to select and screen out participants whose results are a biased null-effect with greater accuracy. Having said this, there is some discussion in the psychology literature concerning the validity of self-reports (Brener et al. 2003). A further problem with our qualitative data capture is that we can't relate

self-reported experiences to particular audio-visual separations. For instance, some of the participants who reported that they primarily used audio information to make their responses also said that they very occasionally did see visual depth changes. We are unable to assess whether these rare occasions were random, or whether they coincided with particular audio-visual separations. It is for these reasons that we have chosen not to use qualitative data in estimating the effect's psychometric function.

Careful thought was given to the wording of the question put to participants in each test. The aim was for the question to be as neutral as possible, in order to avoid directly biasing the participant's perception. "Which phone appears nearest?" is a very different question to that adopted in the preliminary trial: "Which image shows the phone to be nearest?" The latter may result in the participant purposefully blocking out the audio and focusing purely upon the image, reducing the strength of the cross-modal effect. On the other hand, if the effect can still be observed when participants are encouraged to ignore the audio, we can be more confident that a cross-modal perceptual effect is occurring. The danger of being neutral with questions is that you may appear vague; and so participants may not approach the the task in a consistent manner. For instance, the qualitative data revealed that some participants consciously resorted to audio information when they were unsure of a visual difference, whilst others consciously ignored audio, as they had assumed it was there simply to distract from the task at hand.

7.4 Conclusions

This study has extended the preliminary work reviewed in Chapter 4, using our results and equipment from Chapter 6. The MLE and Bayesian models of the ventriloquist effect suggests that whilst vision does bias our spatial perception of audio, the same is also true vice-versa. The amount of audio and visual bias depends upon various factors relating to the stimuli and environment within which they are perceived. In Chapter 4 we presented evidence that audio depth can bias perceived visual depth in S3D images. In this chapter we have sought to confirm this result, whilst gaining a greater insight into the effect's psychometric function and qualitative nature.

The experiment outlined in this chapter is an amalgamation of the experiments in Chapters 4 and 6. Participants were required to sit through a series of tests in which they consecutively viewed two pictures of a mobile telephone, each accompanied by a ringing sound. The audio depth changed in each pair, whilst the visual depth

remained constant. Moreover, the audio could either be played from the same depth as the visual stimulus, or from a given distance in front of the visual stimulus, which we refer to as the audio-visual separation. The two possible orderings of the audio depth (near then far, and far then near) give rise to two types of test which were randomly executed with equal frequency. The participants were then asked the question "Which of the two phones appears nearest?" If the significant majority of participants' responses agree with the audio component of the stimulus, then it was concluded that the audio component was influencing their perception of the stimulus' depth.

We investigated the effect of four different audio-visual separations upon the perception of the phone's depth: 0 cm, 25 cm, 50 cm and 75 cm. Each participant responded to 16 tests for each audio-visual separation. A mean and 95% confidence interval were calculated for each audio-visual separation using the fraction of responses that were consistent with the audio depth change. The psychometric function, which plots the dependency of task performance upon audio-visual separation, was estimated by fitting a logistic curve to these four measurements. This function is shown in Figure 7.4.

The psychometric function was used to determine the limits of the cross-modal effect – a primary conclusion of this experiment. The logistic model has asymptotic limits in task performance at 50% (chance performance) and 74%. By fitting further unconstrained logistic curves to the upper and lower limits of the 95% confidence intervals, we were able to interpolate a region of 95% confidence above and below the psychometric function. We consider the effect to be limited by the range of audio-visual separations for which our interpolated region of 95% confidence remains entirely within the performance limits of 50% - 74%. This corresponds to effect limits of 33.27 cm and 50.71 cm.

As far as the authors are aware, this is the first time participants have been screened for RAD perception acuity. This was deemed necessary because of the significant variation between participants that was found in our work outlined in Chapter 6. The screening test required participants to be blindfolded whilst responding to four tests for each of the non-zero audio-visual separations: 25 cm, 50 cm and 75 cm. The test took just a few minutes to execute and appeared to succeed in attaining a crude estimation of their acuity in RAD perception. Furthermore, we found evidence that participants' screening test scores were related to their scores in the experimental task.

The qualitative data was used to gain a deeper understanding of effect. Participants were divided into those who used vision primarily to give their responses, those

who used sound primarily and those who said they used both. When the participants who used sound primarily were removed from our group of screened participants, results for the 50 cm and 75 cm audio-visual separation remained significantly different from chance. This result gives us greater confidence that a cross-modal perceptual effect was occurring, instead of a null result being biased by sound. Participants were also grouped according to whether they felt the depth difference informing their responses varied. We found that the vast majority of participants (83%) did feel that the depth difference varied, suggesting that this effect is not simply "turned on" as if modelled by a step function. We also probed participants to estimate, in units of length, the maximum size of visual depth difference that informed their responses. Out of the 37 participants who contributed to the qualitative analysis, 10 felt unable to do this. The median estimate of the the maximum visual depth difference they perceived was 1 cm, with an interquartile range of 0.5 cm to 3.5 cm. Whilst we recognise the limitations of the methodology used here, this analysis does provide evidence that the auditory bias of visual depth in S3D images that we are exploring may be of a useful size.

This study opens up a number of avenues for further research. There is a wealth of psycho-physical experimentation to be undertaken concerning the auditory bias of visual depth in S3D images. Such experimentation should seek to gain further confidence that a cross-modal effect is being observed instead of a biased null effect. It should also seek to quantify the amount of bias that occurs and reconcile this with existing models of the ventriloquist effect, such as the MLE model and the Bayesian model discussed in Section 3.2.3. There are a variety of factors that may influence the effect that remain unexplored, including the nature of the auditory and visual stimulus, the viewing environment, the display technology and audio-visual scene complexity.

CHAPTER 8

Measuring the size of the cross-modal bias

The cross-modal effect observed in the preliminary study outlined in Chapter 4 has been confirmed in Chapter 7 using calibrated equipment from Chapter 6 and an extended experimental design. The results from these experiments show that the majority of participants perceive a ringing mobile phone (a cross-modal stimulus) to be nearer if the ring (the auditory component of the stimulus) is played from a nearer depth. These experiments have measured the impact of the cross-modal effect upon a forced choice depth comparison task, without actually measuring the size of the cross-modal bias induced by the effect. The size of the cross-modal bias is a crucial factor when considering the effect's value in application scenarios. In this final experimental chapter, we investigate the potential value of this effect in application scenarios through measuring the perceived cross-modal bias.

We use the term "cross-modal bias" to refer to the difference in perceived depth (measured in units of distance) of a *visual* stimulus, induced by a spatially disparate but seemingly congruent *auditory* stimulus. Various studies addressing the ventriloquist effect have proposed models for predicting this value (discussed in Section 3.2.3). The majority of these studies consider the bias of the simulus' auditory component towards its visual component. In this thesis we consider the value of reversing the ventriloquist effect – biasing the position of the visual component towards the auditory component – for use in application scenarios such as S3D cinema, gaming, simuluation and data visualisation.

This chapter is structured in the usual manner and begins by outlining the experimental method used in Section 8.1. In Section 8.2 we present the results of the experiment which are discussed in Section 8.3. We summarise our work and draw out its salient conclusions in Section 8.4.

8.1 Method

As in Chapter 6, this section begins by briefly discussing the preliminary trials in Section 8.1.1 that contributed towards formulating the final experimental design, presented in Section 8.1.2. In Section 8.1.3 we specify the experimental setup used, followed by detailing the participant samples in Section 8.1.4. We finish presenting our method in Section 8.1.5 by outlining how we captured qualitative data to support our analysis of the quantitative data.

8.1.1 Preliminary trials

One of the initial aims for this final experiment was to demonstrate a quantitative effect in a potential use-case scenario. We sought to design a game-like task centred around a cross-modal depth judgement, and show that a participant's performance could be significantly altered by changing the auditory depth. A game that modelled an arcade claw/crane machine fulfilled these requirements neatly. The task required the participant to position a claw directly above a cross-modal stimulus, so that when the claw dropped it could clamp onto the stimulus and lift it up. The task was simplified by limiting the claw's degrees of freedom, so that the claw could only be moved in or out of the screen. This meant that the only judgement used in the task was a relative depth judgement between the stimulus and the claw. We hoped that by changing the auditory depth associated with the cross-modal stimulus we could alter a participant's choice of claw depth and thus also damage their task performance (if we define success to be the correct selection of the cross-modal stimulus' visual depth).

The relative depth judgement in this task is different in nature to the task used in Chapters 4 and 7. Previously, we asked participants to compare the depth of two cross-modal stimuli shown consecutively, whereas in this task we asked them to compare the depth of a visual stimulus (the claw) with a cross-modal stimulus (the phone) shown concurrently. We therefore adjusted the task slightly. The participant was asked to pick up two phones, shown consecutively at notionally different depths. The initial depth of the claw was random, but was not reset after picking up the first phone, so that the difference in selected claw depths for each phone should be a measure of the perceived visual depth difference between the phones. In practice, just as in the previous cross-modal experiments, there was no visual depth difference between the two phones – just an auditory depth difference in the ring. The difference in the selected claw depths should therefore be a measure of

the cross-modal bias perceived by the participant.

The mobile phone and phone ring from the previous cross-modal studies were used again in this experiment. The experimental setup was designed to match the setup of the previous experiment as outlined in Section 7.1.2 and Figure 7.1 (page 119). The same four audio-visual separations from the previous experiment were used: 0cm, 25cm, 50cm and 75cm. For each audio visual separation the participant was asked to complete the task three times.

Before recruiting the participants to run the full experiment, we decided to "dry-run" the design on a small group of participants who had no prior knowledge of the research. Three female participants, sourced from the first-year undergraduate students at Durham University, contributed to this dry-run. Two of the participants were aged 18 and the other aged 19. They were all screened for stereo vision (the Titmus test), 20/20 vision (the Snellen eye test) and hearing (the BSHAA online hearing hest) prior to their participation in the dry-run.

The means of the nine measurements collected for each audio-visual separation are shown in Figure 8.1 with the corresponding 95% confidence intervals and p-values for a one-sample t-test against the null hypothesis that the mean bias equals zero. The results give no indication of a cross-modal effect occuring. Whilst we acknowledge that the the lack of evidence for an effect could be due to the small sample size, it still prompted us to question our design resulting in a number of further issues coming to light.

Participants generally ignored the fact that the claw depth didn't change between the first and second part of each test. When viewing the second phone, they would usually move the claw to an extermity and then re-select the depth as if it were a separate task. This suggests that they were not considering the relative depths of the cross-modal stimuli when completing the task; they were making absolute depth judgements instead. Therefore, the results of the previous studies don't suggest that we should expect to see an effect in the results collected using this experimental design. Because of this, we chose to alter the experimental task before collecting more data.

