THE ANALYSIS OF FREQUENCY DEPENDENT VERTICAL LOCALISATION THRESHOLDS AND THE PERCEPTUAL EFFECTS OF VERTICAL INTERCHANNEL CROSSTALK

Rory Wallis

Applied Psychoacoustics Lab

School of Computing and Engineering

University of Huddersfield

March 2017

A thesis submitted to the University of Huddersfield in partial fulfillment of the requirements for the degree of Doctor of Philosophy

COPYRIGHT STATEMENT

i. The author of this thesis (including any appendices and/or schedules to this thesis) own any copyright in it (the "Copyright") and s/he has given the University of Huddersfield the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

ii. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the University of Huddersfield Library. Details of these regulations may be obtained from the Librarian. This page must form any part of any such copies made.

iii. The ownership of any patents, designs, trade marks, and any and all other intellectual property rights except for the Copyright title (the "Intellectual Property Rights") and any other reproduction of copyright works, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

ABSTRACT

In the context of microphone techniques for recording three-dimensional (3D) sound in an acoustic space, vertical interchannel crosstalk occurs when the height layer of microphones capture direct sound. This effect can cause sound images to be formed as vertically oriented phantom images, at positions intermediate between the main and height layer of loudspeakers, as opposed to at the desired position of the main layer. Additional spatial and timbral effects will also be perceived, although these have not yet been examined in the literature.

Previous research has examined the minimum amount of attenuation of direct sound in the height layer necessary to prevent vertical interchannel crosstalk from affecting the perceived location of the main channel signal, which has become known as the 'localisation threshold'. However, existing methods of applying this have not considered the frequency dependency of median plane localisation. The present thesis therefore examined if localisation thresholds could be applied through the frequency dependent manipulation of the direct sound in the height layer (band reduction), as well as the most salient perceptual effects of vertical interchannel crosstalk. The operation of the precedence effect in the median plane was also considered.

A review of human localisation mechanisms was first conducted, with a particular focus on how such characteristics might be able to be exploited for the development of a band reduction method. Additionally, consideration was also given to how secondary vertical sources might affect direct sounds, in order to gain further understanding of what the most salient effects of vertical interchannel crosstalk might be.

The frequency dependency of localisation thresholds was considered in anechoic conditions, with subsequent localisation experiments being conducted to assist in explaining the results. Following this, localisation thresholds using blanket reduction (attenuation of the direct sound in the height layer evenly across the spectrum) were analysed. The frequency dependency of localisation threshold was subsequently examined in a natural listening environment, with a series of band reduction methods being developed based on the

results. The band and blanket reduction thresholds were then verified in localisation tests. The final experiment considered the most salient effects of vertical interchannel crosstalk, how these were affected when the different localisation threshold methods were applied and which was the most preferred method by subjects.

The results showed that localisation thresholds are frequency dependent in both anechoic and natural listening environments. In particular, more level reduction was necessary for the mid-high frequencies compared to low frequencies. Additionally, a series of different band reduction methods were found to be effective. Elicitation experiments showed that the most salient effects of vertical interchannel crosstalk were increases in vertical image spread, source elevation, loudness and fullness, with the perception of these when the localisation threshold was applied being dependent on the method being used. Moreover, although subjective preference could not discriminate between the methods tested, the presence of direct sound in the height layer was consistently preferred compared to situations where it was absent. Furthermore, no evidence was found to support the existence of either the precedence effect or localisation dominance in the median plane.

ACKNOWLEDGEMENTS

The author would like to thank the following people for their help with respect to the completion of this thesis; Hyunkook Lee for his incredible levels of support and guidance throughout and for always being available when called upon; Daisy for supporting me through such a long project and for putting up with the years of neglect, as well as keeping me sane; my parents for continued support throughout my life and for putting me in the position to be able to reach the level that I have; all of the members of the Applied Psychoacoustics Lab past and present, especially Mark Wendl, Dale Johnson and Chris Gribben for various levels of support, assistance and guidance; John Francombe for his assistance with threshold detection methods and general advice; all of the staff and students from the University of Huddersfield's Music Technology department for their assistance with sitting listening tests; the staff at Bluerooms, particularly Ben Evans, for their technical support; the peer reviewers at the Journal of the Audio Engineering Society and Applied Sciences for their feedback on the journal papers, which has inevitably helped to improve the overall quality of the work; and the staff at David Brown Santasalo, particularly Andy, Rich, Danika and Emma, for their ongoing support throughout the duration of the project.

CONTENTS

WORD COUNT: 82,674

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

GLOSSARY

Apparent Source Width (ASW): A spatial attribute of a sound source corresponding to a broadening of the sound image.

Band Reduction: A method of applying localisation thresholds whereby the direct sound in the height layer undergoes frequency-dependent attenuation.

Blanket reduction: A method of applying localisation thresholds whereby the direct sound in the height layer is attenuated evenly across the frequency spectrum.

Directional Bands: Based on the work of Blauert [1969]. Describes a localisation phenomenon whereby 1/3-octave bands are localised on the basis of frequency when presented from the median plane. Certain frequency bands have been shown to relate to specific directions irrelevant of the position of the emitting loudspeaker.

Head-Related Transfer Function (HRTF): A localisation mechanism that includes the spectral filtering of the pinnae and shoulder and torso reflections. When sounds are incident from the median plane, HRTF cues are the sole mechanism used for localisation.

Horizontal Interchannel Crosstalk: When recording for conventional surround sound, the phantom imaging of sound sources at the reproduction stage is achieved based on time and level differences between pairs of microphones covering the sector in which a given source lies. Horizontal interchannel crosstalk occurs when microphones other than the intended pair record the signal from a source. Work undertaken by Lee [2006] demonstrated that the effects of horizontal interchannel crosstalk include an increase in source width and a decrease in locatedness.

Horizontal Plane: The plane about the listener that bisects the head at the entrances to the ear canals and runs parallel to the floor.

Interchannel Level Difference (ICLD): The difference in amplitude between two loudspeakers emitting a coherent source.

Interchannel Time Difference (ICTD): The time difference between two loudspeakers emitting a coherent source.

Listener Envelopment (LEV): A spatial attribute of a sound source that relates to the sensation that the listener is surrounded by the source.

Localisation Threshold: Based on the work of Lee [2011]. In the case that vertically arranged stereophonic loudspeakers emit a coherent signal, the localisation threshold refers to the minimum amount of attenuation of direct sounds necessary in the height layer for the resultant phantom image to be localised at the position of the main layer.

Masking threshold: Based on the work of Lee [2011]. In the case that vertically arranged stereophonic loudspeakers emit a coherent signal, the masking threshold refers to the minimum amount of attenuation of direct sounds necessary in the height layer for the height layer's signal to become inaudible.

Median Plane: The plane about a listener that bisects the head symmetrically

Phantom Image Elevation Effect: A localisation phenomena that refers to the increase in perceived elevation of a phantom image when the base angle between stereophonic loudspeakers on the horizontal plane increases.

Pitch-Height Effect: A localisation phenomena that refers to the systematic spatial arrangement of band-limited and tonal stimuli based on their frequency. Increases in frequency correspond to an increase in perceived elevation.

Precedence Effect: Localisationn phenomenon caused when stereophonic loudspeakers on the horizontal plane emit a coherent signal. A delay greater than 1.1 ms will result in the phantom image position corresponding to the exact location of the earlier loudspeaker.

Vertical Image Spread (VIS): A spatial attribute of a sound source corresponding to the increased vertical spread of the sound source. Conceptually similar to ASW.

Vertical Interchannel Crosstalk: Interference effect caused when direct sounds are present in the height layer of a 3D audio configuration. Can result in main channel images being formed as vertically oriented phantom images between the main and height layers. Additional spatial and timbral effects will also be perceived depending on the time and level relationship between the direct sound in the main and height layers.

0 INTRODUCTION

0.1 BACKGROUND TO THE RESEARCH

0.1.1 3D Audio and Vertical Interchannel Crosstalk

Audio reproduction systems for surround sound are currently in a state of evolution. Engineers are increasingly looking to improve on the spatial impression offered by conventional 5.1 systems through the incorporation of loudspeakers in the vertical domain. The implementation of these so-called 'height channels' has seen audio reproduction systems move into the third dimension, with systems such as Auro-3D [Auro Technologies 2016] and Dolby Atmos [Dolby Laboratories 2016] becoming more widely utilised. Such developments inevitably have implications for the recording process, as additional height layers of microphones are required alongside the pre-existing main channel layer in order to capture the necessary spatial information.

In conventional microphone techniques for horizontal surround sound, pairs of microphones are positioned to capture specific areas of the recording angle [Rumsey 2005]; examples of this being the 'critical linking' technique developed by Williams and Le Du [2000] and the 'OCT' technique from Theile [2001]. For such techniques, the phantom imaging of a given sound source in the reproduction stage is achieved based on the time and level differences between the source signal arriving at each of the microphones covering the recording sector in which the source lies. However, should microphones other than the intended pair pick up the direct sound from a source, which is referred to as interchannel crosstalk, then its phantom imaging at the reproduction stage may be affected [Theile 2001]. Experiments conducted by Lee [2006] showed that the most salient effects of interchannel crosstalk are an increase in source width and a decrease in locatedness

(Fig. 0.1). For the duration of this thesis, the term ‘horizontal interchannel crosstalk’ will be used to refer to interchannel crosstalk oriented between horizontally arranged microphones.

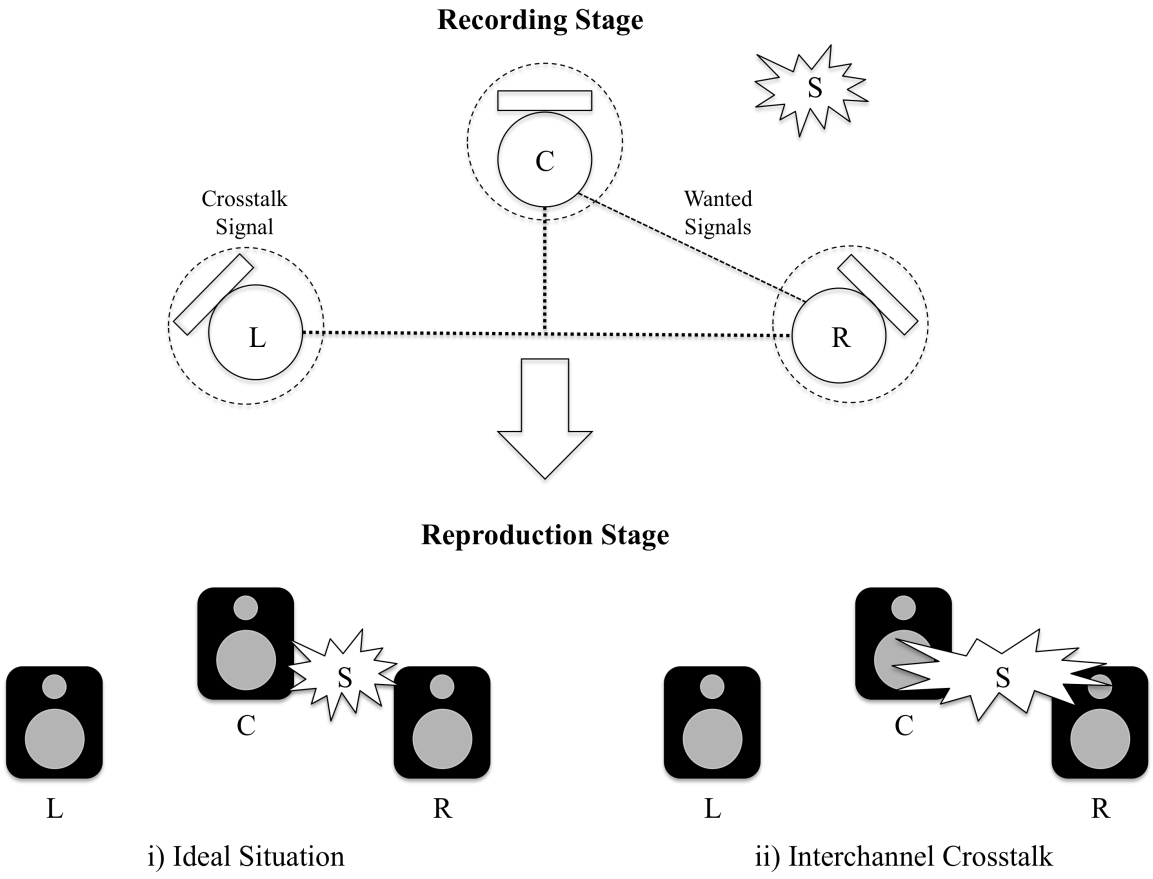

Fig. 0.1: Illustration of the cause and effect of horizontal interchannel crosstalk.