8.1.2 Final design

The task used in the final experimental design ensured that the depth judgement was a sequential comparison of two cross-modal stimuli, as used in Chapters 4 and 7. In each test the participant could switch between two different images of the same mobile phone used in our previous cross-modal studies and shown in Figure

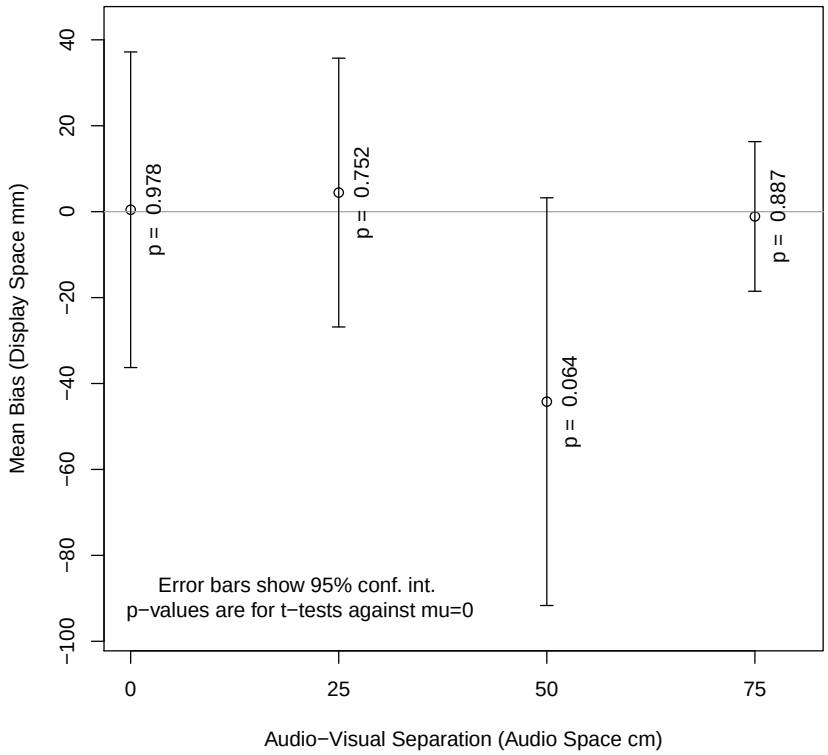

Figure 8.1: The results from the preliminary trials. The mean bias for each audio-visual separation are plotted with the corresponding 95% confidence interval calculated using a Student's t-test. The p-values indicate the proability that we can reject the null hypothesis that the median is equal to zero. All points prove insignificantly different from zero and no significant differences are found between audio-visual conditions.

4.1. They did this as many times as they wished, using the space bar. Each image of the phone was accompanied by the same telephone ring sound used in our previous cross-modal studies. One of the images was labelled "reference" in the top left-hand corner, whilst the other was labelled "selector" in the top right-hand corner. In each test the participant was asked to position the selector phone at the same depth as the reference phone.

The height and lateral position were fixed and the same for both phones, so that the only degree of freedom controlled by the participant was the depth of the phone. The phones were horizontally aligned in the centre of the screen at approximately eye height. The depth of the selector phone's visual and auditory components could then be altered by 2 mm (in the software's visual co-ordinate system) with each press of the up and down arrow keys, such that the audio-visual depth separation satisfied one of four different audio-visual conditions. The auditory component could

be positioned at distances of 25cm, 50cm or 75cm in front of the the visual component (as in our previous experiment), by using the motorised rails to move the appropriate loudspeaker. The fourth audio-visual condition satisfied the "null case" by keeping the auditory component fixed at the same depth as the reference stimulus (independant of the selector phone's visual depth). Our experiments use physical loudspeakers which cannot pass through the 3DTV screen so we were unable to explore an audio-visual separation of 0cm. The test ended when the participant pressed [c] to confirm their selected depth for the selector phone.

The conceptual idea behind this design is that the disparate auditory component will pull the perception of the selector phone's cross-modal depth forward, thus resulting the visual component of the selector phone being positioned behind the visual component of the reference phone. The depth difference between the two phones at the end of the test is taken as a measurement of the cross-modal bias. The four different audio-visual conditions were replicated three times for each participant, making a total of twelve tests for which the data was recorded. The order of these tests was randomised for each participant. A further four training tests, one for each audio-visual condition, were randomly ordered for each participant and undertaken at the begining of the experiment. The purpose of these training tests, was to ensure the participant was comfortable with the test procedure and to reduce any impact of a learning effect. The results of these training tests were therefore discarded.

The stereo depths of the phones were controlled using the algorithms outlined by Jones et al. (2001). In this experiment it is important to approximately match values of depth in display space, or in our experiment visual space, to those in physical space, or in our experiment auditory space. Due to pysiological differences between human visual systems, we cannot assume that all participants will perceive the same amount of visual depth when subject to the same amount of binocular disparity. We chose to perform a simple calibration task prior to each participant undertaking the experiment, in order to map between auditory space and the participant's visual space. We assumed that the mapping could be treated as linear:

$$D_{visual} = k.D_{audio}$$

Where D indicates a depth and k denotes the linear mapping constant. We measured k by asking the participants to position the mobile telephone directly above a marker placed 10cm (in audio space) in front of the screen. They did this three times and the mean selected depth was divided by the audio space depth of 10 cm to calculate k. This mapping was then used, along with the depth control algorithms, to calculate

Figure 8.2: The experimental setup was the same as outlined in Figure 7.1, except that the participant was given access to a keyboard to control the software. The keyboard was lit by a small torch when the lights were turned off.

the binocular disparity required to make the phone appear at the right depth in front of the screen.

8.1.3 Experimental setup

We used the same experimental setup in this study as in our previous experiment. A diagram detailing this setup can be found in Figure 7.1 (page 119). The only notable difference was that the participant had access to a keyboard to control the software. A photograph of the arrangement is shown in Figure 8.2, though the room was dark whilst the participant undertook the calibration and experimental tasks. As in the previous experiment, a thin black cloth was used to conceal the loudspeaker arrangement from the view.

An LG BM-LDS302 47 inch 3DTV screen was used to display the visual stimulus, whilst two K-Array KT20 loudspeakers, driven by a Cambridge Audio Topaz AM1 amplifier, were used to play the auditory stimulus. The loudspeakers had been identified by the manufacturer as having well matched frequency response curves. The experiment was performed in the same semi-reverberant laboratory with a background noise of approximately 41.5 ± 0.3 dB. The volume of the auditory stimulus was set to be a maximum of 70.0 dB at the approximate listening position, with the loudspeakers positioned on the rail at the furthest possible point from the par-

ticipant. All equipment was placed upon a desk that extended from the screen to the participant. The centre of each loudspeaker's front face stood 14 cm above the desk, whilst the chin rest stood 37 cm above the desk.

The display's crosstalk at the center of the screen was measured using a Sekonic L-758 Cine light meter fixed to a tripod in the approximate position of the viewers' eyes. Application of the method outlined by Liou et al. (2009) concludes that the display's crosstalk is no more than $0.95\% \pm 0.01\%$ in the left eye and $0.50\% \pm 0.01\%$ in the right eye. We say these are upper bounds because our equipment was not sensitive enough to measure the non-zero black-black level, meaning it was below $0.25cd/m^2$. According to Liou *et. al* this black-black level should be subtracted off the black-white leakage measurements to obtain a more accurate measurement of the display's crosstalk. The lower bound for the crosstalk could therefore be calculated as $0.46\% \pm 0.01\%$ in the left eye and $0.09\% \pm 0.01\%$ in the right eye.

The motorised rails, loudspeakers and display were all controlled by a computer program written in the Java programming language, using the JOGL to render the graphics. The visual stimulus was rendered using the wavefront file and texture map from Chapter 4. The program procedure was controlled by the participant who acknowledged they were ready to begin each test by pressing space bar. For each participant, the program generated a text file which stored their results and test details.

8.1.4 Participants

The participants were sourced from the undergraduate student group at Durham University and did not include the authors. Thirty-five participants took part in the study, each contributing three measurements for each audio-visual separation, making 105 measurements for each audio-visual separation in total. The participants' ages ranged from 18-24 with an interquartile range of 18-21 and a median of 19. The sample was made up of 60% male and 40% female participants. All participants were screened for stereo acuity, 20/20 vision and hearing using the Titmus stereo test, a Snellen eye chart and the BSHAA online hearing test. The auditory depth perception screening test that we designed and outlined in Section 6.3.3 was used after the main body of the experiment in order to avoid it causing the participant to hypothesis guess.

8.1.5 Qualitative data capture

Each participant was required to fill out a post-experiment questionnaire. The purpose of this questionnaire was to find qualitative evidence supporting the quantitative data and identify threats to the validity of the participant's results. Using a questionnaire ensures that each participant's data is captured in a consistent and repeatable manner. The questionnaire recorded responses to the following questions:

1. Did you understand the task required of you? (Yes/No)
2. Do you feel that your answers were a correct representation of what you perceived? (Yes/No) Please add any comments that explain your answer.
3. Please give any thought you have on how you aligned the two phones
4. Do you have any other significant comments that may be worth recording, regarding the execution of the test?

The participant was given a box in which to enter their answer for questions 2, 3 and 4 as prose. Participants were given as much time as they required to fill out the form to their desired level of detail.

8.2 Results

As in the previous chapter, we asked participants to take the screening test for RAD perception that we designed using our work in Chapter 6. We provide analysis of the screening test's value to this study that is analgous with the analysis provided in Section 7.2.1. Figure 8.3 shows the relationship between scores in the screening test (number of correct responses out of four) and performance in the experiment. As we discuss in Section 3.2.3, audio-visual bias depends upon how clearly the brain can distinguish the audio and visual cues. Therefore, we might expect that better performance in the screening test would indicate larger bias in the experiment. Each graph in Figure 8.3 includes a least squares linear fit to the data. Only the linear fit corresponding to the 25 cm depth difference matches expectation by yielding a positive gradient. This appears to contradict our expectations. Finite sample f-tests against the null hypothesis that the gradient equals zero were used to determine whether the gradients were significantly positive or negative. The gradients, standard errors and p-values for these tests are shown in table 8.1, along with the value of the Pearson's product-moment correlation coefficient. Non of the cases prove statistically significant. We therefore chose not to use the screening test data in the analysis of our results.

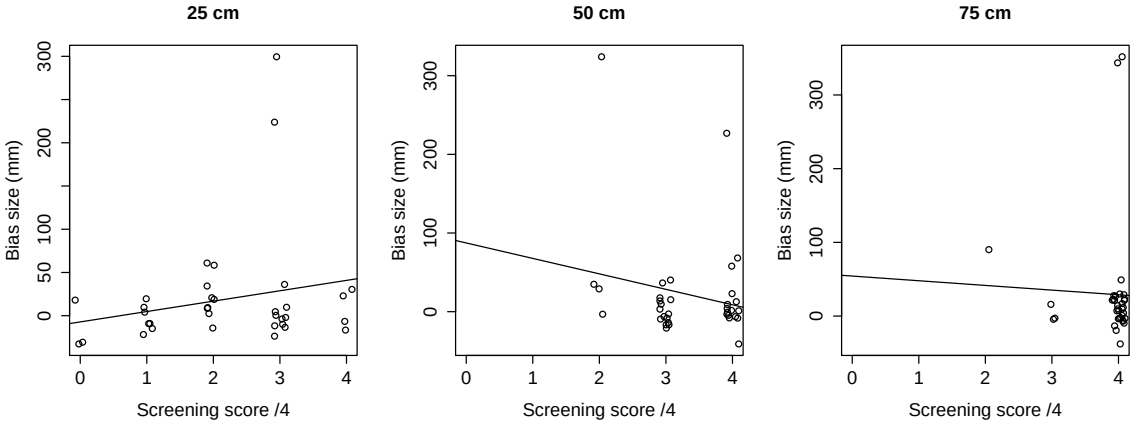

Figure 8.3: The relationship between the participants' performance in the experiment and their scores in the screening test, for each of the three non-zero depth differences. As the screening test data is heavily discretised, we decided to "jitter" the data points in the x-dimension so as to make the data distribution visibly clearer. In other words, for display purposes only, a small amount of random noise has been added to the x-values so as to minimise the number of overlapping data points.