In the context of microphone techniques for recording three-dimensional (3D) sound in an acoustic space, interchannel crosstalk is also oriented between vertically arranged microphones. Consider a 3D microphone array consisting of two vertical layers of microphones (Fig. 0.2). The lower (main) layer would be typically used for horizontal source imaging, whilst the upper (height) layer would be used to enhance perceived listener envelopment (LEV). When recording for such formats, it is necessary to pay close attention to the amount of direct sound present in the height layer signal. The reason for this is as follows. Should there be excessive direct sound in the height layer then, at the reproduction stage, sound sources may be perceived as

vertically oriented phantom images at positions intermediate between the main and height loudspeaker layers. Additional spatial and timbral effects may also be perceived depending on the time and level relationships between the direct sounds in the respective layers. Collectively these properties comprise an interference effect referred to as 'vertical interchannel crosstalk'.

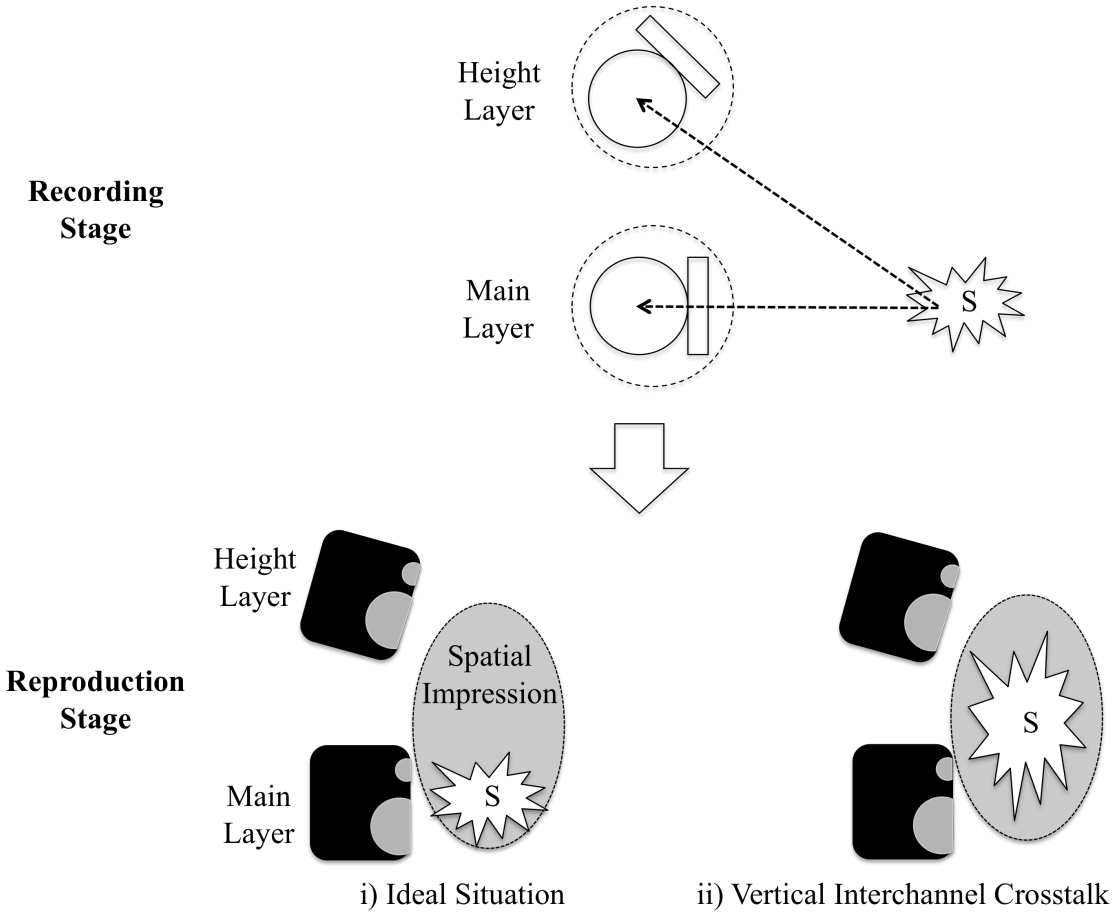

Fig. 0.2: Illustration of the cause and potential effects of vertical interchannel crosstalk.

From a practical standpoint, vertical interchannel crosstalk can be considered as being problematic. Should the effect be present in a recording then its influence would theoretically be lessened with sufficient attenuation of the height layer's amplitude at the mixing stage. However, an issue with this approach is that the resultant mix may be too dry, as engineers would be forced to sacrifice ambience levels in favour of

sound source location on the horizontal plane. It is clear then that the amount of direct sound recorded in the height layer needs to be limited at the recording stage in order that vertical interchannel crosstalk effects are minimised.

0.1.2 Localisation Thresholds

To date, vertical interchannel crosstalk has garnered little attention in the literature, with those studies that have addressed the effect being primarily concerned with the prevention of the vertical migration of the main channel signal from the position of the main layer. Lee [2011] presented anechoically recorded bongo and cello excerpts to subjects from a pair of vertically arranged loudspeakers directly in front of the listening position. The lower loudspeaker was not elevated, whilst the upper loudspeaker was elevated by 30°. Stimuli were presented as vertically oriented phantom images. The experiments subjectively measured the minimum amount of attenuation necessary in the upper loudspeaker for the resultant phantom image to be localised at the position of the lower loudspeaker. This was referred to as the 'localisation threshold'. Additionally, delays, ranging from 0 to 50 ms were applied to the upper loudspeaker with respect to the lower. The results showed that the localisation threshold for both sources was between -6 and -7 dB for interchannel time differences (ICTDs) up to 5 ms. This suggests that, should the path difference between the direct sound arriving at each layer be less than 1.7 m (corresponding to an ICTD of 5 ms), vertical interchannel crosstalk would not affect the perceived location of the main channel signal provided the amplitude of the direct sound in the height layer was reduced by between 6 and 7 dB.

A further study conducted by Stenzl et al. [2014] also considered localisation thresholds. Their study differed from that conducted by Lee [2011] in that phantom images were oriented between diagonally, as opposed to vertically, arranged stereophonic loudspeakers (i.e. left and height right). Localisation thresholds were considered for sound localisation at the main layer, with the height layer delayed (localisation bottom), and vice versa (localisation top). Delay times ranging from 0-50 ms were tested, with the test stimuli being male

speech, bongo and cello. The experimental data obtained generally showed good agreement with Lee [2011]. For the localisation bottom condition the threshold was between -4 and -8 dB for ICTDs up to 10 ms, whilst for localisation top the threshold was between -6 and -9 dB in the same range of delays. Furthermore, no significant effect of sound source was found for either condition, which also agreed with Lee [2011].

The aforementioned literature has demonstrated that, generally, vertical interchannel crosstalk will not affect the perceived elevation of the main channel signal when the amplitude of the direct sound in the height layer is attenuated with respect to that in the main layer by around 6 dB. However, there are numerous other factors that have not yet been considered. For example, as was mentioned by Lee [2011], the localisation threshold is not a complete masking of the vertical interchannel crosstalk signal. This implies that additional spatial and timbral effects would be perceived as a result of the interaction between the main channel and vertical interchannel crosstalk signals when the localisation threshold is applied. At present, the perceptual effects of such an interaction have not yet been explored. Furthermore, it is not yet known what the most salient perceptual effects of vertical interchannel crosstalk are, even before the application of localisation thresholds. Each of these points arguably requires attention, as this may dictate whether the direct sound in the height layer should be reduced only to the localisation threshold or further to the point where its influence on the main channel signal is entirely inaudible (the masking threshold).

0.1.3 The Band and Blanket Reduction Methods

An interesting feature of the Lee [2011] and Stenzl et al. [2014] studies was their approach to the application of localisation thresholds. In each case, the amplitude of the signal in the height layer was reduced as a whole (i.e. even attenuation across the frequency spectrum). Henceforth, this method will be referred to as ‘blanket reduction’. However, the literature demonstrates that vertical localisation is not consistent for all frequencies. Instead, frequency-based vertical localisation phenomena exist such as the pitch-height effect [Cabrera and Tiley 2003] and directional bands [Blauert 1969], both of which are discussed in detail in

Chapter One. If the vertical localisation of sound sources is frequency dependent, then this would seemingly indicate that localisation thresholds might be too. This in turn leads to the following question: would it be possible to apply localisation thresholds through the frequency-dependent attenuation of the vertical interchannel crosstalk signal? This is a method that will henceforth be referred to as 'band reduction'. Naturally, such a methodology would be difficult to achieve in a practical recording environment, however there are implications for the rendering of vertical stereophonic images in 3D sound mixing, particularly if the tonal and spatial effects of direct sounds featuring in the height layer are considered as being preferable.

0.1.4 The Precedence Effect

The studies conducted by both Lee [2011] and Stenzl et al. [2014] each found that the attenuation of the vertical interchannel crosstalk signal was always required in order that the localisation threshold was perceived to have been met. In neither study was a condition observed in which ICTD alone was sufficient. Such results indirectly suggest that the precedence effect is not a feature of vertical stereophony. The effect is considered in detail in Chapter Two, however a brief summary is offered thus. When two loudspeakers located on the horizontal plane radiate a coherent signal, an ICTD of 1.1 ms or more will result in the perceived location of the phantom image being at the position of the earlier loudspeaker [Blauert 1997]. The existence of the precedence effect naturally has implications for vertical interchannel crosstalk. If it can be demonstrated that the precedence effect operates between vertically arranged loudspeakers then this would indicate that vertical interchannel crosstalk would not affect the perceived position of the main channel signal provided there was sufficient ICTD. However, if the effect does not operate, as was suggested by Lee [2011] and Stenzl et al. [2014] then interchannel level differences (ICLD) would always be required.

0.2 RESEARCH QUESTIONS

From the above background, the following research questions, to be addressed in the present thesis, were derived:

1. How do localisation thresholds vary across the frequency spectrum and is there a sound source dependency for more natural stimuli?
2. Can any evidence be found to support the existence of the precedence effect for vertically arranged loudspeakers?
3. What are the most salient perceptual effects of vertical interchannel crosstalk?
4. How are the most salient effects affected when applying the localisation threshold using the band and blanket reduction methods and which method is more subjectively preferred?

Alongside the above, additional research questions and aims will arise as the thesis progresses. Each of these are discussed at the beginning of the relevant chapter.

The primary purpose of the work is to influence techniques both for the recording and rendering of 3D images. For example, one of the primary applications for the blanket reduction technique is in practical recording environments. In this regard, Lee [2011] has already proposed a series of techniques. It should be noted, however, that the scope of his experiments was somewhat limited, with only two sound sources and one loudspeaker configuration being tested. It can be argued then that a more rigorous analysis of the blanket reduction technique will increase understanding about how the microphones in the height layer should be positioned to prevent vertical interchannel crosstalk from affecting the perceived location of the main channel signal. Equally, knowledge about the operation of the precedence effect in this context will help to determine the types of microphones that should ideally be used in the height layer.

In addition, the band reduction technique will be predominantly applicable to image rendering techniques for 3D audio (although the blanket reduction technique will remain valid in this context). Should it be demonstrated that there exists a frequency dependency of localisation thresholds then band reduction will represent a creative way in which the location-based effects of vertical interchannel crosstalk can be minimised. This point is especially salient given that the height layer has an audible influence on the main channel signal when the localisation threshold is applied [Lee 2011]. It might therefore be that the most salient effects of vertical interchannel crosstalk at the localisation threshold are dependent on the threshold method being used (i.e. blanket or band reduction). Should this be the case then an analysis of such effects, how their audibility changes and subjective preference would be valuable in the context of 3D image rendering. It could be, for example, that there are benefits to having direct sound present in the height layer at the localisation threshold. In this case, the results could be used to inform the design of a 3D image-rendering tool that looks to enhance these effects, whilst keeping the main channel signals at the position of the main layer.

Throughout the thesis, the practical applications of the results with respect to both image rendering and recording techniques for 3D audio are discussed. The structure of the thesis is outlined below.

0.3 THESIS STRUCTURE

The subsequent chapters of the present thesis are arranged in the following way.

Chapter One considers the mechanisms used to localise elevated sound sources. The chapter begins with a brief overview of the general mechanisms used for sound localisation. Following this, sound localisation in the median plane is considered, with particular emphasis on spectral cues, head rotations and torso reflections. Subsequently, the frequency dependency of median plane localisation is discussed, which includes overviews of the pitch-height effect and directional bands. The chapter concludes with consideration of the phantom image elevation effect and its implications for vertical interchannel crosstalk.

Chapter Two explores the perceptual effects of secondary vertical sources. This chapter is divided into three broad sections. In the first, location based effects, including the precedence effect, are discussed. The second section considers the timbral effects of secondary vertical sources and, in particular, the perceptual effects of comb filtering. The final section discusses how secondary vertical sources can influence perceived spatial impression including apparent source width, listener envelopment and vertical image spread.