Depth /cm	Gradient	Std.Err.	p-Value	Pearson's
25	6.202	4.704	0.1964	0.2237
50	-11.41	8.89	0.2083	-0.2181
75	-2.2	16.82	0.8967	-0.02276

Table 8.1: The gradients for the linear regression in Figure 8.3 with their standard errors and p-values for finite sample f-tests against the null hypothesis that the gradient equals zero. The Pearson product-moment correlation coefficient is also shown.

We have already explained in Section 8.1.2 why display space and audio space may not be related by a one-to-one mapping. Prior to reporting the results here, we have converted all the data, which was measured by the computer in display-space units of depth, back to audio-space units of depth. This was done by dividing each participant's data by their mapping constant calculated during the calibration phase of the experimental procedure.

Before starting to analyse the results, we also discarded one participant's data. This was due to their responses in the questionnaire; specifically their response that said they did not feel their answers formed a correct interpretation of what they perceived. This participant's comments are discussed further in Section 8.3.

Shapiro-Wilk tests fail to conclude that any of our data distributions are normally distributed. This conclusion is supported by plots of the frequency histograms and calculations of the values for kurtosis and skewness. Our data is leptokurtic with a

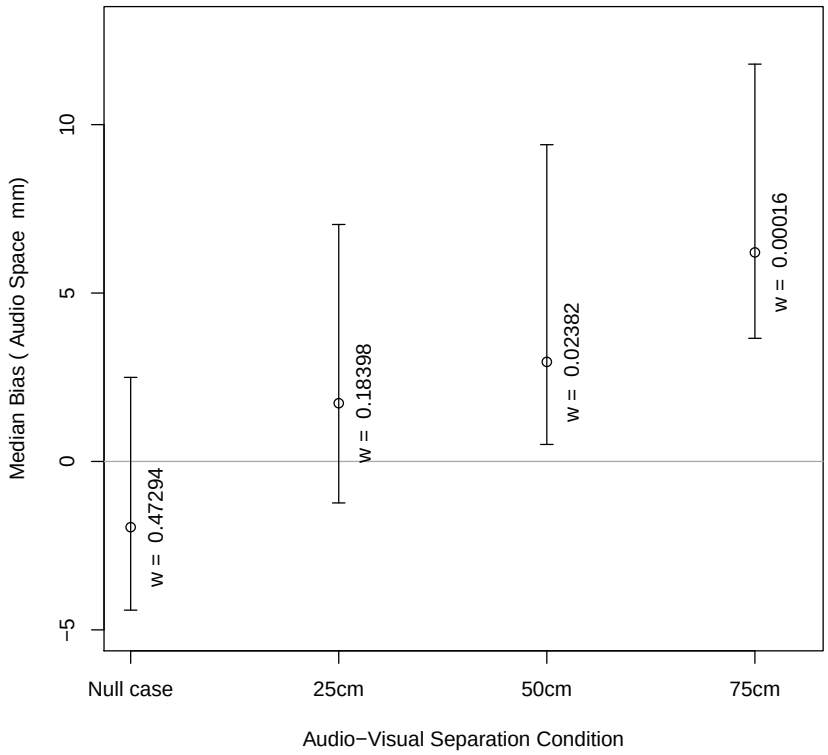

Figure 8.4: The median cross-modal bias for each audio-visual separation, with the corresponding 95% confidence intervals calculated using Wilcoxon signed rank tests. The p-values indicate probability that we can reject the null hypothesis that the median is equal to zero.

positive skew. The skewness and kurtosis values which ideally should lie between -1 and 1, ranged between 2.36 and 4.07 for skewness, and 8.97 and 19.21 for kurtosis. In such situations it may be possible to transform the data and attain better values for the skewness and kurtosis. We found that a recipricol transformation yielded the best possible results and almost eradicated any skew, however the data remained quite leptokurtic unless we discarded 18 data points as outliers. As we discuss in Section 8.3, we would not necessarily expect our results to be normally distributed. We have therefore decided to apply a non-parametric analysis to our data.

Figure 8.4 plots the median bias measured for each audio-visual condition in the direction of the auditory component's depth. Median bias sizes in audio-space for the null case and the 25 cm, 50 cm and 75 cm audio-visual separations were measured to be -1.95mm, 1.73mm, 2.96mm and 6.21mm respectively. Each value is accompanied by a 95% confidence interval calculated using the non-parametric Wilcoxon Signed Rank test. P-values for these tests against the null hypothesis

that the median bias is equal to zero are also shown on the graph. These tests prove significant for both the 50cm and 75cm audio-visual separations.

The Friedman test is the non-parametric equivalent of a repeated measures ANOVA. It is implemented for replicated blocked data in the R-Project package *muStat* (Wittkowski and Song 2012). Grouping the data by audio-visual condition and blocking the data by participant yields a strongly significant p-value of 0.0103. We therefore conclude that there are significant Bias differences between each audio-visual condition.

Two sample Wilcoxon signed rank tests allow us to identify where these significant differences occur. The difference between the null case and the 50cm audio-visual separation proves strongly significant with a p-value of 0.0480. The difference between the null case and the 75cm audio-visual separation also proves strongly significant with a p-value of 0.00185. The difference between 25cm audio-visual separation and the 75cm audio-visual separation proves weakly significant with a p-value of 0.0840. All the other differences prove insignificant.

8.3 Discussion

The failure of the screening test to predict anything about the participants' performance in the experiment came as a suprise to us, particularly when it proved valuable in our previous experiment. It is important to note that the 2AFC paradigm, used in the previous cross-modal experiments and the screening test, collects binomial data - a response could be either correct or incorrect - whereas in this experiment we collected continuous data. The task is fundamentally different, which may offer some indication of why the screening test failed to be valuable in this experiment.

There are various reasons that might explain why our data is non-normally distributed. Firstly, without applying an effective screening test for audio acuity in the task, we cannot assume that audio performance is normally distributed. We would expect this to effect the distribution of cross-modal performance. Furthermore, we do not know that the human method of combining of audio and visual depth cues is normally distributed across participants.

The data collected in the post-experiment questionnaire did reveal that the majority of participants self reported as using visual cues to complete the task. Specifically, visual size was the most popular cue with 54% of the participants giving comments in the questionnaire suggesting its use. Such comments varied from *"Used mainly the idea of size in order to gauge consequent distance,"* to, *"Sometimes I tried to match the size of the icons on the mobile phone to help match their depth."*

One participant commented saying, *"Some of the harder/closer ones I compared by distance from bottom of the screen."* Thirty-seven percent of particpants classified as using size cues to complete the task, made such references to using the "height" of the phone to judge depth. The distance between the bottom edge of the phone and the bottom of the screen decreases as the phone moves towards the participant and increases in size. Another 7% of participants who used size to complete the task did so by focusing on the position of particular points and edges in the image. One participant said *"I used points of reference as well as comparing sizes of parts of the phones to determine when they were an equal distance away."* Whilst another participant said, *"By looking at the placement of the top and bottom lines of the phone"*. None of the participants made direct reference to the binocular depth cue, although a couple of participants did make comments about a generic depth cue that likely referred to it, such as *"Often [judged] by size, but was able to consider "depth" most of the time,"* and *"Judging the sizes of the phone, the difference from base to floor, and by how "close" they appeared."*.

Only one participant responded to question 2 with a *"No"* (Do you feel that your answers were a correct representation of what you perceived?). They explained their answer by saying *"Wasn't sure on the 2nd task whether I was moving the mobile phone to look the same on the screen as the reference or whether it was getting the phone to sound the same, because the reference phone always sounded quieter."* From this we concluded that the participant may not have been responding to the task consistently. This participant's responses were removed from the data before presenting the results in Section 8.2.

The post experiment questionnaires reveal that 14% of participants consciously used sound in matching the depths of the two phones. Participants commented that, *"I was trying to match the depth visually as well as how far away they both sounded but sometimes I thought the pitch was different for the phones,"* and, *"The ringing sound impacted how close I thought each one was, and that was mainly how I aligned them."* When these participants are excluded from the data, we still get a weakly significant result from the Friedman test with a p-value of 0.0874. Further analysis with Wilcoxon signed rank tests reveal that just the difference between the null case and the 75 cm audio-visual separation proves significant with a p-value of 0.0150. Figure 8.5 shows the replotted medians for the data, excluding those participants reported using sound to complete the experimental task. The newly calculated medians for the null case, 25cm, 50cm and 75cm audio-visual separations were found to be -1.95mm, 0.96mm, 1.81mm and 4.67mm respectively.

The effect sizes observed in this study are satisfyingly similar to the self reports

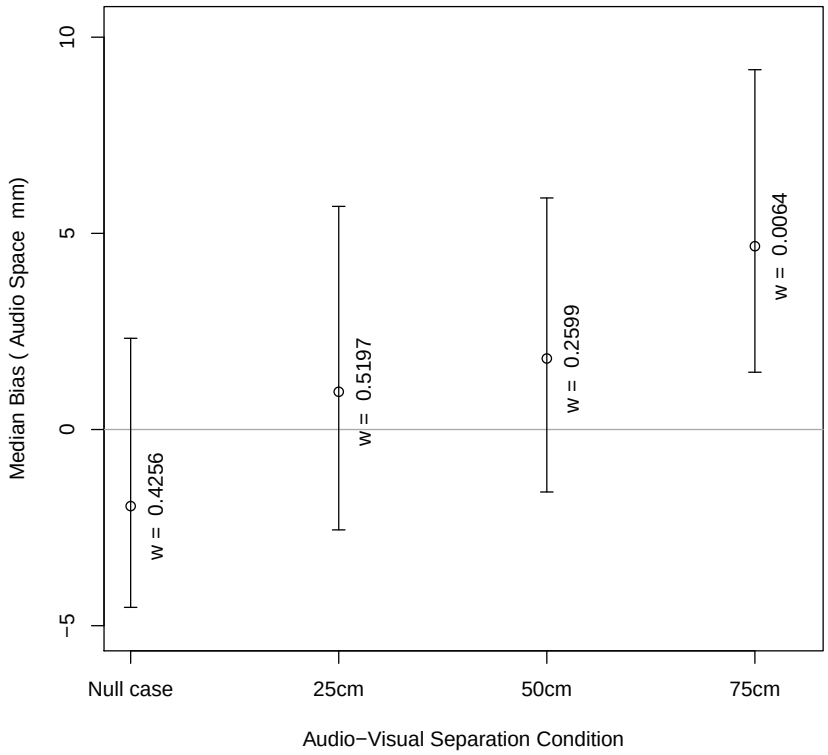

Figure 8.5: The median biases for the participants who didn't report using sound in the questionaire. The corresponding 95% confidence intervals, calculated using Wilcoxon signed rank tests, are shown with p-values specifiying the probability with which we can reject the null hypothesis that the median is equal to zero.

of perceived depth difference that are discussed in Section 7.2.4. In the previous experiment, the median maximum visual depth difference participants believed they had percieved was 1 cm. This value is similar to 0.6 cm and sits within the interquartile range that we measured for the 75cm audio-visual depth seperation, during this Chapter's experiment. This match between subjective and quantitative data gives us further confidence that our results are indicative of a perceptual effect.