Chapter Three describes two experiments. The first experiment (Experiment One) considered localisation thresholds for octave bands of pink noise in an anechoic chamber. The purpose of this experiment was to determine whether or not there exists a frequency-dependency of localisation thresholds in anechoic conditions. The second experiment (Experiment Two) was a localisation test that was conducted in order to help understand the underlying location mechanisms relating to vertical interchannel crosstalk. The primary purpose of undertaking this experiment was to assist in understanding the results obtained in Experiment One.

Chapter Four describes a series of experiments that were conducted as a full analysis of both the band and blanket reduction methods. The first experiment in the chapter (Experiment Three) explored the blanket reduction method in more detail than had been done previously. The second experiment (Experiment Four) was an analysis of the frequency-dependency of localisation thresholds in a natural listening environment, which was important in order that a band reduction method could be developed. The final experiment in the chapter (Experiment Five) was a verification test. Firstly, a series of band reduction methods were derived based on the experimental data obtained from Experiment Four. These thresholds were then tested alongside the blanket reduction thresholds obtained from Experiment Three.

Chapter Five considers the perceptual effects of vertical interchannel crosstalk and the subjective preference of localisation thresholds. In Experiment Six, these effects were first elicited, with audibility tests then being conducted to determine those that were the most salient. Following this, grading tests were conducted in which the band and blanket localisation thresholds derived from Chapter 4 were applied, with subjects

grading how the most salient perceptual effects varied for each method. The chapter concludes with an analysis of the subjective preference of the localisation threshold methods developed during the thesis.

Chapter Six summarises the work conducted in the thesis and also discusses plans for future work.

0.4 ORIGINAL CONTRIBUTIONS

- It has been demonstrated that there exists a frequency dependency of localisation thresholds that is maintained for both anechoic and natural listening environments (Experiments One and Four).
- The frequency-dependency of localisation thresholds has been used to propose and test a series of methods of applying the localisation threshold whereby the direct sound in the height layer undergoes frequency-dependent attenuation. This is otherwise known as 'band reduction' (Experiment Five).
- The 'blanket reduction' method, whereby the direct sound in the height layer undergoes equal attenuation across the full frequency range, has been explored in further detail than had been done previously (Experiment Three).
- The perceptual effects of vertical interchannel crosstalk have been elicited, with further study conducted on how the perception of these attributes is affected by different methods of applying the localisation threshold (Experiment Six).
- The pitch-height effect has been shown to operate when octave band stimuli are presented from vertically arranged stereophonic loudspeakers, with the effects of ICTD on this phenomenon also being explored (Experiment Two).
- A novel threshold detection method, known as the 'adaptive method of adjustment' (AMOA) has been developed and used for testing (Experiments Three and Four).

- Novel hypotheses regarding the primary mechanisms used to determine the localisation threshold for both band-limited and complex stimuli have been proposed (Experiments One, Two, Three, Four and Five).
- No evidence could be found to support the operation of either the precedence effect or localisation dominance in the median plane (Experiments One, Two, Three, Four and Five).
- A relationship has been shown between directional bands and the pitch-height effect (Appendix A).

0.5 PUBLICATIONS

0.5.1 Journal Papers

Wallis, R. and Lee, H. [2015]: 'The Effect of Interchannel Time Difference on Localisation in Vertical Stereophony', *Journal of the Audio Engineering Society*, 63(10), pp. 767-776.

Wallis, R. and Lee, H. [2016]: 'Vertical Stereophonic Localisation in the Presence of Interchannel Crosstalk: The Analysis of Frequency-Dependent Localisation Thresholds, *Journal of the Audio Engineering Society*, 64(10), pp. 762-770.

Wallis, R. and Lee, H. [2017]: 'The Reduction of Vertical Interchannel Crosstalk: Analysis of Localisation Thresholds for Natural Sound Sources', *Applied Sciences,* 7(3), 278, pp. 1-20.

0.5.2 Conference Papers

Lee, H., Gribben, C., and Wallis, R. [2014]: 'Psychoacoustic Considerations in Surround Sound with Height', 28th Tonmeistertagung.

Wallis, R. and Lee, H. [2014]: 'Investigation into Vertical Stereophonic Localisation in the Presence of Interchannel Crosstalk', Audio Engineering Society 136th Convention, Preprint 9026.

Wallis, R. and Lee, H. [2015]: 'Directional Bands Revisited', Audio Engineering Society 138th Convention, Preprint 9278.

Wallis, R. and Lee, H. [2016]: 'The Reduction of Vertical Interchannel Crosstalk, The Analysis of Localisation Thresholds for Musical Sources', Audio Engineering Society 140th Convention, Preprint 9513

Wallis, R. and Lee, H. [2016]: 'The Frequency Dependency of Localisation Thresholds in the Presence of Reflections', 29th Tonmeistertagung.

1 THE LOCALISATION OF ELEVATED SOUND SOURCES

One of the primary aims of the present thesis is to develop methods in which vertical interchannel crosstalk can be prevented from affecting the perceived location of the main channel signal. This will be achieved through the analysis of localisation thresholds. In order that this can be achieved it is first necessary to gain an understanding of human localisation mechanisms, in particular those used to localise elevated sound sources. Blauert [1997] offered the following definition of the term 'localisation':

""Localisation" is the law or rule by which the location of an auditory event (e.g., its direction or distance) is related to a specific attribute or attributes of a sound event, or of another event that is somehow correlated with the auditory event. Examples include the relation of the position of the auditory event to the position of the sound source; the relation of the direction of the auditory event to the interaural sound level difference of the ear input signals; and the relation of the direction of the auditory event to the amplitude of head motion." [Blauert 1997]

The present chapter discusses human localisation mechanisms in detail. Firstly the general mechanisms used for horizontal sound localisation are considered, before particular focus is given to those cues used specifically for elevation perception. Additional localisation phenomena, which are considered as being relevant to the issue of vertical interchannel crosstalk, are also discussed.

1.1 GENERAL MECHANISMS USED FOR HORIZONTAL SOUND LOCALISATION

1.1.1 The Duplex Theory of Sound Localisation

First proposed by Strutt [1907], the 'duplex theory' of sound localisation considers the mechanisms used to determine the horizontal location of sound sources. The theory states that sound localisation in the horizontal plane is reliant on interaural cues, which can be thought of as being the differences between the same sound source arriving at each ear. Such cues can also be referred to as 'binaural' cues, as they necessitate the presence of both ears in order to function. Primarily, two interaural cues exist, interaural time difference (ITD) and interaural level difference (ILD). ITD is the result of a sound source arriving at one ear delayed with respect to the other (Fig. 1.1). ILD is a result of the shadowing effect of the head and results in the amplitude of the signal being greater at one ear [Howard and Angus 2009].

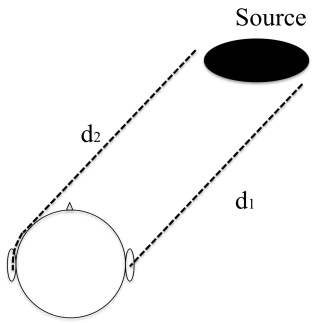

Fig. 1.1: A source to the right of the listener will stimulate the right ear before the left (ITD).

It has been shown in the literature that the predominant interaural cue used to localise a given sound source in horizontal space is dependent on the frequency of the source. In early localisation experiments conducted by Sandel et al. [1955] and Mills [1958], who used wide band noise and tonal stimuli respectively, it was

identified that ITD cues are the predominant mechanism used to localise low frequency sources. In both studies, a maximum frequency for the mechanism to be effective was reported as being around 1.5 kHz. Similarly, Scherer [1959, cited in Blauert 1997] found that subjects tasked with localising sinusoidal signals had an error rate of 50% at 800 Hz, which decreased steadily to 100% as the frequency increased above 1.6 kHz. On the other hand, ILD cues are predominantly utilised for the localisation of high frequency sources. Moushegian and Jeffress [1959] suggested that the intensity differences between the ears are low below 3-4 kHz and, as a result, frequencies between this threshold and the threshold at which the ITD cue begins to function are not localised as accurately. Further, it was proposed by Blauert [1972, cited in Blauert 1997] that low frequency signals, lacking frequency components above 1.6 kHz, are predominantly localised based on ITD, with ILD cues being used when energy above 1.6 kHz is present in the signal.

1.1.2 The Head-Related Transfer Function

ITD and ILD cues are essential for determining the azimuth of a given sound source. However, both Wallach [1940] and Perrett and Noble [1997] noted that such cues are ambiguous with respect to elevation perception and further that they contain no information regarding the front-back location of sound sources. Each of these is naturally important with respect to accurate localisation. Absence of the latter cue, for example, can result in front-back confusions, which occur when a subject identifies a sound incident from in front of them as being in the rear and vice versa [Wightman and Kistler 1999]. In addition to this, a sound source arriving from a point in space can yield the same interaural cues as another source at any location on a cone centered on the interaural axis, which is known as the 'cone of confusion' [Rottger et al. 2007]. The human auditory system therefore contains additional cues in order that accurate localisation is possible in the presence of these ambiguities. These cues comprise reflections from both the shoulder and torso, as well as the directional-dependent filtering of the outer ear (pinna), each of which are unique for each direction of incidence [Hebrank and Wright 1974a, Wightmann and Kistler 1999, Algazi et al. 2001]. Collectively these cues are known as the head-related transfer function (HRTF).

In general the ILD, ITD and HRTF cues co-exist in order to determine the location of a given sound source. However, a unique situation arises when sound sources are incident from the median plane (Fig. 1.2), the plane at which the head is bisected symmetrically [Blauert 1997]. Under such conditions, sound sources reach the ears at the same time and with equal amplitude, rendering binaural cues absent [Middlebrooks and Green 1991]. This leaves HRTF cues, in particular the spectral filtering of the pinna, as the primary mechanism for localisation [Roffler and Butler 1968a, Hebrank and Wright 1974a, Morimoto and Nomachi 1982]. Analysis of median plane localisation mechanisms is therefore beneficial, as the cues used for elevation perception are isolated. This is considered in the next section.

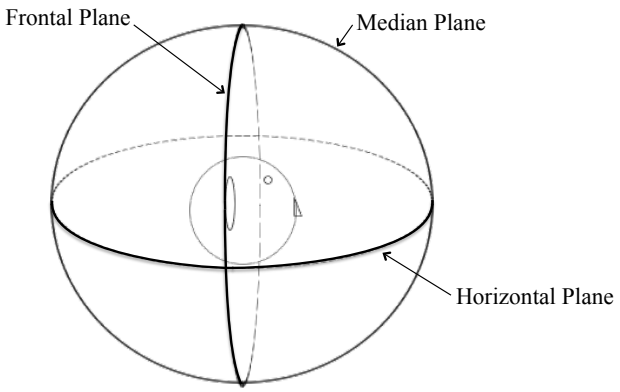

Fig. 1.2: Diagram of the locations of the median, frontal and horizontal planes in relation to a listener.

1.2 SOUND LOCALISATION IN THE MEDIAN PLANE

The following section presents a review of the mechanisms used to localise sound sources that are incident on the median plane. First, the prerequisites for accurate median plane localisation are discussed. Following this, attention is given to the role of the pinnae, head rotations and shoulder and torso reflections in the perception of elevation. The cues used to resolve front-back confusions are also discussed.

1.2.1 The Prerequisites for Accurate Vertical Localisation

Independent of the localisation mechanisms themselves, it has been demonstrated in the literature that numerous factors are necessary for a given sound source to be accurately localised in vertical space. Roffler and Butler [1968a] presented low and high-passed broadband noise and tonal stimuli to subjects from four loudspeakers located directly in front of, and 1.5 m away from, the listening position, with elevation angles of -13°, -2°, 9° and 20°. Subjects gave localisation judgments both in normal conditions and when the pinna cues were removed. For the unrestricted pinna condition, localisation was found to be accurate (i.e. at or near the position of the emitting loudspeaker) only for the filtered noise that contained frequencies above 7 kHz. No localisation judgments were accurate when the pinna cues were absent. From this study, the authors concluded that three criterion had to be satisfied in order for a sound source to be accurately localised in vertical space:

1. The pinnae must be present.
2. The sound source must be complex.
3. The sound source must contain frequencies above 7 kHz.

Subsequent experiments conducted by Gardner and Gardner [1973] further scrutinized the conclusions made by Roffler and Butler [1968a]. Nine loudspeakers were arranged directly in front of the listening position in an arc that spanned ±18°, at 4.5° intervals, on the median plane. The distance between each loudspeaker and the listening position was 10 feet. Subjects were required to localise broadband and band passed noise (2-10 kHz) with either free pinnae or with the pinnae blocked with a mold. The results generally showed good agreement with those reported by Roffler and Butler [1968a]. Firstly, localisation accuracy was poor for the pinnae blocked condition, which shows that the pinnae are necessary for elevated sound sources to be accurately localised. In addition, localisation judgments for the broadband source were more accurate than for the band passed sources, showing that complex sources are more accurately localised in vertical space.

Interestingly, despite the aforementioned agreements between the results of the two studies, there were notable differences with respect to the minimum spectral content necessary in the signal. Although Roffler and Butler [1968a] suggested that frequencies above 7 kHz were necessary, Gardner and Gardner [1973] found that judgments for the 3 kHz band passed source were accurate in 60% of cases. A subsequent study conducted by Hebrank and Wright [1974a] also suggested a minimum frequency markedly lower than 7 kHz. For their experiments, low and high pass filtered white noise was presented to subjects from a loudspeaker that could be moved to one of nine positions on an arc in a darkened anechoic chamber. With respect to the listening position, the arc spanned the region between -30° and +210°. From this experiment, the authors concluded that the features used in median plane localisation lay in the region between 3.8 and 16 kHz; this agreeing with Gardner and Gardner's [1973] results. Hebrank and Wright [1974a] hypothesised that the differences with respect to the Roffler and Butler [1968a] study were due to differing loudspeaker positions. Whereas Roffler and Butler [1968a] positioned loudspeakers in front of the listening position, Hebrank and Wright [1974a] used an arc on the upper median plane. It was therefore suggested that the necessity of 7 kHz for localisation determined by Roffler and Butler [1968a] was only relevant to localisation in front of the subject. However, it should be noted that this hypothesis is inconsistent with the results of Gardner and Gardner's [1973] study, in which median plane localisation was considered only for loudspeakers located in front of the listening position. Despite these differences, it is clear that accurate median plane localisation is not possible in the absence of the spectral cues provided by the pinnae, which are explored in the next section. In addition to this, sound sources must be complex and also feature energy in the high frequency region.

1.2.2 The Cues Used for Median Plane Localisation

1.2.2.1 The Role of the Pinnae

A key component of median plane localisation is the directional-dependent filtering of the pinnae, which act as 'linear filters whose transfer function depends on the distance and direction of the sound source' [Blauert

1997]. More simply, the pinnae can be thought of as placing a spectral 'fingerprint' on sound sources that is unique to the direction from which each source is incident. It was hypothesised by Hebrank and Wright [1974a] that such filtering is a result of delayed reflections from the concha interfering with sound directly entering the ear canal. Whereas ITD and ILD can be considered as being binaural cues, as both ears are necessary, the spectral filtering of the pinna is possible with a single ear and as a result can be considered as being a monaural cue [Gardner 1973, Blauert 1997]. Examples of how the ear input spectra varies when source elevation increases from 0° to 30° (0° azimuth) can be seen in Fig. 1.3.

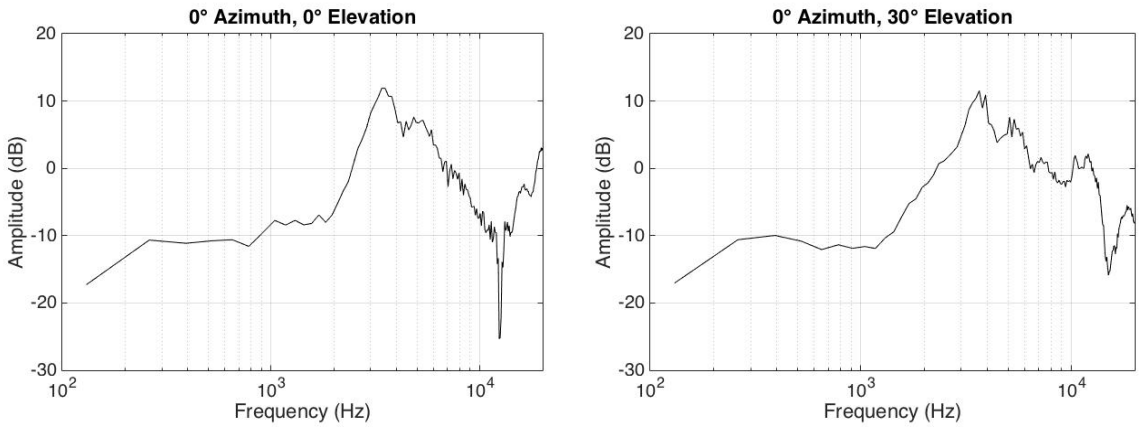

Fig. 1.3: The ear input spectra measured for sources at 0° azimuth with 0° (left) and 30° (right) elevation, using MIT'S KEMAR dummy head database [Gardner and Martin 2000].

With respect to median plane localisation, the spectral cues provided by the pinnae have been much explored in the literature. This has been achieved through a combination of binaural measurements and localisation studies. Shaw and Teranishi [1968] measured changes in HRTF with variations in source elevation by placing microphones in the ear canals of subjects. Source elevation with respect to the subject varied from 45° to -45°. Their measurements showed that the response level above 5 kHz was strongly dependent on elevation. In particular, a notch in the spectrum, centered on 6 kHz, was apparent for source elevation of -45°. This notch gradually increased to around 10 kHz as source elevation approached 45°. Shaw and Teranishi

[1968] therefore concluded that the ear is ‘best adapted to receive sound from an elevated source in the 5-10 kHz band’.

A subsequent study conducted by Hebrank and Wright [1974a], who analysed the filtering of the pinna for sounds incident from different directions on the median plane, confirmed the results obtained by Shaw and Teranishi [1968]. The stimuli used for their experiment were high pass (3.8-15.3 kHz), low pass (3.9-16.0 kHz) and unfiltered white noise, as well as white noise filtered with 1/12-octave band pass peaks (4.0-14.5 kHz) and bandstop notches (6.2-17.8 kHz). Stimuli were presented from loudspeakers arranged on an arc in the upper median plane. In addition, objective measurements were made by placing a microphone at the position of the eardrum in several artificial ears. From both experiments, the authors deduced that the key spectral cues for median plane localisation are as follows. A notch in the spectrum between 4 and 8 kHz cues frontal sounds, with additional increased energy above 13 kHz. Overhead perception relates to a 1/4-octave peak between 7 and 9 kHz, with a low pass slope at 10 kHz. Also, sounds in the rear are cued by a 10 kHz high pass slope, with a peak near 12 kHz. Alongside these results, it was also identified that the perception of elevation corresponds to a 1-octave notch in the spectrum in the region between 4 and 10 kHz. Further, an increase in the centre frequency of this notch resulted in the perception of increased elevation. This result therefore supported Shaw and Teranishi’s [1968] conclusions regarding both the range in which the elevation cues lie and also how the spectrum is modified as the elevation of the source increases.

A series of experiments conducted by Butler and Belendiuk [1977] were able to confirm the existence of an elevation-dependent notch in the frequency spectrum, in line with that reported by Hebrank and Wright [1974a]. During an initial experiment, objective measurements of the external ear transfer function were made by placing microphones in the ear canals of eight different subjects and averaging the overall spectrum. The test stimuli were 1/3-octave bands of noise, with centre frequencies ranging from 4-10 kHz. Stimulus presentation was from one of five loudspeakers positioned directly in front of the listening position with elevation angles of 0°, ±15° and ±30°. The objective measurements revealed the following. Firstly, for any vertical source position, the 4 kHz band always contained the most energy, with the 10 kHz band always

containing the least. However, the energy for the frequencies between these two extremes was found to vary with source elevation, with the most notable result being a correlation between source elevation and the energy in the 6.3 kHz band. Further objective measurements showed that a notch centered around 7 kHz was present in the spectrum when source elevation was 15°. In line with what was reported by both Shaw and Teranishi [1968] and Hebrank and Wright [1974a], the centre frequency of this notch decreased to around 5.5 kHz as source elevation decreased to -30°.

In a later study, Asano et al. [1990] presented time-stretched pulses to subjects from a loudspeaker located 1.5 m from the listening position. HRTF measurements were made at every 10° on the median plane, with the exception of 90°. Initial measurements showed that there was little directional dependence in the frequency region below 4-5 kHz, which agreed with what had been previously reported by Shaw and Teranishi [1968]. Additionally, notches in the spectrum were observed between 6 and 10 kHz, with a peak also being measured between 11 and 14 kHz for sounds incident from the front of subjects. As source elevation increased, decreases in this peak were observed, whilst the centre frequency of the 6-10 kHz notch also increased. This latter result shows good agreement with what was reported in the aforementioned studies. In general, the measurements made by Asano et al. [1990] demonstrated that the power at 0.5 kHz and between 3 and 5 kHz increased with frontal incidence, whilst 1-2 kHz increased for the rear. Further experiments were conducted in which HRTFs were convolved with sound sources to give the impression of median plane incidence. The method of simulation allowed the peaks and dips in the HRTFs to be simplified. This simulation led the authors to conclude that elevation perception cues are macroscopic and exist in the region above 5 kHz.

1.2.2.2 Head Rotation Cues

Alongside the spectral filtering of the pinnae, additional median plane localisation cues are provided as a result of head rotations. Wallach [1939, 1940] discussed their role in localisation, suggesting both that they could provide a dynamic cue and further that the effectiveness of this cue would lessen with increases in

source elevation. Experiments conducted by Thurlow and Runge [1967] demonstrated this to an extent. In their study, when high and low frequency noise and pulses were presented to subjects on the horizontal plane, head rotations were found to significantly reduce the localisation error compared to when no head movement was permitted. In addition, some limited improvements were also observed for low frequency sources incident on the median plane, which suggests that the head rotation cue for elevation perception operates in the low frequency range. However, as this study did not consider the effect of source elevation, it was not possible to evaluate whether or not Wallach's [1939, 1940] hypothesis regarding the weakening of the head rotation cue with increases in source elevation was accurate. It could at least be suggested that perhaps greater improvements in localisation accuracy for the low frequency median plane sources would have been observed in the study had the source elevation been lower than the 41° tested.

That head rotation cues for the localisation of sources incident on the median plane are only effective when the source contains low frequencies agrees with the results of a later study conducted by Perrett and Noble [1997]. Broadband and filtered white noise were presented to subjects from seven loudspeakers arranged in an arc on the upper median plane (30° intervals). Subjects were required to identify the perceived location of the stimuli both when no movement was permitted and when the head was rotated in a 60° arc on the horizontal plane. For both cases, normal and distorted pinna conditions were tested. With respect to localisation accuracy, elevation was only perceived for the low frequency stimuli when head movements were permitted. In addition, when low frequency energy was progressively removed there was a notable decrease in the effectiveness of head rotations for both the distorted and normal pinna conditions. From this study, a number of conclusions regarding head movements were made. Firstly, head rotations are able to increase median plane localisation accuracy provided that low frequencies are present in the source; this agreeing with Thurlow and Runge [1967]. Moreover, head rotations can provide comparable median plane localisation accuracy in the absence of the pinna. As a result of these conclusions, it was hypothesised that the dynamic elevation cues provided by head rotation, which they called the 'Wallach' cue after the studies of Wallach [1939, 1940], rely of the rate of change of ITD as a function of elevation; a hypothesis that was later confirmed by Rao and Xie [2005]. Interestingly, although Perrett and Noble [1997] identified that some

spatial regions are benefitted more from head rotations than others, no degradation of elevation perception was reported with increases in elevation, which disagrees with Wallach [1939, 1940].

In a more recent study, Morikawa et al. [2013] permitted free head movement and analysed the motions that subjects made in order to make median plane localisation judgments. Seven loudspeakers were arranged on the upper median plane at 30° intervals, presenting broadband, 500 Hz low pass and 12 kHz high pass white noise to subjects. Localisation was tested both under free head movement and head restrained conditions. When no head motion was permitted, the broadband stimulus was localised with 70% accuracy, whilst accuracy for the band-limited stimuli was less than 40%. However, for all stimuli localisation accuracy increased by 20% when head movements were permitted. This result is somewhat expected for the broadband and low pass noise, which each contained low frequencies. It is interesting though that accuracy increased for the high pass noise, which was devoid of frequencies below 12 kHz. This would seemingly indicate that head rotations have some effect in the high frequency range, although it should be noted that this disagrees with Perrett and Noble's [1997] results. The reasons for the differences are, at present, unclear.

1.2.2.3 Shoulder and Torso Reflection Cues

Gardner [1973] conducted median plane localization experiments in which random noise was presented from one of nine loudspeakers arranged in an arc (±18°) on the median plane in front of the listening position. For the test, subject's pinnae were blocked in the following ways; a) no obstruction; b) one pinna cavity blocked; c) both cavities blocked. The results of the study showed good agreement with those that have been discussed previously; localisation judgments were generally not accurate in the absence of either the pinnae or of frequencies above 4 kHz. A further result of note in the study was that around half of the subjects tested were able to accurately localise a 3 kHz ½-octave band source, which is at odds with the discussion in Section 1.2.1, that sound sources must be complex in order to be accurately localised in vertical space. It was therefore hypothesised by Gardner [1973] that additional median plane localisation mechanisms may exist in

the region below 3 kHz, but that generally such cues were overshadowed by the cues provided by the pinnae at high frequencies.