The bias sizes we have measured appear rather small, so it is important to analyse their contextual significance. It is first important to note that the result for the 25cm, 50cm and 75 cm audio-visual separations correspond to being wrong by 2, 3 and 6 down-arrow key presses respectively. Figure 8.6 shows a simple S3D viewing arrangement of a single point in front of the screen with a disparity of g between left and right eye images. The viewer has an eye separation of e and is positioned at a viewing distance of z from the screen. If we assume that the perceived depth d occurs at the intersection of left and right eye rays, we can use the definition of the

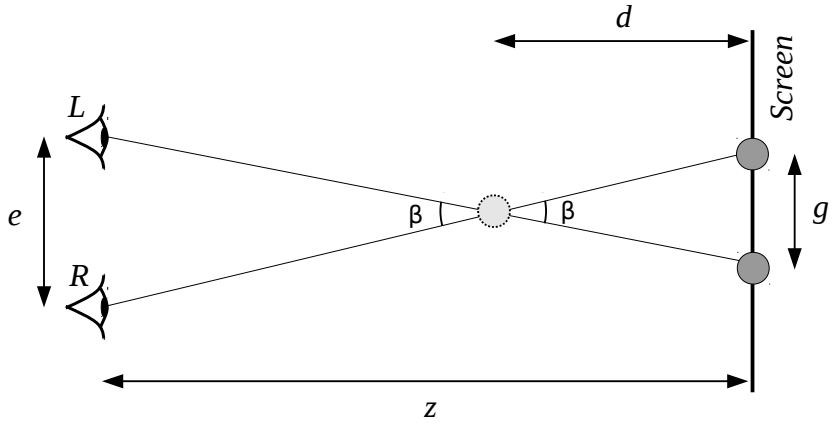

Figure 8.6: A simple S3D veiwing arangement of a point in front of the screen with corresponding screen disparity g. The viewer, with an eye spearation e, views the screen from a distance z and should perceive depth of approximately d. Using simple geometry we can derive equation 8.1 that predicts the perceived depth from a given disparity.

Sine function to write:

$$Sin\frac{\beta}{2} = \frac{g/2}{d} = \frac{e/2}{z-d}$$

Rearranging this gives us an equation that allows us to theoretically predict perceived depth from the screen disparity in our experimental setup:

$$d = \frac{z}{\frac{e}{g}+1} \tag{8.1}$$

For our calculations we use 0.06 m as a nominal approximation of the eye-separation (Dodgson 2004) and the viewing distance of 1.91 m that we used in our experiments.

A pixel in our screen is 0.000542 m wide. Using the above equation we can calculate the perceived depth corresponding to a single pixel's disparity in our TV viewing scenario to be 0.0171 m, or 1.71 cm. This is larger than the 0.621 cm effect size we have measured for an audio-visual separation of 75 cm. So the audio-visual effect bias is smaller than a visual depth difference coresponding to a single pixel of disparity in a 3DTV viewing scenario. In fact, using the above equation we can calculate that 0.621 cm of perceived depth corresponds to just 0.0196 cm of disparity, or 36% of a single pixel's disparity. This disparity would subtend an angle of 21 arcseconds. Howard (1919) found that some participants could distinguish disparity differences as small as 1.8 arcseconds, a value that is supported by several other studies (Langlands 1926; Yeh and Silverstein 1990; Julesz et al. 2006). Hence, we

expect that many of our participants would be able to perceive the small effect size, particularly as they had all been screened for stereo acuity using the Titmus stereo test (meaning they could perceive at least 40 arcseconds of disparity). These results support those from Chapters 4 and 7, which show that the audio-visual effect creates a depth difference that influences participants' relative depth judgments.

The small bias size does limit the practical application of this effect, but there may be scenarios when this effect would be useful. Particularly in situations where the full ranges of other depth cues have been exhausted and a little more is needed to distinguish between two points. Furthermore, due to limited time and resources we have been unable to draw conclusions concerning the external validity of the effect, so we know relatively little about the possible use-case scenarios. For instance, the effect may, under certain conditions, create a measurable subjective impact in a similar manner to our results in Chapter 5. It's also important to explore the impact of this effect in different experimental environments, such as desktop viewing, tablet viewing and cinema viewing. In light of the small bias size, we might expect this effect to have greater value in smaller screen arrangements. As mentioned in Section 4.1.2, the stimulus was chosen carefully as one which we might expect to yield an audio-visual effect. However, we don't know whether other stimuli would respond better or not. We also don't know whether motion would be more suseptible to the effect than static images. There is much further work to be completed.

The construct validity of this study is threatened by the lack of control for fidelity in RAD perception. We know from our previous work in Chapters 4 and 7 that there is considerable variation between participants in RAD perception acuity. Our results in this chapter include the results of participants who may have very poor ability to match the depths of the reference and selector stimuli in an auditory-only case. We would expect these participants to experience a smaller bias, or even no bias. Therefore, our results could be an under-estimate of the bias size experienced by those we would expect to perceive the effect.

Designing a S3D image that is perceived by participants to have depth cues that match its real world counterpart is as yet non-trivial. This poses a threat to the construct validity of our results, as the majority of literature addressing the ventriloquist effect uses real world data, in which one can know exactly where the visual stimulus is positioned. Various studies have found that perceived depth does not accurately match up with the widely accepted geometric models for depth perception in S3D images (e.g. Renner et al. 2015; Tai et al. 2013). In this study, we use a calibration task to improve the mapping of perceived depth in the display to perceived depth in the real world. Whilst this is a helpful "first order" improvement,

in practice we cannot be sure that the resulting camera projection used to create the visual stimuli was related to the real world by a true one-to-one mapping.

8.4 Conclusions

This chapter reports an experiment seeking to measure the size of the audio-visual bias whose impact we have observed in Chapters 4 and 7. The experimental task required participants to match the depth of a cross-modal stimulus with spatially disparate audio and visual components, called the "selector", to a static replica of the same stimulus but with spatially congruent audio and visual components, called the "reference". The stimulus was the same ringing mobile telephone used in our previous experiments. The reference and selector phone were viewed sequentially with a 0.5 second gap between each viewing, and for each matching task the participant could switch as many times as they liked between the two phones using the space key. The participant controlled the depth of the selector phone using the up and down arrows. Three different audio-visual separations were investigated: 25cm, 50cm and 75cm. As the participant altered the selector phone's depth, the audio and visual components moved such that the audio-visual separation remained fixed throughout the matching task. We also investigated the null case in which the audio depth was fixed at the depth of the reference phone and did not vary as the participant varied the selector phone's depth.

We expected that the nearer audio stimulus would "pull" the perception of the phone's visual depth forwards, so that when matching the two stimuli, the participant would believe that they were matched whilst the selector phone's visual component was still positioned behind the reference phone's visual component. The depth difference between the two phones is therefore a measurement of the audio-visual bias (albeit a noisy one). Thirty-five participants, screened for vision, stereo vision and hearing, contributed three measurements for each audio-visual condition.

The results for each audio-visual condition are plotted in Figure 8.4. The data was found to be non-normal, so a non-parametric analysis was applied to the data. A Friedman test looking for differences between the audio-visual conditions, gives a strongly significant result. Further analysis with two-sample Wilcoxon signed rank tests reveal that significant differences exist between the null case and the 50cm and 75cm audio-visual separations, and that a weakly significant differences exists between the 25 cm and 75 cm audio-visual separations. Median biases of -1.95mm, 1.73mm, 2.96mm and 6.21mm were measured for the null case and the 25cm, 50cm and 75cm audio visual separations respectively. One sample Wilcoxon signed rank

tests allow us to conclude that the 50cm and 75cm audio-visual separations gave results that were significantly different from a bias of 0 cm.

Six millimeters of depth is smaller than the depth corresponding to one pixel's disparity in our TV viewing scenario. The effect is therefore very small. However, it is significantly larger than the most accurate levels of stereo acuity reported in literature (Howard 1919) and corresponds to six down-arrow key presses in our software. This may explain the strong results in Chapter 7 and allows us to conclude that that there could be uses for this small effect. Further work should seek to explore possible use-cases as well as offer a better understanding of the scope of this effect. In particular they should establish whether other experimental conditions might lead to a larger bias sizes relative to the display's depth budget.

CHAPTER 9

Conclusions

This chapter draws together all the work presented in this thesis, and seeks to to answer the over-arching research question raised in Chapter 1: *Is it important to quality-control audio-visual depth by considering audio-visual interactions in depth perception when designing content for integrated S3D display and spatial sound systems?* Our work began with a preliminary experiment exploring the hypothesis that audio depth could affect participants' judgements of visual depth in a S3D image. Building upon the positive results of this experiment, we assembled a number of key research questions that would lead us towards answering our over-arching research question. Each research question was directly addressed by at least one chapter in this thesis. In total, four distinct experimental studies have been undertaken, each making novel contributions to the field.

This chapter begins in Section 9.1 by summarising the work undertaken and the conclusions made that enable us to answer the research questions raised in Chapter 1. We then offer an answer, in Section 9.2 to the over-arching research question. We draw attention to the novel contributions of our work in Section 9.3 before finally discussing the questions posed by our work that further research could address.

9.1 Research questions

Here, we provide a summary of the work undertaken in order to answer each research question raised in Chapter 1.

1. What is there to be learnt from the literature?

In Chapter 2 we review the literature detailing the cues to visual and auditory depth perception as well as the engineering of S3D displays and spatial sound systems.

Two dimensional displays present depth in a scene using a variety of pictorial cues, such as perspective, size and occlusion. Despite offering binocular disparity as an additional visual depth cue, depth perception in S3D displays cannot be considered a "natural" viewing experience, due to factors such as the vergence-accommodation conflict. Depth perception in a S3D image is therefore degraded, when compared with depth perception in a natural environment, although markedly enhanced when compared with depth perception in a conventional 2D image. This results in S3D displays having a limited range of depth which is comfortable to view, called the "depth budget" or "zone of comfort". This range is very limited for small screens with small viewing distances.

The primary cues to auditory depth perception are loudness, reverberation, frequency spectrum and inter-aural differences. Spatial sound systems utilise these cues (and others) to present a source's position in 3D space. Developments in the engineering of spatial sound systems have turned attention to S3D media as a showcase for the technology.

Chapter 3 reports on studies concerning the interaction between audio and visual perception. There are many examples of visual and auditory perception influencing each other. One of the most famous examples of an audio-visual interaction, or cross-modal effect, is the ventriloquist effect. Models of the ventriloquist effect suggests that, whilst most commonly vision affects auditory spatial perception, audition can conversely affect visual spatial perception under certain conditions. Indeed, this has been empirically confirmed in some studies. The conditions under which this occurs include degraded visual stimuli, bi-stable visual stimuli, or an impaired human visual system.

Our preliminary experiment, described in Chapter 4, suggests that auditory depth cues can influence perception of depth in S3D displays. The experiment showed that a 2AFC relative depth judgement between two cross-modal stimuli could be influenced significantly by varying the depth of the auditory component, if the visual components are positioned at the same depth. So in answer to our first research question, the literature presents ideas motivation for our work, is well-aligned with the results of our preliminary experiment, and suggests where we should look for a stronger cross-modal effect.

2. Does viewing S3D content with quality-controlled binocular cues create measurable positive changes in the audience's subjective attitudes towards S3D media?

Our first experimental study, reported in Chapter 5, sought subjective evidence of an enhanced viewing experience offered by high-quality S3D media, in which the binocular cue was quality-controlled using the algorithms outlined by Jones et al. (2001). Subjective evidence was collected in the form of responses to 5 questions. Each response was given on a five point Likert scale, shown in Figure 5.1, that was subdivided into 100 values and printed such that it satisfied the specifications outlined in ITU-R Recommendation BT.500-12 (2009). With regard to viewing S3D media, our five questions investigated the concepts of: viewing experience, suitability for displaying complex information, viewing comfort, naturalness and knowledge retention. We implemented a one group pre-test post-test quasi-experimental design, in which these questions were asked before and after an intervention – in our case the viewing of a short, high-quality S3D film. Any significant change in responses to the questions are considered to arise as the consequence of the intervention. Our first experiment was undertaken using the technology we most expected to yield significant positive effects: our low cross-talk, large screen, active shutter-glasses display. We then performed a series of replications in which we varied the viewing technology, content and location of the experiment.

We concluded that our high quality S3D films, Cosmic Origins and Cosmic Cookery, create measurable, repeatable changes in audience attitude towards the medium. These changes remain significant when varying the content, display technology and site location used in the experiment. Both large and small screens were tested, as well as national (UK) and international sites. The current popular attitude towards S3D content may be improved by the wider distribution of high-quality content, created using quality-controlled depth cues. Our positive answer to this research question supports the importance of quality-controlled depth cues and the benefit of restricting binocular cues to a depth budget, thereby providing motivation for our further research questions.

3. What is the MAD in our experimental setup?

In Chapter 6 we measure the MAD between our two loudspeakers using a two-down one-up transformed adaptive experiment with parameter estimation by sequential testing. This sensory threshold informs the design of our later cross-modal ex-

periments and forms an important value that spatial sound content designers and systems engineers should consider in their work. In the experiment, participants answered a series of 2AFC tests in which they were required to choose which of two sounds appeared nearer to them. The depth difference between the speakers in each test depended on whether the participant answered the previous tests correctly or not. The experimental design specifies rules that result in the depth differences converging on a measure of the participant's MAD.

From the results we concluded that there is significant variation of the MAD between participants, suggesting that, where appropriate, researchers may wish to screen participants for RAD perception prior to including their result in analysis. We propose a model for this screening test, implementing and evaluating this in Sections 7.1.4, 7.2.1 and 8.2. The PDH, which theoretically predicts the MAD to be 5% of the reference distance, was found to approximate the performance of the participants with the best acuity. The worst performing participants were found to have acuity values ranging from 20% to 35% of the reference distance. Our results match the reciprocal trend between the MAD and the reference distance that was found to exist in previous studies. This was then used to build a model for the upper-quartiles of our data sets (M) from the reference distance (D):

$$M = \frac{10.7}{D} + 14.6$$

For content designers, different limits are important in different design contexts. If auditory depth is used as a medium to deliver important information, a larger value for the MAD should be considered to ensure the majority of the audience can perceive that delivery. In such a case, we suggest using the upper-quartile value as estimated using our model above. There are many scenarios when using auditory depth to create specific sensation or effect in film might require this, such as gun shots approaching behind the camera, or footsteps approaching the camera in the dark. This might also be important when using audio depth to improve or distract performance in a S3D gaming environment, or even when using audio to improve comprehension of scientific data visualisation. On the other hand, if the purpose of auditory depth is to provide the scene with a degree of fidelity that is valuable for listeners with the best levels of acuity, then one should aim for the smallest MAD values of approximately 5%. Doing so will provide audience members who have the best acuity with a level of auditory spatial detail they can appreciate.

The data and analysis presented in Chapter 6 therefore provides an answer to this particular research question. For a reference distance of 1.81 m, the median

MAD in our TV viewing scenario is approximately 25 cm.

4. Does auditory depth influence our perception of relative depth in S3D images?

There were a number of weaknesses in the execution of the preliminary experiment that meant we could only use its results as a motivation for answering our overarching research question, rather than as part of the answer. We therefore designed a new experiment, based upon the design of the preliminary experiment, that would give a more robust answer to this particular research question. This experiment is reported in Chapter 7, and answers both this research question and the succeeding research question: *How does the cross-modal effect vary with auditory depth?*

Participants were asked to make a relative depth judgement between two cross-modal stimuli whose visual depths were the same but whose auditory depths varied. The stimuli, representing a mobile phone, were presented consecutively, and the participant was asked: "Which phone appears nearest?" The possible responses were "the first" or "the second". This study differed from the preliminary experiment in that it used a calibrated experimental setup and a more rigorous experimental design. The new experimental design involved screening participants for acuity in RAD perception, and used an improved method for collecting qualitative information from a participant. In order to simultaneously answer the succeeding research question, multiple audio-visual separations were investigated for a viewing distance of 1.91m corresponding to a 30° viewing angle. This is widely quoted as the SMPTE recommended viewing distance (Rushing 2004).

The results showed a significant effect for auditory depth differences of 50 cm and 75 cm. That is, in a significant majority of cases, participants believed that the nearer stimulus was the stimulus accompanied by the nearer sound. Such an effect was still observed after removing all the participants who reported that they used sound to determine their response. In the preliminary study the audio-visual separation was set at 20% of the reference distance, for which 65% of responses said that the visual stimulus accompanied by the nearer auditory stimulus appeared nearer. By interpolation, we can use the results from this new study to conclude that the corresponding result for an audio-visual separation of 20% of the reference distance used would be 57% of responses. We wouldn't necessarily expect such a similar result, given that the experimental setup has been altered substantially to give us greater confidence in our result e.g. we have used a different reference/viewing distance. As in the preliminary experiment, a significant cross-modal effect has been

observed in our results with a slightly smaller effect size. So in answer to our research question, auditory depth therefore can influence our perception of relative depth in S3D images.

5. How does the cross-modal effect vary with auditory depth?

The experiment reported in Chapter 7, which answered the previous research question, was also designed to investigate how the cross-modal effect varies with auditory depth (or audio-visual depth separation). As mentioned in the previous section, four different audio-visual separations, separated by the median MAD as measured in the previous experiment, were incorporated into the experimental design: 0 cm, 25 cm, 50 cm and 75 cm.

For the 0cm, 25cm, 50cm and 75cm audio-visual separation we found that 50%, 52%, 66% and 73% of particicipants believed that stimulus with the nearer auditory component was the nearer stimulus. In the null case, where the audio-visual separation was 0 cm, there was no nearer auditory component, so we expected and measured a 50% chance divide in responses. For the 50 cm and 75 cm audio-visual separations, the response divide was significantly different from chance. The effect's dependancy on the audio-visual separation can be modelled neatly using a logistic curve, from which we can draw estimates for the effect's limits. The effect appears to only vary significantly for audio-visual separations between 33.27 cm and 51.71 cm.

6. How large is the cross-modal bias?

When answering the previous research question, we measured the impact of a cross-modal bias upon performance in a relative depth judgement task. We did not measure the size of the cross-modal bias, which is important when considering potential applications of the effect. In Chapter 8 we measure the size of the bias using a variation of the experimental task used in Chapters 4 and 7. In the new task, participants could switch as many times as they liked between two stimuli - one labelled the "selector" and the other labelled "the reference". The audio and visual depths of the reference stimulus were matched and fixed at 10 cm in front of the screen. The audio and visual depths of the selector stimulus were separated by a fixed value, and could be controlled in a synchronous manner by the participant (such that the audio-visual depth separation remained constant). The participant was asked to match the perceived depth of the selector stimulus (with disparate auditory and visual components) to the fixed depth of the reference stimulus. The

difference between the participant's selected depth and the reference stimulus' depth was taken as a measure of the cross-modal bias created by the audio-visual depth separation in the selector stimulus.

Four different audio-visual separation conditions were explored. The audio and visual depths of the selector stimulus could be separated by one of three fixed distances: 25 cm, 50 cm and 75 cm, or in the null case, the audio depth was tied to the same depth as the reference phone. Significant differences in the bias size were found to exist between the null case and the 50 cm and 75 cm audio visual separations. Another significant difference was found to exist between the 25 cm and 75 cm audio visual separation. The maximum bias size observed was 6.21 mm, for an audio visual separation of 75 cm from a viewing distance of 1.91 m. This value is small, but bigger than the minimum visible stereo depth difference, which we consider to be the reason we observed the impact of the effect in Chapters 4 and 7. We have shown that when the full range of binocular depth is used, sound has the potential to offer a small but noticable amount of further perceived depth.

9.2 The over-arching research question

The over-arching research question that directed the work in this thesis is presented in section 1.2 as: *Is it important to quality-control audio-visual depth by considering audio-visual interactions in depth perception when designing content for integrated S3D display and spatial sound systems?* We have identified three different parts to this research question. Firstly, the importance of quality-controlling audio-visual depth; secondly, the consideration of audio-visual interactions; and finally, the integration of S3D displays and spatial sound systems.

The approach we have taken to answer this research question begins with an experimental study that shows the importance of quality-controlling the binocular depth cue alone, by restricting it to a particular depth budget. This study serves as a motivation for our later work, which explores the possibility of extending the depth budget using audio depth. The experiments we then report each address a part of the over-arching research question, in reverse order. We build a calibrated audio and visual display system and use it to observe the impact of an audio-visual interaction upon a relative depth perception task, before finally measuring the size of the audio-visual bias in order to discuss the effect's potential application and significance.

The first part of the research question – addressing the importance of quality-controlling audio visual depth – is addressed by both our first and last experiments.

We began by collecting subjective evidence suggesting it is valuable to quality-control the binocular cue. By extension we might expect that it is important to quality-control other depth cues. Furthermore, we have also shown that there is qualitative value in restricting binocular depth to a given depth budget for a display, which our further work then seeks to extend using auditory depth. This experiment therefore suggests, in multiple ways, that it could be important to quality-control audio-visual depth. This is confirmed in the final experiment, where we show a cross-modal effect that could extend the binocular depth budget in certain application scenarios. However, the small size of the effect seems likely to restrict the number of possible application scenarios.

The final two experiments we report directly address the second part of our over-arching research question – the consideration of audio-visual interactions. We have shown that audio depth can influence viewers' performance in a relative depth perception task. Audio depth differences, larger than approximately 33 cm (for a viewing distance of 1.91m), can cause a significant majority of viewers to believe they saw a depth difference between stimuli. For an audio depth difference of 75 cm, this visual depth bias is approximately 6 mm, which is larger than the minimum visible binocular depth difference that has been measured. The popularity of the algorithms used to design Cosmic Origins and Cosmic Cookery (Jones et al. 2001; Holliman 2004; Holliman et al. 2006; Holliman 2010), suggests it is important to quality-control visual depth in S3D displays. Furthermore, we have shown that audio depth can influence visual depth perception in S3D displays. We therefore conclude that there could be scenarios where it is important to quality-control audio-visual depth in S3D media.

All this experimentation was undertaken using a calibrated experimental setup for which the value of the MAD was known. In preparing this experimental setup, we addressed the third part of the over-arching research question – the integration of spatial sound and S3D display systems. We used an LG BM-LDS302 47 inch 3DTV display and two frequency matched K-Array KT20 loudspeakers, positioned using motorised rails. This is a very simple spatial sound system that served the purpose of delivering reliable localisation cues to the participant for all the degrees of spatial freedom our experiments required. The MAD for our experimental set up was found to be approximately 25 cm, when the participant was positioned at a listening distance (between participant and back speaker) of 1.81 m.