Further experiments relating to Gardner's [1973] hypothesis showed that the error rate of localization judgments for wideband noise that was low passed filtered around 3 kHz was relatively low. Additionally, there was not a notable improvement in the error rate when the pinnae were added or removed, which demonstrated that the localization cues were being provided by something other than the pinnae. Binaural measurements were subsequently conducted in which both a head and a torso were present. These measurements showed that a notch was formed in the spectrum between 0.7 and 2 kHz that was absent when the torso was removed. As a result of this Gardner [1973] concluded that the torso provided additional median plane localization cues in the region between 0.7 and 3.5 kHz.

Gardner's [1973] findings have been rigorously tested in the literature. Measurements made by Kuhn and Guernsey [1983] showed that the torso disturbs incident sound waves at low frequencies, leading Kuhn [1987] to conclude that median plane localization is governed by pinnae cues above 4 kHz and by torso reflections below 3.5 kHz. However, despite these conclusions, Searle et al. [1976] found that reflections from the shoulder, which they called 'shoulder bounce', contribute little overall to the perception of elevation. It can therefore be suggested that, although the torso modifies the spectrum below 3.5 kHz, it does not necessarily provide a useful localization cue.

A later study conducted by Algazi et al. [2001] considered localization mechanisms at low frequencies. In their first experiment, broadband and 3 kHz low pass noise were convolved with the HRTFs of subjects and were subsequently presented through headphones. Azimuths of 0°, -25°, -45° and -65°, both in front of and behind the subject, were simulated, with subjects required to identify the perceived elevation of each source. The test considered twelve elevation angles, which ranged from -45° to 78.75° in 11.25° steps. The results for the broadband source showed that elevation perception was accurate in all cases. For the low pass noise, subjects were still able to perceive source elevation, although elevation judgments were generally

underestimated. Judgments for these stimuli were also more accurate when source presentation was away from the median plane. The authors hypothesized that the perception of elevation in the absence of frequencies above 3 kHz was related to torso reflections. A series of binaural measurements were conducted to test this further. The primary finding was that elevation-dependent notches were measured in the spectrum as low as 700 Hz that were not caused by the pinna reflections. Instead, the reflections were found to come from below the pinnae (i.e. the torso), with delays of at least 0.17 ms being necessary to observe the effect. The authors therefore hypothesized that the torso reflections act as a comb filter, with an inverse relationship between the delay and the frequencies at which the notches occur. Such cues were concluded to benefit median plane localization the least, which agrees with Searle et al. [1976].

A subsequent study by Kirkeby et al. [2007] further considered the role of the torso in localization. In their study, three different HRTFs were simulated by combining a hard head with i) a hard torso; ii) a moderately absorbing torso; iii) no torso. The sound field was tested in 50 Hz intervals in the range of 50 Hz to 24 kHz. When sound sources from above the subject were simulated, reflections from the shoulder were found to arrive at the ear around 0.8 ms after the direct sound and caused ripples in the spectrum up to around 7 kHz. It is interesting to note that this is much higher than the 3.5 kHz maximum indicated by the results of Gardner [1973] and Algazi et al. [2001]. This is perhaps related to Kirkeby et al. [2007] choosing to simulate all measurements, as opposed to the practical measurements made in the aforementioned studies. Their model for the torso, for example, may not have been representative of a real torso and therefore potentially would have yielded less accurate results.

1.2.2.4 Summary of Elevation Cues

From the aforementioned studies, it can be concluded that accurate localisation of sound sources incident from the median plane necessitates the presence of spectral cues provided by the pinnae. In this regard, the key localisation cue is a one octave notch in the spectrum centered in the range between 4 and 10 kHz that

increases in centre frequency with increases in source elevation. In addition to this, certain frequency regions are boosted for a given source direction, which is necessary for determining the directionality of a sound source. Additional median plane localisation cues are provided as a result of head rotations and these are particularly beneficial when low frequency content is present in the source. Moreover, the torso has been shown to modify the spectrum in the range between 700 Hz and 2 kHz. However, the localisation cue provided by this is generally thought to be weak, particularly compared to the cues provided by the pinnae.

1.2.3 The Cues Used to Resolve Front-Back Confusions

As was discussed earlier, a front-back confusion occurs when a sound source incident from in front of a subject is localised behind them and vice versa. Within the literature it has been shown that, much in the same way as elevation perception, there are certain spectral components necessary in a source to limit the number of front-back confusions. In addition, head rotations also function as an effective mechanism. However, the cues provided as a result of shoulder and torso reflections are much weaker [Theile and Spikovski 1982, cited in Algazi et al. 2001]. This was shown experimentally by Kirkeby et al. [2007], who found that, when sources directly in front of or behind the subject were simulated, the spectrum was not found to change significantly between when the torso was and was not present (although some ripples in the spectrum were still observed).

1.2.3.1 The Spectral Content of Sound Sources

With respect to the necessary spectral content in a source to minimize front-back confusions, the literature has reported the following. In the aforementioned study conducted by Asano et al. [1990], it was hypothesised that the numerous peaks and dips in the spectrum, as provided by the pinnae, might not all be necessary. Instead, it was thought that the human auditory system looked for more general trends in the spectrum. This was tested by recording the HRTFs of subjects at various locations on the median plane and

convolving them with sound sources in order to give the impression of median plane incidence. Localisation tests were subsequently conducted whereby the rate of front-back confusions was analysed as the convolved spectra were progressively simplified. The results of the experiment revealed the following. Firstly, when the HRTF was simplified across the full spectrum, increased levels of simplification resulted in increases in the rate of front-back confusions. Furthermore, when the HRTFs were only simplified below a specific boundary frequency (3, 2, 1 and 0.5 kHz) it was found that 2 kHz was a 'critical boundary frequency for front-rear judgments'. Asano et al.'s [1990] results were discussed in light of a previous study conducted by Wettschurek [1973, cited in Asano et al. 1990]. In that study, it was found that the rate of front-back confusions increased when high frequencies were absent from the test stimuli. With this result in mind, Asano et al. [1990] concluded the following with respect to front-back confusions. Firstly, the spectral cues to resolve them are microscopic peaks and dips in the region below 2 kHz. In addition to this, high frequencies also have some importance, although in this case macroscopic patterns are utilised.

Perrett and Noble [1997] produced results that supported Asano et al.'s [1990] conclusions regarding the spectral cues used to resolve front-back confusions. In localisation experiments conducted using low and high pass (1, 2 and 4 kHz) and broadband noise, the following results were found in the case that no head motion was permitted and the pinnae were not obstructed (Fig. 1.4). Firstly, for the low pass noise the number of front-back confusions was highest when the cut-off frequency was 1 kHz (38% of all judgments). This fell to 30% when the cut-off increased to 4 kHz. Equally, for the high-pass noise the error rate rose from 14% for 1 kHz to 25 % for 4 kHz. These results agree with Asano et al.'s [1990] in that the sources whose spectral content included both high frequencies and the region around 2 kHz were localised with fewer front-back confusions compared to those in which either of these were absent.

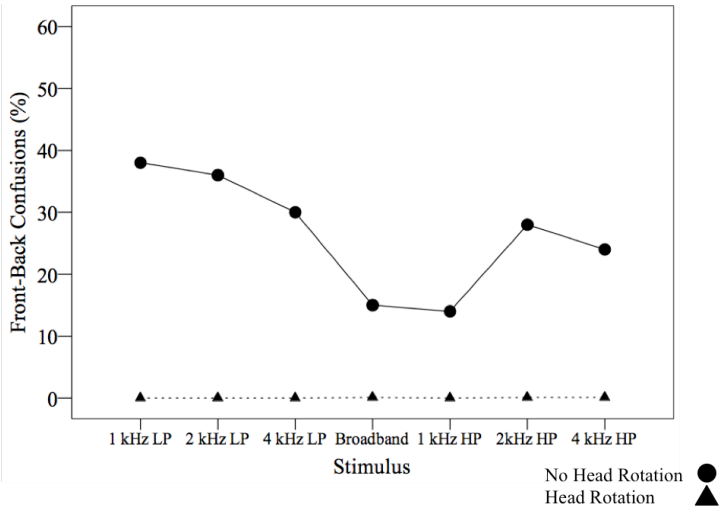

Fig. 1.4: Results from Perrett and Noble [1997] showing the effect of both frequency and head rotations on the number of front-back confusions [after Perrett and Noble 1997].

1.2.3.2 Head Rotations

Blauert [1997] stated that front-back confusions 'almost never occur' in situations whereby the subject is able to move their head freely. This has been demonstrated experimentally in numerous studies including in the aforementioned experiments conducted by Perrett and Noble [1997], whose study showed that the rate of front-back confusions fell from up to 38% (1 kHz low-pass) to no greater than 0.6% in the case that head movements were permitted (Fig. 1.4). In addition, Wightmann and Kistler [1999] presented white noise to subjects from loudspeakers arranged in a vertical arc, with the effect of head movements on front-back confusions at varying azimuths being tested. When no head movement was permitted, three out of the seven listeners recorded a high number of front-back confusions, whilst two recorded very little; this suggesting that the number of front-back confusions depends on the subject. However, when head movements were permitted few front-back confusions were recorded, which agreed with both Blauert [1997] and Perrett and Noble [1997]. In addition, Wightmann and Kistler [1999] also found that the dynamic cues provided by head rotations are only effective if they are under the control of the subject; moving the source around the subject had little effect on the resolution of front-back confusions unless the subject was aware of the direction of source movement. Also, as a greater number of front-back confusions were reported with respect to the

number of up-down confusions, it was hypothesised that the spectral filtering of the pinna is more effective for elevation perception than it is for the resolution of front-back confusions.

A subsequent study conducted by Iwaya et al. [2003] further demonstrated the effect of head rotation on the resolution of front-back confusions. As with Wightmann and Kistler [1999] this study was not confined to the median plane, with loudspeakers being arranged at every 30° on the horizontal plane. 1 kHz low pass, 3 kHz high pass and broadband pink noise were used as the test stimuli. The study considered not only the effect of head rotation on the resolution of front-back confusions but also the effect of signal duration, with 0.5 and 3 s being considered. Subjects were tested with free and restricted head motion conditions. The results of the experiment showed that front-back confusions were greatest for the low pass pink noise. This result agrees closely with the conclusions of both Asano et al. [1990] and Perrett and Noble [1997] with respect to the importance of both the 2 kHz range and high frequency content. In addition, as with the aforementioned studies, the number of front-back confusions was reduced when head movements were permitted. Furthermore, signal duration was found to have an effect, with the number of front-back confusions for the 3 s stimuli being notably reduced compared to the 0.5 s stimuli. This result indicates that the effects of head rotation on the resolution of front-back confusions are greater when the source is continuous in nature.

1.3 THE FREQUENCY DEPENDENCY OF MEDIAN PLANE LOCALISATION

It has been discussed thus far that sound sources incident on the median plane must be complex in order for localisation to be accurate. From this, it can be deduced that non-complex (i.e. tonal and band-limited) sources are localised independently of the position of the emitting loudspeaker when incident from the median plane. Under such conditions, numerous studies have reported that localisation judgments are made according to the frequency of the stimulus. Two such observed phenomena are discussed here, ‘directional bands’ and the ‘pitch-height effect’.

1.3.1 Directional Bands

1.3.1.1 The Parameters for Directional Band-Like Localisation

Blauert [1969] presented 1/3-octave band pulses to subjects in a darkened anechoic chamber from three loudspeakers arranged on the upper hemisphere of the median plane (directly in front of, above and behind the subject). The noise pulses were between 200 ms and 1 s in length, with the amplitude of stimulus presentation being between 30 and 60 dB (10 dB steps). Subjects were required to identify the location of each stimulus using a scale that divided the median plane into three regions, 'front' (345° to 45°), 'above' (45° to 135°) and 'behind' (135° to 195°). A finer division of the scale was opted against in order that the test did not become too time consuming or difficult for subjects. The experimental data obtained in the experiment showed that, rather than being related to the position of the emitting loudspeaker, localisation judgments were made on the basis of frequency, with specific frequency 'bands' being related to specific regions on the median plane. These bands were referred to as 'directional bands' (Fig. 1.5). A directional band was defined as having occurred if more than half of subjects gave an answer in a specific region more often than they gave both of the remaining answers combined. The key directional bands identified in the study were as follows. The frequency ranges between 250 and 500 Hz were associated with localisation in front of the subject, 0.7 – 2 kHz behind and 8 kHz above. Interestingly, Blauert [1969] also concluded that there existed a relationship between 4 – 7 kHz and localisation in front of the subject. Although it can be seen in Fig. 1.5 that the relationship is strong for 4 kHz is strong (>72%) is is notable that the relationship at 7 kHz is much reduced, being below 52%. It is therefore unclear why Blauert [1969] concluded that the relationship between the 4 Hz region and the front extended as far as 7 kHz.

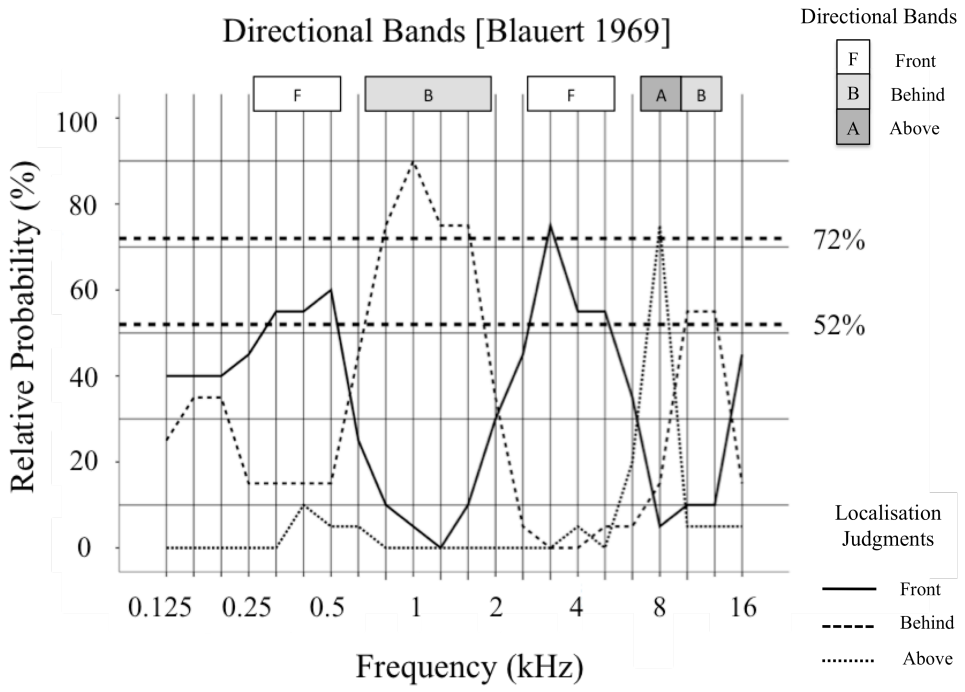

Fig. 1.5: The results of Blauert's [1969] directional band experiment, showing the relationship between the centre frequency of 1/3-octave bands and their perceived location on the median plane [after Blauert 1969].

Morimoto and Aokata [1984] found that directional band-like localisation was not confined solely to the median plane and in fact was prevalent in planes parallel to it. Loudspeakers were positioned by combining three lateral angles (30°, 60° and 90°), with three rising (elevation) angles (0°, 90° and 180°) on the right quadrant of the upper hemisphere. Loudspeakers were each 1.5 m away from the listening position. The stimuli were six 1/3-octave band noise (1 s duration), with centre frequencies ranging from 1-10 kHz, and were presented at 50 dB. It was found that, with respect to lateral angle, localisation was generally accurate for all stimuli. With respect to rising angle, the judgments followed directional-band like localisation, although the overall effect was not as strong as that reported by Blauert [1969]. With respect to this result, Morimoto and Aokata [1984] suggested that directional bands might have operated less strongly in their experiment due either to the number of subjects tested or the test methodology utilised. However, they did not consider the different localisation mechanisms available to subjects when test stimuli are incident from the median plane compared to when they are presented from any other region around a subject. As was

discussed in Section 1.1.1, localisation generally relies on a combination of binaural and HRTF cues, with the latter being the only cue available for localisation when sources are incident from the median plane. It could be argued then that the presence of binaural cues served to increase localisation accuracy, which would have reduced the strength of the directional band relationship, Despite this, it is apparent from the results in the Morimoto and Aokata [1984] study that localisation accuracy in the presence of binaural cues was still low and therefore the effect of binaural cues on directional bands would require further study. Nevertheless, the results do at least demonstrate that directional band-like localisation is maintained away from the median plane.

A further study conducted by Itoh et al. [2007] considered the effect of bandwidth on directional band-like localisation. 1/3 and 1/6-octave band white noise (800 Hz-12.5 kHz) was presented to subjects from one of three loudspeakers located directly in front of (0°), behind (180°) and above (90°) the listening position. Stimuli were presented at 60 dBA. Subjects were required to plot the perceived elevation of each source using computer software. As with Blauert [1969], for the purposes of statistical analysis the median plane was divided into 'front', 'above', and 'behind', with binomial tests being conducted to identify directional bands. Fewer directional bands were reported for this study than were reported by Blauert [1969], however there remained a number of similarities between the results of the two studies. Itoh et al. [2007] reported that 0.8-1.6 kHz was related to behind localisation, 2-5 kHz to the front and 6.3-8kHz to above. With respect to the effect of bandwidth, the experimental data obtained showed no significant differences in localisation judgments between the 1/3 and 1/6-octave bands. In addition, individual differences were found for directional bands indicating that they are not consistent for all subjects.

A more recent study conducted by the author [Appendix A] also considered the effect of bandwidth on the perception of directional bands. Octave bands, 1/3-octave bands and tonal stimuli with centre frequencies of 0.5, 1, 4 and 8 kHz were presented either continuously or in 200 ms bursts to subjects. Stimulus presentation was from one of two loudspeakers; one located directly in front of the listening position, with the other directly behind. Subjects were provided with a scale that split the median plane into 8 identical regions (Fig.

1.6). This scale was designed to be a more refined version of that used by Blauert [1969] and was used with the intention of trying to identify location differences between directional bands that had previously been considered to be in the same position as one another (e.g. 500 Hz and 4 kHz). The results of the study showed that bandwidth had a notable influence on localisation judgments. For example, the 8 kHz bursts showed directional band-like localisation for the tonal and 1/3-octave band stimuli. Conversely, the 1 kHz bursts showed directional band-like localisation for the octave bands and not the tones. In addition, the directional band effect was generally found to operate most strongly for the 1/3-octave bands compared to the tonal and octave band stimuli. The results of this study therefore suggested that not only is directional band-like localisation dependent on bandwidth but also that the bandwidth depends on the centre frequency of the stimulus. Another interesting point of note was that, generally, the 4 kHz stimuli were perceived as being elevated with respect to the 0.5 kHz stimuli. This therefore showed evidence of the pitch-height effect, which is discussed in detail in Section 1.3.2. Blauert [1969] was unable to make such a connection, arguably due to the use of a scale that was less refined.

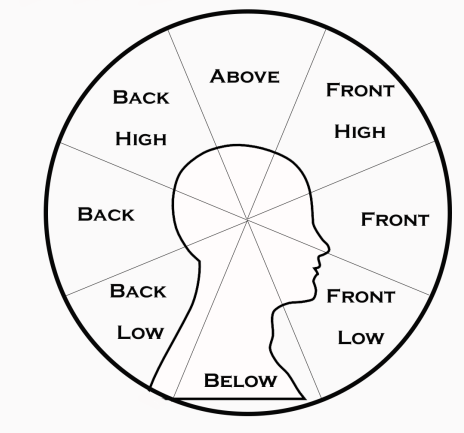

Fig. 1.6: Scale used for the author's study on directional bands [Appendix A].

Studies considering directional bands have not been confined to simply studying the localisation of band-limited stimuli. A study conducted by Chun et al. [2011] demonstrated that the manipulation of directional bands in a complex signal could alter the perceived elevation of the sound source in the median plane.

Speech and musical sources were presented to subjects from stereophonic loudspeakers in the horizontal plane (azimuth ±30°), with the resultant sound image being formed on the median plane. The stimuli underwent both HRTF and spectral notch filtering, as well as directional band boosting. The experimental data showed that when directional band boosting alone was applied in the 4.5-9 kHz region (i.e. in the region in which the spectral cues for elevation lie) the resultant signals were perceived as being elevated by up to 20° with respect to the loudspeaker position. The results of this experiment have important implications for vertical interchannel crosstalk. If it is the case that the boosting of directional bands can cause the perceived elevation of sound sources to increase then, hypothetically, the opposite (i.e. directional band reduction) could cause perceived source elevation to decrease. It could therefore be argued that vertical interchannel crosstalk would not affect the perceived elevation of the main channel signal if the energy in the 4.5-9 kHz region was reduced for the direct sound in the height channel. This is considered as part of the band reduction method in the present thesis, particularly in Chapters Three and Four.

1.3.1.2 The Relationship between Directional Bands and Spectral Cues

Following the identification of directional bands, Blauert [1969] conducted objective measurements of the transfer function of the ear in order that the phenomenon might be explained. The measurements considered the difference in the ear input spectra for sound incident from the front compared to sound incident from the rear (Fig. 1.7). From the measurements, it was identified that certain frequency regions (250-500 Hz, 4-7 kHz) were 'boosted' for sound incident from the front, whilst others (0.7-2 kHz) were boosted for the rear. These 'boosted bands' were found to agree closely with the directional band results. An additional boosted band was found at 8 kHz for sound incident from above. Blauert [1969] concluded that, with respect to sound localisation in the median plane, judgments might be resolved based on the directional band in which the signal has the most power.

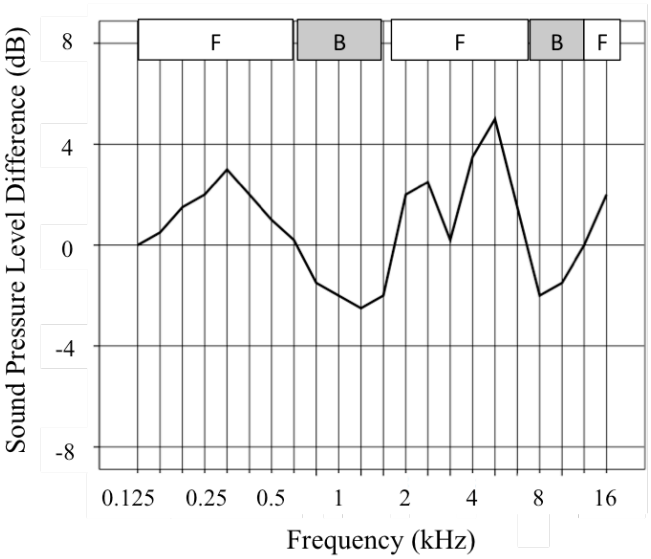

Fig. 1.7: Ear input spectra for sound incident from the front minus the spectra for sound incident for the rear, showing Blauert's boosted bands [after Blauert 1969].

Both Blauert's [1969] 'boosted band' hypothesis and his objective measurements of the external ear's transfer function show good agreement with the later studies of Hebrank and Wright [1974a] and Asano et al. [1990], who measured the directional-dependence of the external ear for elevated sound sources in the median plane. Hebrank and Wright [1974a] showed directional dependencies in the spectra including a ¼-octave notch between 7 and 9 kHz for above perception, whilst Asano et al. [1990] found that, in general, the power at 0.5 and 3-5 kHz increased with frontal incidence, whilst 1-2 kHz increased for the rear.

The results of the aforementioned studies are valuable in ascertaining the reasons for the directional band phenomenon. It is clear that median plane localisation relies on the direction-dependent filtering of the external ear and further that there exists a cognitive relationship between frequency and direction of incidence (more specifically source elevation). It would therefore seem that the centre frequency of band-limited sources is taken as being a localisation cue, therefore rendering the position of the emitting loudspeaker irrelevant. Arguably then, directional band-like localisation can be considered as being a by-product of the external ear transfer function. It should also be noted that the maximum bandwidth for directional band-like localisation is open to further investigation. It is known from the studies of Roffler and Butler [1968a], for example, that localisation judgments are generally more accurate for sound sources that

are ‘complex’. This would therefore indicate that there is a threshold bandwidth whereby directional band-like localisation breaks down and the position of the emitting loudspeaker becomes dominant in determining the perceived location of the sound source. At present no study, of which the author is aware, has considered this issue. However, given that the bandwidth of HRTF cues is dependent on frequency [Hebrank and Wright 1974a] (i.e. a 1-octave notch in the 4-10 kHz region relates to front localisation, whilst the above cue is a ¼-octave peak between 7 and 9 kHz) it can at least be suggested that the bandwidth at which directional bands break down is frequency-dependent. This agrees with the data reported by the author [Appendix A].