We have shown that audio depth can influence perception of depth in an S3D image. But this is different to assessing the importance of considering audio-visual interactions when designing content and engineering systems. Determining the im-

portance should be driven by conceiving, implementing and evaluating possible application scenarios for the effect, in an analogous manner to the work we present in Chapter 5 concerning the evaluation of quality-controlling the binocular effect. Such work could be seen as a focus of further work, as discussed in Section 9.4. In this thesis we have presented evidence that suggests the subjective evaluation of an application scenario could prove positive. Though the number of potential applications of the effect reported in this thesis may be limited by its size, the literature suggests that under certain conditions the audio-visual bias could be larger (such as degraded and bistable visual stimuli as explained in Section 3.2.3 and discussed in Section 3.3). To conclude, the work presented in this thesis gives us greater confidence in answering the over-arching research question with a clear and resounding, "Perhaps."

9.3 Novel contributions

There are novel contributions made by the work presented in this thesis.

- A review of literature related to audio-visual depth perception.
- An evaluation of subjective impressions of high-quality S3D media with quality-controlled binocular cues.
- An environmentally valid measurement of the MAD for a TV viewing scenario.
- A consideration of the implications of the MAD for content designers and systems engineers.
- A proposal and evaluation of a RAD perception screening test.
- An observation of audio depth influencing depth perception in S3D images.
- A plot of the psychometric function showing how the impact of this audio-visual interaction depends upon the auditory depth.
- A measurement of cross-modal bias created by the audio-visual interaction.

9.4 Further work

There are a number of further research questions that have arisen from our work, which we did not have the time or resources to address in this thesis. HCI is a very applied field of science, so it is right for related research to be directed by a focus upon the application of the research in relevant industries. For this reason, the most

obvious matters that still need to be addressed are those that stand in the way of applying this research to the industries of display system engineering and content design. We therefore propose three main points of focus for future work. Firstly, this further work could build upon the points made in our literature review to look for conditions under which audio can have a greater influence upon visual depth. Secondly, it could seek to implement possible application scenarios and investigate the impact of commercially available spatial sound systems upon audio-visual depth interactions. Finally, it could seek to evaluate such implementations of application scenarios in a manner similar to our evaluation of media that implements quality-controlled binocular depth cues.

Our literature review concluded that the majority of strong audio-visual interactions occur when the visual stimuli are either degraded, or bi-stable. It was a conscious decision for our work to focus on unadulterated S3D visual stimuli in the hope that the effect observed might have a wider scope. Depth perception in S3D displays is a degradation of natural depth perception, despite it being an improvement upon depth perception in traditional 2D displays. Now that we have shown a small effect does exist for unadulterated S3D stimuli, researchers may wish to look at other conditions for which the effect is larger. Specific suggestions that can be drawn from our literature review include: stimuli appearing in the periphery of the viewer's vision; stimuli where the binocular depth has been compressed significantly, in a manner that is inconsistent with the pictorial cues; S3D images that have been degraded significantly by crosstalk (Tsirlin et al. 2011); blurred or noisy stimuli; and conventional 2D images instead of S3D images.

The typical cycle of HCI research begins with studies of human behaviour and thinking, which feed into the design of new computing algorithms, design principles and guidance concerning the production of a quality user experience. Implementations of these outcomes are then evaluated in further human studies, and so the cycle often begins again. In this thesis we have progressed as far as performing a set of human trials and deriving some design recommendations. Further work could seek to implement possible application scenarios for the effect using commercially available technology. There could be a number of interesting research questions posed by this, including the impact of different commercially available spatial sound systems upon the effect, and the value of the effect for displays with very small depth budgets (e.g. tablet computers).

The final step in the HCI cycle, before the effect could be applied in industry, would be to evaluate the implementation of an application scenario. If appropriate, this could take a subjective approach, much like our work presented in Chapter

5. If the selected application scenario seeks to alter the user's task performance in something, then a quantitative evaluation may be required. Having identified and measured a perceivable cross-modal effect in depth perception, implemented an application scenario and demonstrated its value through a qualitative or quantitative evaluation, we could then be sure that *it is important to quality-control audio-visual depth by considering audio-visual interactions when integrating S3D displays and spatial sound systems.*

Bibliography

M. A. Akeroyd, S. Gatehouse, and J. Blaschke. The detection of differences in the cues to distance by elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 121(2):1077–1089, 2007.

C. André, J.-J. Embrechts, and G. V. Jacques. Adding 3D sound to 3D cinema: Identification and evaluation of different reproduction techniques. In *2010 International Conference on Audio Language and Image Processing (ICALIP)*, pages 130 –137, 2010.

D. H. Ashmead, D. LeRoy, and R. D. Odom. Perceptions of the relative distances of nearby sound sources. *Perception and Pyschophysics*, 47(4):326–331, 1990.

B. Atal and M. Schroeder. Apparent sound source translator. *U.S. Patent*, 3,236,949, 1963.

Audacity. Audacity: a free, open source, cross-platform software for recording and editing sounds. `http://audacity.sourceforge.net/`, March 2015. (Accessed on this date).

P. W. Battaglia, R. A. Jacobs, and R. N. Aslin. Bayesian integration of visual and auditory signals for spatial localisation. *Journal of the Optical Society of America*, 20(7):1391–1397, 2003.

D. Batteau. The role of the pinna in human localization. *Proceedings of the Royal Society of London*, 168(11):158–180, 1967.

A. Berkhout. A holographic approach to acoustical control. *Journal of the Audio Engineering Society*, 36:977–995, 1988.

A. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America*, 93(5):2764–2778, 1993.

J. Berry. Using 3D sound to influence perception of depth in 3D stereoscopic images, 2011. Undergraduate Thesis, School of Engineering and Computing Sciences, Durham University.

J. Berry, D. Budgen, and N. Holliman. Evaluating subjective impressions of quality controlled 3D films on large and small screens. *Journal of Display Technology*, 2015.

J. S. Berry, D. A. T. Roberts, and N. S. Holliman. 3D sound and 3D image interactions: a review of audio-visual depth perception. In *Proceedings of Human Vision and Electronic Imaging XIX - SPIE*, volume 9014, pages 901409–16, 2014.

B. Block and P. McNally. *3D Storytelling: How stereoscopic 3D works and how to use it*. Taylor & Francis, 2013.

A. D. Blumlein. Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems. *British Patent*, 34657, 1933.

M. Boone. Acoustic rendering with wave field synthesis. In *Proceedings of the ACM SIGGRAPH and Eurographics Campfire on Acoustic Rendering for Virtual Environments*, pages 37–45, 2001.

M. Boone, D. de Vries, and P. van Tol. Spatial sound field reproduction by wave field synthesis. *Journal of the Audio Engineering Society*, 43(12):1003–1012, 1995.

M. M. Boone. Multi-actuator panels (MAPs) as loudspeaker arrays for wave field synthesis. *Journal of the Audio Engineering Society*, 52(7/8):712–723, 2004.

A. L. Bowen, R. Ramachandran, J. A. Muday, and J. A. Schirillo. Visual signals bias auditory targets in azimuth and depth. *Experimental Brain Research*, 214: 403–414, 2011.

N. D. Brener, J. O. Billy, and W. R. Grady. Assessment of factors affecting the validity of self-reported health-risk behavior among adolescents: evidence from the scientific literature. *Journal of adolescent health*, 33(6):436–457, 2003.

A. Bronkhorst. Modeling auditory distance perception in rooms. *Human Factors*, 397(6719):517–520, 2002.

A. W. Bronkhorst and T. Houtgast. Auditory distance perception in rooms. *Nature*, 397:517–520, 1999.

F. P. Brooks, M. Ouh-Young, J. J. Batter, and P. J. Kilpatrick. Project GROPE - haptic displays for scientific visualisation. *Computer Graphics*, 24(4):177–185, 1990.

D. Brungart and W. Rabinowitz. Auditory localization of nearby sources. Head-related transfer functions. *Journal of the Acoustcal Society of America*, 106:1465–1479, 1999.

D. Burr and D. Alais. Combining visual and auditory information. *Progress in Brain Research*, 155:243–258, 2006.

K. Cater, R. Hull, T. Melamad, and R. Hutchins. An investigation into the use of spatialised sound in locative games. In *Proceedings of CHI*, pages 2015–2020, 2007.

M. Chion and C. Gorbman. *Audio-vision: sound on screen*. Columbia University Press, 1994.

E. Y. Choueiri. Optimal crosstalk cancellation for binaural audio with two loudspeakers. `http://www.princeton.edu/3D3A/Publications/BACCHPaperV4d.pdf`, Nov 2011. (Accessed on this date).

J. J. Clark and A. L. Yuille. *Data fusion for sensory information processing systems*. Kluwer Academic, Norwell, Massachusetts, 2001.

F. B. Colavita. Human sensory dominance. *Perception and Physcophysics*, 16(2): 409–412, 1974.

P. D. Coleman. Failure to localize the source distance of an unfamiliar sound. *The Journal of the Acoustical Society of America*, 34(3):354–346, 1962.

P. D. Coleman. An analysis of cues to auditory depth perception in free space. *Psychological Bulletin*, 60(3):302–315, 1963.

P. D. Coleman. Dual role of frequency spectrum in determinaton of auditory distance. *The Journal of the Acoustical Society of America*, 44(2):631–632, 1968.

V. Conrad, A. Bartels, M. Kleiner, and U. Noppeney. Audiovisual interactions in binocular rivalry. *The Journal of Vision*, 10(10):27, 2010.

S. Coren and L. Ward. *Sensation and Perception*, pages 273–294. Orlando: Harcourt Brace Jovanovich Publishers, 1989.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009. ISBN 0262033844, 9780262033848.

D. Corrigan, M. Gorzel, J. Squires, and F. Boland. Depth perception of audio sources in stereo 3D environments. In *Proceedings of Stereoscopic Displays and ApplicationsXXIV - SPIE*, volume 8648, pages 864816–13, 2013.

B. Cullen, D. Galperin, K. Collins, B. Kapralos, and A. Hogue. The effects of audio on depth perception in S3D games. In *Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound*, pages 32–39, 2012.

B. Cullen, D. Galperin, K. Collins, A. Hogue, and B. Kapralos. The effects of 5.1 sound presentations on the perception of stereoscopic imagery in video games. In *Proceedings of Stereoscopic Displays and Applications XXIV - SPIE*, volume 8648, pages 864815–864815, 2013.

B. Cumming and G. DeAngelis. The physiology of stereopsis. *Annual Review of Neuroscience*, 24:203–238, 2001.

W. P. J. de Bruijn and M. M. Boone. Application of wave field synthesis in life-size videoconferencing. In *114th Convention of the Audio Engineering Society*, pages 1–17, 2003.

N. A. Dodgson. Variation and extrema of human interpupillary distance. In *Electronic Imaging 2004*, pages 36–46. International Society for Optics and Photonics, 2004.

N. A. Dodgson. Autostereoscopic 3D displays. *Computer*, 38(8):31–36, 2005.

N. A. Dodgson. On the number of viewing zones required for head-tracked autostereoscopic display. In *Electronic Imaging 2006*, pages 60550Q–60550Q. International Society for Optics and Photonics, 2006.

N. A. Dodgson. Optical devices: 3D without the glasses. *Nature*, 495(7441):316–317, 2013.

R. Duda and W. Martens. Range dependence of the response of a spherical head model. *Journal of the Acoustcal Society of America*, 104:3048–3058, 1998.