1.3.2 The Pitch-Height Effect

1.3.2.1 The Effect for Tonal Stimuli

In this section, ‘tonal stimuli’ refers to the use of sine waves of single frequencies as the test stimuli. In an experiment conducted by Pratt [1930] tones of frequency 256, 512, 1024, 2048 and 4096 Hz were presented to subjects from a telephone receiver located behind a numbered scale, directly in front of the listening position. The position of the receiver could be moved between five different vertical locations. Subjects were required to use the scale to identify the perceived location of each test stimulus. The experimental data (Fig. 1.8) showed that, for every observer, localisation judgments were made on the basis of frequency, with an increase in frequency correlating with an increase in perceived source elevation. In subsequent years, this phenomenon has been heavily scrutinized and has been described both as the ‘pitch-height effect’ [Cabrera and Tiley 2003] and ‘Pratt’s effect’ [Cabrera and Morimoto 2007].

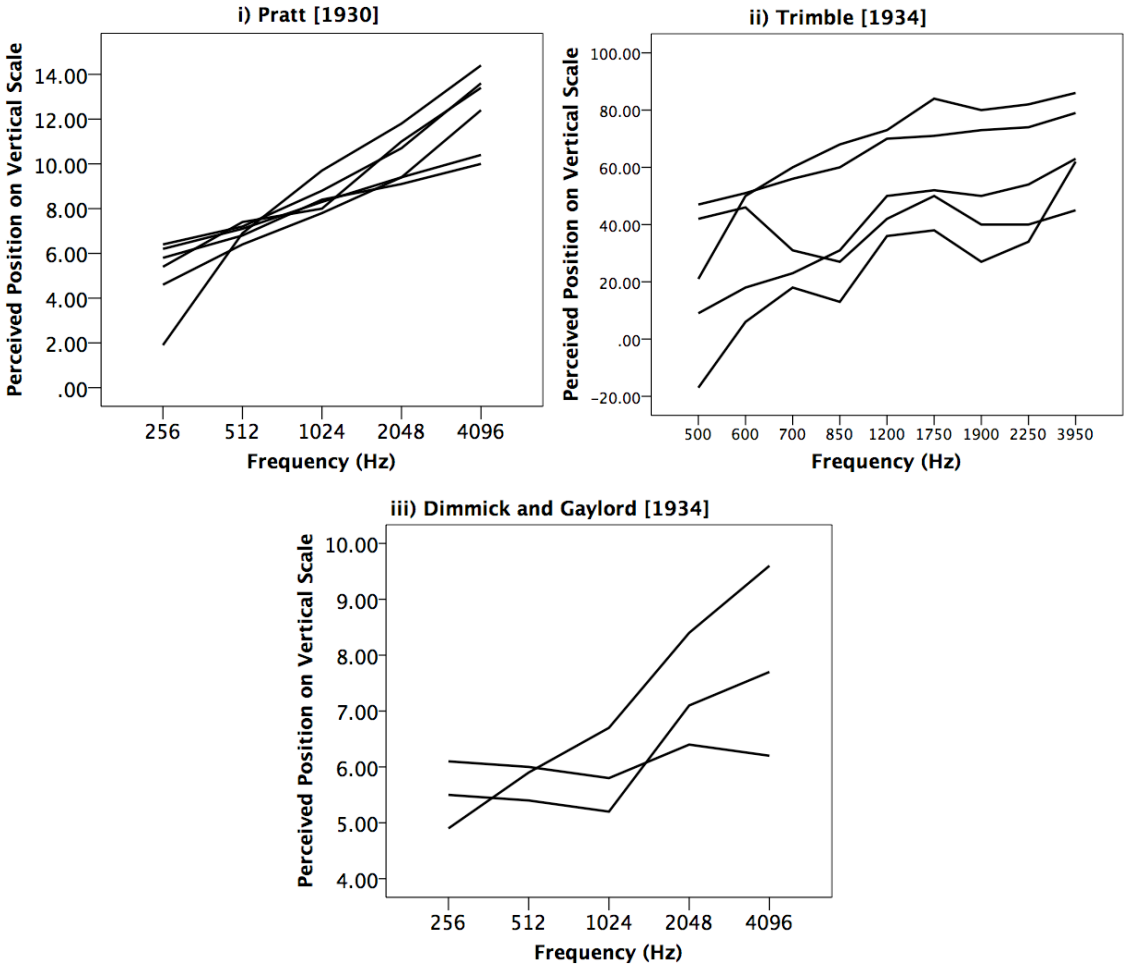

Fig. 1.8: Results of pitch-height experiments for tonal stimuli from the studies of i) Pratt [1930], ii) Trimble [1934] and iii) Dimmick and Gaylord [1934] [after Pratt 1930, Trimble 1934, Dimmick and Gaylord 1934]. In each case, the graphs show localisation judgments for tonal stimuli on scales that were located in front of the listening position, with the results of i) and ii) in particular showing a relationship between frequency and height (the 'pitch-height effect'). Note that each study utilised a different scale (evidenced by the differing y-axes), making a direct comparison between results difficult.

Trimble [1934] further analysed Pratt's [1930] conclusions, presenting nine tonal stimuli to subjects from receiving phones located 15 cm from their ears. The frequency of the test stimuli ranged from 500 to 3950 Hz. Subjects were required to identify the angular displacement of each tone from the horizontal plane. The results showed that, for most subjects, the high tones were perceived as being elevated with respect to the low tones. Trimble [1934] suggested that vertical localisation 'may be a function of wave-form of phase in

some dimension other than the sine dimension'. This was due to the fact that the low tones in the experiment featured greater complexity than the high tones. In further experiments, Trimble [1934] presented tonal stimuli, with frequencies ranging from 1200-3950 Hz, as a series, in both ascending and descending order, and asked subjects to identify the perceived displacement from the horizontal at the start and the end of the sequence. The majority of observers perceived the movement of sound in an arc as the frequency changed, again with increases in frequency resulting in increases in perceived height. Trimble [1934] described this as the 'arc-effect', concluding that it operates in the order of phase, time and intensity effects seen in horizontal localisation.

A further study conducted by Dimmick and Gaylord [1934], that directly replicated Pratt's [1930] experiment, was unable to obtain similar results. Instead, the relationship between pitch and height was only observed for one subject. Differences in the experimental method and instructions given to subjects were postulated by Dimmick and Gaylord [1934] as being the reason for the conflict in data. However, they did not consider inter-subject differences for the operation of the effect. Between the 14 subjects tested amongst the three experiments conducted by Pratt [1930], Trimble [1934] and Dimmick and Gaylord [1934], three were reported as not perceiving a correlation between pitch and height, two for Dimmick and Gaylord [1934] and one for Trimble [1934]. Based on this, it can be argued the pitch-height effect is not perceived for all subjects. Had Dimmick and Gaylord [1934] used a larger number of subjects then it might have been that they would have been more likely to obtain results more in-line with the pitch-height effect.

Roffler and Butler [1968b] undertook a more thorough analysis of the pitch-height effect for tonal stimuli. Their study was concerned, not only with replicating the effect, but also with determining whether or not it was maintained when cues associated with high and low were removed or manipulated. In initial studies, the correlation between pitch and height was observed for tonal stimuli ranging from 250-7200 Hz. It should be noted, however, that one subject did not perceive the systematic spatial arrangement of stimuli, with all stimuli appearing to originate from a narrow region. This agrees with the present hypothesis, that the pitch-height effect is not prevalent for all subjects. Further studies were conducted as follows. Firstly, the effect of

source distance was considered. Under these conditions, the frequencies were found to originate from specific sections of the panel that was used for localisation judgments, irrelevant of distance. The authors therefore concluded that judgments were based on the subject's frame of reference. Additional experiments were conducted in which the effects of listener orientation, visual bias and knowledge of the terms 'high' and 'low' in describing musical pitch were analysed. For each condition, the correlation between pitch and height was maintained. Roffler and Butler [1968b] therefore concluded that the data they obtained supported the conclusions made by Pratt [1930] regarding the existence of an intrinsic spatial character in tonal stimuli.

1.3.2.2 The Effect for Band-Limited Stimuli

In more recent years, it has been demonstrated that the pitch-height effect is not confined to tonal stimuli. Cabrera and Tiley [2003], for example, showed that the effect is maintained for octave bands, whilst Ferguson and Cabrera [2005] obtained similar results when simultaneously presenting high and low frequency band-limited pink noise to subjects. Furthermore, Cabrera and Morimoto [2007] reported that the pitch-height effect is maintained when complex tones (fundamental frequencies ranging from 55-1244.4 Hz) are presented to subjects both from the median and lateral planes. This latter result shows that the pitch-height effect is maintained when binaural cues are present, much in the same way that Morimoto and Aokata [1984] demonstrated that directional band-like localisation is preserved away from the median plane.

An interesting result was reported in a recent study conducted by Lee [2016]. Octave band stimuli (63 Hz-16 kHz) were presented to subjects from pairs of non-elevated (0°) and elevated (30°) left and right loudspeakers (60° base angle). For both conditions, the resultant phantom image was formed on the median plane. The loudspeakers were obscured by an acoustically transparent curtain, which featured a vertically oriented scale; subjects were to use the scale to make their localisation judgments. The experimental data revealed a 'double' pitch-height effect that formed separately in the regions between 63-500 Hz and 1-8 kHz (Fig. 1.9). That this was not reported in the aforementioned studies may be related to differences in the experimental setup, with Lee [2016] presenting his stimuli as horizontally arranged phantom images, rather

than from fixed sources. It is known, for example, that presentation using the former configuration results in the phantom image elevation effect [De Boer 1947], which is considered in Section 1.4. How this would influence the localisation of narrowband and tonal stimuli is less clear and requires further study.

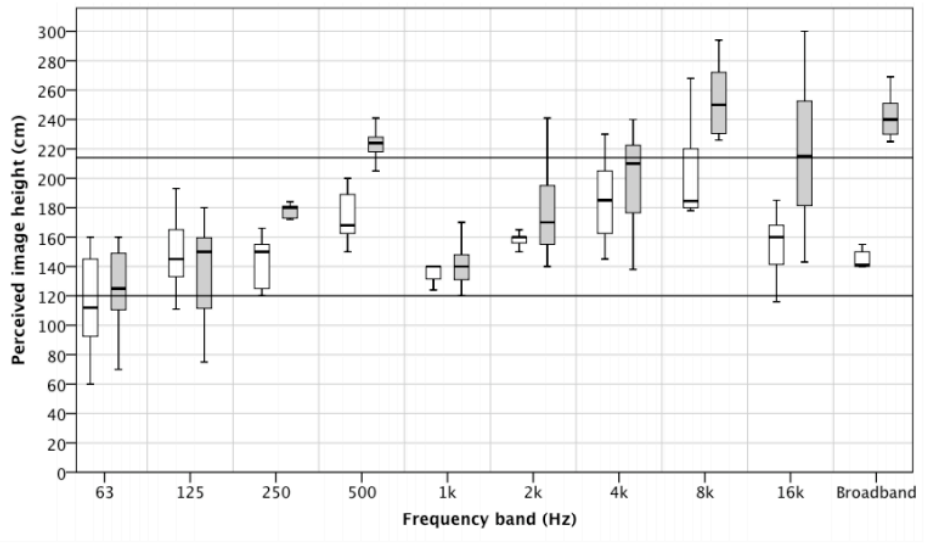

Fig. 1.9: The results of localisation experiments conducted by Lee [2016] showing a double pitch height effect for octave band stimuli. The white and grey boxes are for source presentation from the main and height layers respectively [courtesy of Lee 2016].

From the aforementioned literature, it is clear that there exists a relationship between pitch and height that is maintained both for tonal and band-limited stimuli. Furthermore, as with directional bands, the effect is not confined solely to the median plane. It is unclear at present how this would affect localisation thresholds across the frequency spectrum. Hypothetically more level reduction might be necessary in the high frequency region as perceived source elevation would be greater than for the low frequency stimuli (i.e. the sources would be further away from the main channel layer). This remains to be seen, however, as localisation thresholds across the frequency spectrum have not yet been considered. This is approached primarily in Chapters Three and Four of the present thesis.

1.3.2.3 Cognitive Explanations for the Pitch-Height Effect

Along with demonstrating that the pitch-height effect is maintained for octave bands, Cabrera and Tiley [2003] hypothesised the reasons for the phenomenon. With respect to low frequencies, they proposed a relationship with floor reflections. They argued that, in normal listening conditions, in which a reflective floor is present, the spectral notches caused by comb filtering between the direct and reflected sounds would extend to low frequencies for elevated sources. As a result, more bass energy would be present as proximity to the floor decreased. It was considered that this floor reflection cue could be learned in the same way as median plane pinna cues, hence the relationship between low frequencies and low elevation. Despite this, no explanation was offered for the pitch-height effect at high frequencies.