A. Ecker and L. Heller. Auditory-visual interactions in the perception of a ball's path. *Perception*, 34:59–75, 2005.

A. S. Edwards. Accuracy of auditory depth perception. *The Journal of General Psychology*, 52(2):327–329, 1955.

W. H. Ehrenstein and A. H. Reinhardt-Rutland. A cross-modal aftereffect: Auditory displacement following adaptation to visual motion. *Perceptual and Motor Skills*, 82:23–26, 1996.

M. O. Ernst and H. H. Bulthoff. Merging the senses into a robust perception. *Trends in Cognitive Science*, 8(4):162–169, 2004.

M. Evrard, C. R. André, J. G. Verly, J.-J. Embrechts, and B. F. G. Katz. Object-based sound re-mix for spatially coherent audio rendering of an existing stereoscopic-3D animation movie. In *Proceedings of Audio Engineering Society Convention 131*, 2011.

F. Fontana and D. Rocchesso. Auditory distance perception in an acoustic pipe. *ACM Transactions on Applied Perception*, 5(3):1–15, 2008.

M. B. Gardner. Proximity image effect in sound localization. *Journal of the Acoustical Society of America*, 43:163, 1968.

G. V. Glass, P. D. Peckham, and J. R. Sanders. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3):pp. 237–288, 1972.

B. Goldstein. *Sensation and Perception.* Belmont : Thompson Wadsworth, 7th edition, 2007.

M. Grohn, T. Lokki, and T. Takala. Comparison of auditory, visual and audio-visual navigation in a 3D space. In *Proceedings of the 2003 International Conference on Auditory Display*, 2003.

D. Hairston, P. Laurienti, G. Mishra, and M. W. Jonathan Burdette. Multi-sensory enhancement of localisation under conditions of induced myopia. *Experimental Brain Research*, 152:404–408, 2003a.

W. Hairston, M. Wallace, J. Vaughan, B. Stein, J. Norris, and J. Schrillo. Visual localisation ability influences cross-modal bias. *Journal of Cognitive Neuroscience*, 15(1):20–29, 2003b.

E. H.A.Lagendijk and A. W. Bronkhorst. Fidelity of three-dimensional sound reproduction using a virtual auditory display. *Journal of the Acoustical Society of America*, 107(1):528–537, 2000.

R. Hartley and T. Fry. The binaural location of pure tones. *Physical Review*, 18: 431–442, 1921.

M. Hassenzahl and N. Tractinsky. User experience – a research agenda. *Behaviour and Information Technology*, 25(2):91–97, 2006.

S. Hecht and E. Smith. Intermittent stimulation by light : vi. area and the relation between critical frequency and intensity. *Journal of General Physiology*, 19(6): 979–989, 1936.

S. Hidaka, Y. Manaka, W. Teramoto, Y. Sugita, R. Miyauchi, J. Gyoba, Y. Suzuki, and Y. Iwaya. Alternaton of sound location induces visual motion perception of a static object. *PLoS ONE*, 4:e8188, 2009.

N. Holliman. Mapping perceived depth to regions of interest in stereoscopic images. In *Proceedings of Stereoscopic Displays and Virtual Reality Systems XI - SPIE Volume 5291*, 2004.

N. Holliman. Cosmic origins: experiences making a stereoscopic scientific movie. In *Proceedings of Stereoscopic Displays and Applications XXI - SPIE Volume 7237*, 2010.

N. Holliman, C. Baugh, C. Frenk, A. Jenkins, B. Froner, D. Hassaine, J. Helly, N. Metcalfe, and T. Okamoto. Cosmic cookery: making a stereoscopic 3D animated movie. In *Proceedings of Stereoscopic Displays and Virtual Reality Systems XIII - SPIE Volume 6055*, 2006.

N. S. Holliman, N. A. Dodgson, G. E. Favalora, and L. Pockett. Three-dimensional displays: A review and applications analysis. *IEEE Transactions on Broadcasting*, 57(2):362–371, 2011.

H. Howard. A test for the judgement of distance. *American Journal of Opthalmology*, 17:656–675, 1919.

ITU-R Recommendation BT.500-12. Methodology for the subjective assessment of the quality of television pictures. Technical report, International Telecommunication Union, Geneva, Switzerland, 2009.

G. Jones, D. Lee, N. Holliman, and D. Ezra. Controlling perceived depth in stereoscopic images. *Stereoscopic Displays and Virtual Reality Systems*, 4297:42–53, 2001.

B. Julesz, T. V. Papathomas, and F. Phillips. *Foundations of Cyclopean Perception*. MIT Press, 2006. ISBN 0262101130.

B. Kapralos, M. R. Jenkin, and E. Milios. Virtual audio systems. *Presence*, 17(6): 527–549, 2008.

N. Kayahara. Spinning dancer. `http://www.procreo.jp/labo/labo13.html`, 2003. (Accessed: Dec 2011).

A. Kolarik, S. Cirstea, and S. Pardhan. Evidence for enhanced discrimination of virtual auditory distance among blind listeners using level and direct-to-reverberant cues. *Experimental Brain Research*, 224(4):623–633, 2013a.

A. Kolarik, S. Cirstea, and S. Pardhan. Discrimination of virtual auditory distance using level and direct-to-reverberant ratio cues. *The Journal of the Acoustical Society of America*, 134(5):3395–3398, 2013b.

T. Kuhlen, I. Assenmacher, and T. Lentz. A true spatial sound system for cave-like displays using four loudspeakers. In *Proceedings of the 2nd international conference on Virtual reality*, ICVR'07, pages 270–279, 2007.

A. Kulkarni and S. Colburn. Role of spectral detail in sound source localisation. *Nature*, 396:747–749, 1998.

B. Kunz, L. Wouters, D. Smith, W. Thompson, and S. Creem-Regehr. Revisiting the effect of quality of graphics on distance judgements in virtual environments: A comparison of verbal reports and blind walking. *Attention and Perception Psychophysics*, 71(6):1284–1293, 2009.

C. Kyriakakis, P. Tsakalides, and T. Holman. Surrounded by sound. *IEEE Signal Processing Magazine*, 16(1):55–66, 1999.

M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn. Visual discomfort and visual fatigue of stereoscopic displays: a review. *Journal of Imaging Science and Technology*, 53(3):30201–1, 2009.

N. M. S. Langlands. Experiments on binocular vision. *Transactions of the Optical Society*, 28(2):45, 1926.

E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng. On the minimum audible difference in direct-to-reverberant energy ratio. *The Journal of the Acoustical Society of America*, 124(1):450–461, 2008.

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

T. Levola. Diffractive optics for virtual reality displays. *Journal of SID*, 14(5): 467–475, 2006.

T. Levola. Replicated slanted gratings with a high refractive index material for in and outcoupling of light. *Optical Express*, 15(5):2067–2074, 2007.

R. M. Lindsay and A. S. C. Ehrenberg. The design of replicated studies. *The American Statistician*, 47(3):pp. 217–228, 1993.

J.-C. Liou, K. Lee, F.-G. Tseng, J.-F. Huang, W.-T. Yen, and W.-L. Hsu. Shutter glasses stereo LCD with a dynamic backlight. In *Proceedings of Stereoscopic Displays and Applications XX - SPIE*, volume 7237, pages 72370X–8, 2009.

L. Lipton. *The foundations of stereoscopic cinema*. Van Nostrand Reinhold Company, 1982.

L. M. Lix, J. C. Keselman, and H. J. Keselman. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance f test. *Review of Educational Research*, 66(4):579–619, 1996.

F. Maeda, R. Kanai, and S. Shimojo. Changing pitch induced by visual motion illusion. *Current Biology*, 14(23):R990–R991, 2004.

K. Manabe and H. Riquimaroux. Sound controls velocity perception of visual apparent motion. *Journal of the Acoustical Society of Japan*, 21:171–174, 2000.

S. Mateeff, J. Hohnsbein, and T. Noack. Dynamic visual capture: Apparent auditory motion induced by a moving visual target. *Perception*, 14:721–727, 1985.

H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

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

S. W. Meru. Improving depth perception in 3D interfaces using sound. Master's thesis, University of Waterloo, Canada, 1995.

G. Miller. Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. *The Journal of the Acoustical Society of America*, 19: 609–619, 1947.

R. Minghim and A. R. Forrest. An illustrated analysis of sonification. In *Proceedings of the 6th IEEE Visualisation Conference*, pages 111–117, 1995.

D. R. Moore and A. J. King. Auditory peception: The near and far of sound localisation. *Current Biology*, 9:R361–R363, 1999.

D. S. Moore. *The Basic Practise of Statistics*. Macmillan Education Australia, 2nd edition, 1995.

A. Mouchtaris, P. Reveliotis, and C. Kyriakakis. Inverse filter design for immersive audio rendering over loudspeakers. *IEEE Transactions on Multimedia*, 2(2):77–87, 2000.

P. E. Napieralski, B. M. Alterhoff, J. W. Betraqnd, L. O. Long, S. V. Babu, C. C. Pagano, J. Kern, and T. A. Davis. Near-field distance perception in real and virtual environments using both verbal and action responses. *ACM Transaction on Applied Percepion*, 8(3):18, 2011.

S. H. Nielsen. Depth perception - finding a design goal for sound reproduction systems. In *Procceedings of the 90th Convention of the Audio Engineering Society*, 1991.

S. H. Nielsen. Auditory distance perception in rooms. *Journal of the Audio Engineering Society*, 41(10):755–770, 1993.

Nintendo. Nintendo 3DS - hardware features. `http://www.nintendo.com/3ds/features`, Dec 2011. (Accessed on this date).

Y. Nojiri, H. Yamanoue, A. Hanazato, M. Emoto, and F. Okano. Visual comfort/discomfort and visual fatigue caused by stereoscopic HDTV viewing. In *Proceedings of Stereoscopic Displays and Virtual Reality Systems XI - SPIE Volume 5291*, pages 303–313, 2004.

B. J. Oates. *Researching Information Systems and Computing*. Sage Publications, 2006.

M. Obrist, D. Wurhofer, F. Förster, T. Meneweger, T. Grill, D. Wilfinger, and M. Tscheligi. Perceived 3DTV viewing in the public: insights from a three-day field

evaluation study. In *Proceedings of the 9th international interactive conference on Interactive television*, EuroITV '11, pages 167–176, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0602-7.

M. Obrist, D. Wurhofer, M. Gärtner, F. Förster, and M. Tscheligi. Exploring children's 3DTV experience. In *Proceedings of the 10th European conference on Interactive tv and video*, EuroiTV '12, pages 125–134, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1107-6.

M. Obrist, D. Wurhofer, T. Meneweger, T. Grill, and M. Tscheligi. Viewing experience of 3DTV: An exploration of the feeling of sickness and presence in a shopping mall. *Entertainment Computing*, 4:71–81, 2013.

J. Ohlsson, G. Villarreal, M. Abrahamsson, H. Cavazos, A. Sjostrom, and J. Sjostrand. Screening merits of the lang II, frisby, randot, titmus, and TNO stereotests. *Journal of AAPOS*, 5(5):316–322, 2001.

H. Ohmura. Intersensory influences on the perception of apparent movement. *Japanese Psychological Research*, 29:1–9, 1987.

S. Pala, R. Stevens, and P. Surman. Optical cross-talk and visual comfort of a stereoscopic display used in a real-time application. In *Electronic Imaging 2007*, pages 649011–649011. International Society for Optics and Photonics, 2007.

S. E. Palmer. *Vision science: Photons to phenomenology*, volume 1. MIT press, Cambridge, MA, USA, 1999.

R. Patterson. Human stereopsis. *Human Factors*, 34(6):669–692, 1992.

E. S. Pearson and N. W. Please. Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika*, 62(2):223–241, 1975.

D. R. Perrott and K. Saberi. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, 87(4):1728–1731, 1990.

K. P. Peter Barnecutt. Auditory perception of relative distance of traffic sounds. *Current Psychology*, 73(1):93–101, 1998.

M. Phan, K. Schendel, G. Recanzone, and L. Robertson. Visual spatial localization deficits following bilateral parietal lobe lesions in a patient with Balint's syndrome. *Journal of Cognitive Neuroscience*, 12:583–600, 2000.

M. Pölönen, M. Salmimaa, J. Takatalo, and J. Häkkinen. Subjective experiences of watching stereoscopic Avatar and U2 3D in a cinema. *Journal of Electronic Imaging*, 21(1):011006–1–011006–8, 2012.

H. O. Posten. The robustness of the one-sample t-test over the pearson system. *Journal of Statistical Computation and Simulation*, 9(2):133–149, 1979.

V. Pulkki. Virtual sound source positioning using vector based amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, 1997.

V. Pulkki. Localization of amplitude-panned virtual sources II: Two- and three-dimensional panning. *Journal of the Audio Engineering Society*, 49(9):753–767, 2001.

V. Pulkki and M. Karjalainen. Localization of amplitude-panned virtual sources I: stereophonic panning. *Journal of the Audio Engineering Society*, 49(9):739–752, 2001.

J. C. Read and I. Serrano-Pedraza. Stereo vision requires an explicit encoding of vertical disparity. *Journal of Vision*, 9(4):3, 1–13, 2009.

M. Rebillat, E. Corteel, and B. F. Katz. SMART-I^2: A new approach for the design of immersive virtual environments. In *Proceedings of EuroVR-EVE*, 2010.

G. H. Recanzone. Interactions of auditory and visual stimuli in space and time. *Hearing Research*, 258:89–99, 2009.

R. S. Renner, E. Steindecker, M. MüLler, B. M. Velichkovsky, R. Stelzer, S. Pannasch, and J. R. Helmert. The influence of the stereo base on blind and sighted reaches in a virtual environment. *ACM Trans. Appl. Percept.*, 12(2):7:1–7:18, Mar. 2015. ISSN 1544-3558.

J. A. Rice. *Mathematical Statistics and Data Analysis*, chapter 8.5, pages 267–272. Thomson Brookes/Cole, Duxbury, 3rd ed edition, 2007a.

J. A. Rice. *Mathematical Statistics and Data Analysis*, chapter 8.6, pages 272–278. Thomson Brookes/Cole, Duxbury, 3rd ed edition, 2007b.

W. Richards. Stereopsis and stereoblindness. *Experimental Brain Research*, 10: 380–388, 1970.

C. Richardt, L. Świrski, I. P. Davies, and N. A. Dodgson. Predicting stereoscopic viewing comfort using a coherence-based computational model. In *Proceedings of the International Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, pages 97–104. ACM, 2011.

K. Rushing. *Home Theater Design: Planning and Decorating Media-Savvy Interiors*, chapter 3, page 60. Rockport Publishers, 2004.

C. B. Seaman. Qualitative methods in empirical studies of software engineering. *IEEE Transactions on Software Engineering*, 25(4):557–572, 1999.

R. Sekuler, A. B. Sekuler, and R. Lau. Sound alters visual motion perception. *Nature*, 385(4):308, 1997.

G. Sergi. Knocking at the door of cinematic artifice: Dolby Atmos, challenges and opportunities. *The New Soundtrack*, 3(2):107–121, 2013. doi: 10.3366/sound.2013.0041.

P. J. Seuntiëns. *Visual Experience of 3D TV*. PhD thesis, Eindhoven University of Technology and Phillips Research Eindhoven, 2006.

P. J. Seuntiëns, I. E. Heynderickx, W. A. IJsselsteijn, P. M. J. van den Avoort, J. Berentsen, I. J. Dalm, M. T. Lambooij, and W. Oosting. Viewing experience and naturalness of 3D images. In *Proceedings of Three-Dimensional TV, Video, and Display IV - SPIE Volume 6016*, volume 6016, pages 601605–7, 2005. doi: 10.1117/12.627515. URL `http://dx.doi.org/10.1117/12.627515`.

W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, 2002.

L. Shams and R. Kim. Crossmodal influences on visual perception. *Physics of Life Reviews*, 7:269–284, 2010.

L. Shams, J. Allman, and S. Shimojo. Illusory visual motion induced by sound. In *31st Annual Meeting of the Society for Neuroscience*, volume 27, page 1340, 2001.

L. Shams, Y. Kamitani, and S. Shimojo. Visual illusion induced by sound. *Cognitive Brain Research*, 14:147–152, 2002.

L. Shams, W. Ma, and U. Beierholm. Sound-induced flash illusion as an optimal percept. *NeuroReport*, 16:1923–1927, 2005.

T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks. The zone of comfort: Predicting visual discomfort with stereoscopic displays. *Journal of Vision*, 11(8):1–29, 2011.

S. Shimojo and O. Hikosaka. Visual motion sensation yielded by non-visually driven attention. *Vision Research*, 37:1575–1580, 1997.

W. Simpson and L. D. Stanton. Head movement does not facilitate perception of the distance of a source of sound. *The American Journal of Psychology*, 86(1): 151–159, 1973.

G. S. Smith. Human color vision and the unsaturated blue color of the daytime sky. *American Journal of Physics*, 73(7):590–597, 2005.

S. Soto-Faraco, J. Lyons, M. Gazzaniga, C. Spence, and A. Kingstone. The ventriloquist's effect in motion: Illusory capture of dynamic information across sensory modalities. *Cognitive Brain Research*, 14:139–146, 2002.

J. M. Speigle and J. M. Loomis. Auditory distance perception by translating observers. In *IEEE 1993 Symposium on Research Frontiers inVirtual Reality*, pages 92–99, 1993.

E. S. Spelke, W. S. Born, and F. Chu. Perception of moving, sounding objects by four-month-old infants. *Perception*, 12:719–732, 1983.

J. P. Springer, C. Sladeczek, M. Scheffler, J. Hochstrate, F. Melchior, and B. Frohlich. Combining wave field synthesis and multi-viewer stereo displays. In *Proceedings of the IEEE Virtual Reality Conference*, 2006.

H. E. Staal and D. C. Donderi. The effect of sound on visual apparent movement. *American Journal of Psychology*, 96:95–105, 1983.

J. Strutt. On our perception of sound direction. *Philosophical Magazine*, 13:214–232, 1907.

T. Z. Strybel and D. R. Perrott. Discrimination of relative distance in the auditory modality: The success and failiure of the loudness discrimination hypothesis. *Journal of the Acoustical Society of America*, 76(1):318–320, 1984.

Y.-C. Tai, S. Gowrisankaran, S.-n. Yang, J. E. Sheedy, J. R. Hayes, A. C. Younkin, and P. J. Corriveau. Depth perception from stationary and moving stereoscopic three-dimensional images. In *Proceedings of Stereoscopic Displays and Applications XXIV - SPIE Volume 8648*, 2013.

M. Taylor and C. Creelman. PEST: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America*, 41(4):782–787, 1967.

THX. HDTV set up. http://www.thx.com/consumer/home-entertainment/home-theater/hdtv-set-up/, June 2013. (Accessed on this date).

I. Tsirlin, L. M. Wilcox, and R. S. Allison. The effect of crosstalk on perceived depth from disparity and monocular occlusions. *Special Issue of the IEEE Transactions on Broadcasting: 3D-TV Horizon: Contents, Systems, and Visual Perception*, 57 (2):445–453, 2011.

I. Tsirlin, L. M. Wilcox, and R. S. Allison. The effect of crosstalk on depth magnitude in thin structures. *Journal of Electronic Imaging*, 21(1), 2012.

A. Turner. Enhancing 3D visualisation using 3D sound, 2010. Undergraduate Thesis, School of Engineering and Computing Sciences, Durham University.

A. Turner, J. S. Berry, and N. Holliman. Can the perception of depth in stereoscopic images be influenced by 3D sound? In *Proceedings of Stereoscopic Displays and Virtual Reality Systems XXII - SPIE Volume 7863*, pages 2015–2020, 2011.

K. Ukai and P. A. Howarth. Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations. *Displays*, 29(2):106 – 116, 2008.

A. Valjamae and S. Soto-Faraco. Filling in visual motion with sounds. *Acta Psychologica*, 129:249–254, 2008.

R. van Ee, J. J. A. van Boxtel, A. L. Parker, and D. Alals. Multisensory congruency as a mechanism for attentional control over perceptual selection. *The Journal of Neuroscience*, 29:11641–11649, 2009.

E. Verheijen. *Sound reproduction by wave field synthesis*. PhD thesis, Technical University Delft, The Netherlands, 1998.

F. Volk, U. Muhlbauer, and H. Fastl. Minimum audible distance (MAD) by the example of wave field synthesis. In *38th German Annual Conference on Acoustics (DAGA)*, pages 319–320, 2012.

G. von Bekesy. Uber die entstehung der entfernungsemfindung beim horen. *Akustische Zeitschrift*, 4:21–31, 1938.

R. M. Warren, E. A. Sersen, and E. B. Pores. A basis for loudness-judgments. *The American journal of psychology*, pages 700–709, 1958.

K. Watanabe and S. Shimojo. When sound affects vision. *Psychological Science*, 12 (2):109–116, 2001.

F. A. Wichmann and N. J. Hill. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8):1293–1313, 2001.

V. Wills. An empirical investigation into the human ability to localise audio in depth, 2012. Undergraduate Thesis, School of Engineering and Computing Sciences, Durham University.

K. M. Wittkowski and T. Song. *muStat: Prentice Rank Sum Test and McNemar Test*, 2012. URL `http://CRAN.R-project.org/package=muStat`. R package version 1.7.0.

A. J. Woods. How are crosstalk and ghosting defined in stereoscopic literature? In *Proceedings of SPIE Stereoscopic Displays and Applications Conference XXII*, volume 7863, 2011.

Y. Y. Yeh and L. D. Silverstein. Limits of fusion and depth judgment in stereoscopic color displays. *Human Factors*, 32(1):45–60, 1990.

P. Zahorik. Estimating sound source distance with and without vision. *Optometry and Vision Science*, 78(5):270–275, 2001.

P. Zahorik. Assessing auditory distance perception using virtual acoustics. *Journal of the Acoustical Society of America*, 11(4):1832–1846, 2002.

P. Zahorik, F. Wightman, and D. Kistler. On the discriminability of virtual and real sound sources. In *Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on*, pages 76 –79, oct 1995.

P. Zahorik, D. S. Brungart, and A. W. Bronkhorst. Auditory distance perception in humans: a summary of past and present research. *Acta Acustica United with Acustica*, 91:409–420, 2005.

Z. Zhou, A. D. Cheok, X. Yang, and Y. Qui. An experimental study on the role of 3D sound in the augmented reality environement. *Interacting With Computers*, 16:1043–1068, 2004.