The Cabrera and Tiley [2003] hypothesis for the pitch-height effect at low frequencies arguably has some plausibility. However, there seems to be some conflict with objective measurements made by Algazi et al. [2001]. In their study it was shown that the torso introduces notches in the frequency spectrum and further that the centre frequency of such notches *decreases* with increases in the elevation of the source. This seemingly indicates that there exists a cognitive relationship between a decrease in frequency in the low frequency region and an increase in elevation. However, a key limitation with this measurement is that it offers no explanation of the pitch-height at low frequencies. Despite this, the torso reflection cues demonstrated by Algazi et al. [2001] do not necessarily mean that the Cabrera and Tiley [2003] hypothesis should be rejected. Instead, it is entirely plausible that both relationships exist and further that the floor reflection cue is more dominant than the torso cue, which incidentally has been shown in the literature to be weak in the median plane [Gardner 1973, Searle et. al. 1976]. This would require further study, however seems somewhat reasonable in line with the results of the respective studies.

Explanations for the pitch-height effect at high frequencies are somewhat simpler. From the aforementioned studies considering the spectral cues for median plane localisation, it is clear that there exists a relationship between the region between around 4 and 10 kHz and elevation perception [Shaw and Teranishi 1968,

Hebrank and Wright 1974a, Asano et al. 1990]. Moreover, increases in centre frequency in this region leads to the perception of increased elevation [Hebrank and Wright, 1974a]. With respect to the pitch-height effect at high frequencies then, it is arguable that it is this association that causes the phenomenon, much in the same way that directional bands are likely related to spectral cues. This seems reasonable given that Roffler and Butler [1968b] found that the effect was maintained despite a number of variations in the test conditions. This would require further study.

That both directional bands and the pitch-height effect may both be perceptually related to spectral cues was previously postulated by the author [Appendix A]. In that study not only were directional bands identified, so too was evidence of the pitch-height effect. Based on the results, it was hypothesised that the pitch-height effect and directional bands were both part of the same localisation mechanism. The difference in the experimental data obtained between pitch-height and directional band studies was explained based on differences in experimental context. In particular, the response methods used in the respective studies were thought to have some effect. This latter point was based on the results of a study conducted by Perrett and Noble [1995], in which the results of localisation studies were found to vary when the same stimuli were localised using different response methods. Although this would require further study, it seems plausible based on the experimental data reported in Appendix A.

1.4 THE PHANTOM IMAGE ELEVATION EFFECT

When stereophonic loudspeakers radiate a coherent sound source the resultant phantom image is formed on the median plane [Blauert 1997]. However, it has been demonstrated in the literature that the perceived elevation of the phantom image is a function of the loudspeaker base angle; the greater the base angle the greater the perceived elevation of the phantom image (Fig. 1.10). This effect has become known as the ‘phantom image elevation effect’ [Lee 2017]. The effect was first demonstrated in experiments conducted by De Boer [1947]. Stereophonic loudspeakers were positioned on the horizontal plane and emitted a coherent

signal, with the resultant phantom image being formed on the median plane. Subjects were required to identify the perceived elevation of the phantom image in the case that the distance between the loudspeaker pair and the subject varied. Although the results of the study were found to be inconsistent between subjects, there was a general trend that perceived source elevation increased as the subjects approached the loudspeaker pair. It should also be noted that a decrease in distance between each subject and the loudspeaker pair corresponded to an increase in base angle between the loudspeakers with respect to the subject. In addition, the increase in elevation with decreases in distance was not linear; changes were small initially, with elevation increasing more rapidly the closer subjects were to the loudspeaker pair. Furthermore, when the subjects were directly between the loudspeakers the phantom image was perceived as being behind them.

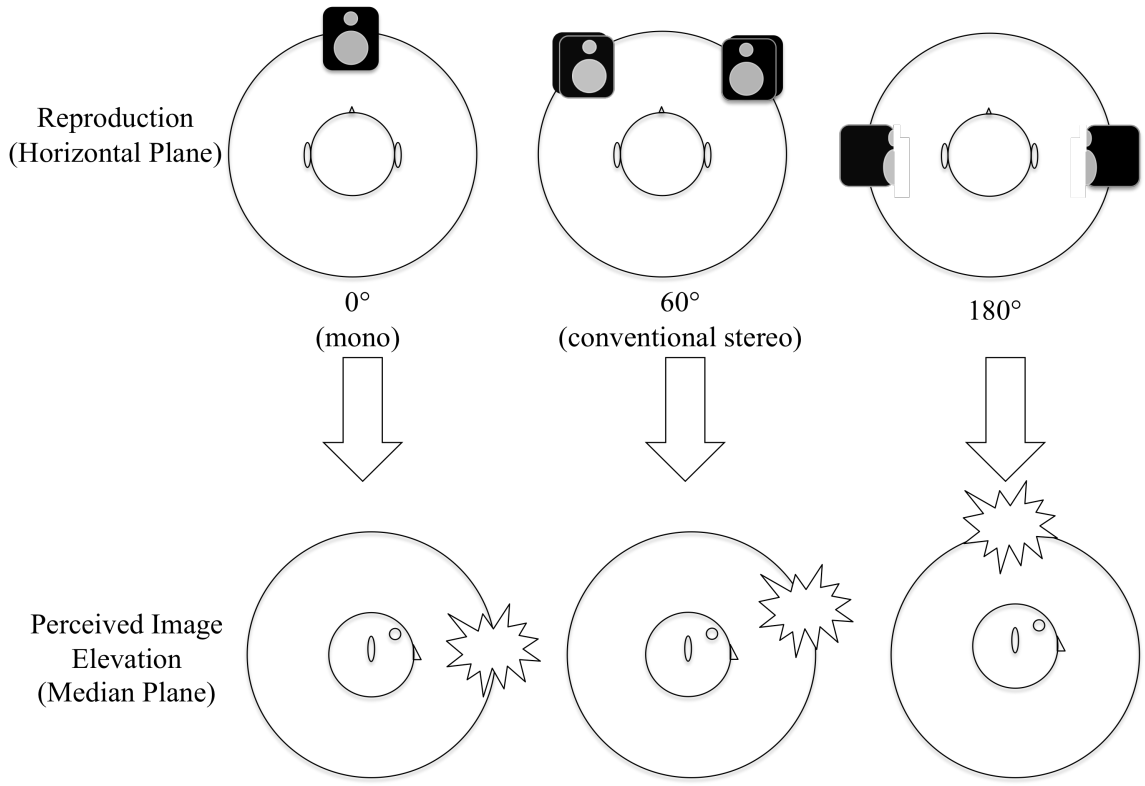

Fig. 1.10: Approximate relationship between loudspeaker base angle and perceived image elevation (the phantom image elevation effect). Based on the results of Lee [2017].

More recently, the phantom image elevation effect has been demonstrated in studies conducted by Jo et al. [2010] and Lee [2017]. In the former study, the effect was considered in the context of 3D audio with broadband white noise (1-16 kHz) being used as the test stimuli. A key result in the study was that little or no source elevation was observed when stimuli were presented from the centre loudspeaker only. However, when stereophonic left and right loudspeakers (±30° azimuth) were added to the centre loudspeaker the resultant phantom images were elevated by up to 10° with respect to the horizontal plane. Moreover, elevation of up to 30° was perceived when the aforementioned configuration also included surround left and right loudspeakers (±110° azimuth). However, it should be noted that this effect was not reported for all subjects, which somewhat agrees with the inconsistent results reported by De Boer [1947]. The results obtained in the study were also found to be consistent when the experiment was conducted in rooms with varying acoustic properties.

Lee [2017] demonstrated that the phantom image elevation effect was source dependent. The stimuli used in his study were 'natural' sources (aeroplane, helicopter, rain, thunder, bird, church bell and male speech) and noise (broadband pink and white noise presented continuously and as bursts, with 10 s and 200 ms duration respectively). The loudspeaker base angles tested in the experiment ranged from 0° to 360°. The most linear phantom image elevation effect was observed for sources with a broad and flat frequency range (white noise, rain), whilst perceived elevation was lessened for sources with low frequency weight or a roll-off in the high frequency range (pink noise, speech, thunder, aeroplane, helicopter). This led Lee [2017] to conclude that 'the broadness and flatness of source spectrum is an important factor for the perception of the phantom image elevation effect'.

Attempts to explain the phantom image elevation effect were initially made by De Boer [1947]. He argued that the brain uses variations in ILD as a fixed plane of reference from which elevation judgments can be made. When the source is presented in the median plane, it is apparent that ILD cues are absent. However, they can be introduced by rotating the head. In this case, the degree of ILD is proportional to the amount of rotation. However, as source elevation increases the ILD for a given head rotation decreases. De Boer [1947]

also noted that the ILD for a given head rotation is less for a phantom image than for the same source presented at the same location from a single loudspeaker. This served as the basis of his hypothesis, which essentially considered the lesser ILD as a result of the source being presented as a phantom image to be misinterpreted as being caused by a real source that was more elevated with respect to the horizontal plane. Thus, a given source presented as a phantom image would be perceived as being more elevated with respect to the same source presented from a single loudspeaker. De Boer [1947] subsequently developed a model in order to predict the phantom image elevation effect, which also considered the effect of increasing the loudspeaker base angle, however this was found to be accurate only when the distance between the subjects and the loudspeaker pair was small.

It is clear that De Boer's [1947] hypothesis was not wholly sufficient. The primary flaw is that it relied on the presence of ILD as a result of head rotations in the case that the phantom image is formed on the median plane. Based on this hypothesis, it could be deduced that if ILD cues were not present (i.e. if the head was not moved) then the phantom image elevation effect would not be observed. However, subsequent research has shown that this is not the case, with Lee [2017] demonstrating that the phantom image elevation effect still operates when no head movements are permitted. It should also be noted that De Boer [1947] did not consider spectral cues when attempting to explain the effect. This is likely because the relationship between spectral cues and source elevation had not yet been explained in the literature, with key studies on the subject such as those by Shaw and Teranishi [1968] and Hebrank and Wright [1974a] coming notably later.

An initial attempt to explain the phantom image elevation effect from the perspective of spectral cues was made by Lee [2017]. Objective measurements showed that energy in the 8 kHz region, which Blauert's directional band hypothesis [1969] related to above perception, increased with increases in loudspeaker base angle up to 240°. Although this closely matched the experimental data in terms of perceived elevation, it did not explain the front-back biases observed with changes in loudspeaker base angle. This therefore suggested that a hypothesis based on the spectral filtering of the pinnae would be inadequate. Instead, Lee [2017] considered the following two mechanisms in order to explain his results. Firstly, when an elevated real

source radiates sound in the median plane the ear receives a shoulder reflection sometime after the arrival of the direct sound. Additionally, as no interaural cues are available when stereophonic loudspeakers radiate a coherent signal on the median plane, the brain might use the acoustic crosstalk delay between the signal from a given loudspeaker arriving at each ear to determine whether the source is real or phantom. Lee [2017] hypothesised that the phantom image elevation effect was essentially the result of the acoustic crosstalk delay being mistaken by the brain for delay between the direct sound and shoulder reflection. This seems reasonable as the delay time between the direct sound and shoulder reflection would increase with increases in source elevation in the same way that the acoustic crosstalk delay would increase with increases in loudspeaker base angle up to 180°.

In relation to vertical interchannel crosstalk the phantom image elevation effect may have some interesting implications. Based on the results of the Jo et al. [2010] study, any direct sounds equally present in the L, C and R loudspeakers would perceptually be elevated by as much as 10° with respect to the horizontal plane. This means that the main channel signal would be in an elevated position with respect to the main layer of loudspeakers, even without any interference from vertical interchannel crosstalk. The effect that the crosstalk signal would then have on the perceived elevation of the main channel signal is unclear. Unfortunately this was not one of the loudspeaker configurations considered in the Jo et al. [2010] study and no other study, of which the author is aware, has approached the issue. If it is the case, for example, that the phantom image elevation effect causes vertical interchannel crosstalk to have less of an influence on the perceived elevation of the main channel signal then it may be either that localisation thresholds would vary or even that the application of localisation thresholds would not be not entirely necessary. A further issue in this regard is that the phantom image elevation effect is source dependent [Lee 2017]. This may result in different thresholds needing to be applied to different sources within the same ensemble, which would be difficult to execute in a practical recording environment. Clearly then, the phantom image elevation effect needs to be understood in the context of vertical interchannel crosstalk, especially in light of the aims of the present thesis.

1.5 SUMMARY

This chapter discussed the mechanisms utilised in the localisation of sound sources, with a particular focus on localisation in the median plane. Firstly, the general mechanisms used in human sound localisation were discussed. Following this, the mechanisms used for sound localisation in the median plane were covered in detail. The frequency dependency of median plane localisation was also given consideration, which included a review of the pitch-height effect and directional band phenomena. Their relationship with median plane localisation cues was also discussed. The chapter concluded with a discussion of the phantom-image elevation effect.

To summarize, the primary mechanisms used in sound localisation are interaural and spectral cues. Interaural cues are the result of differences in the source arriving at each ear and include ITD, for the localisation of frequencies below around 1.5 kHz, and ILD, for the localisation of frequencies above around 3 kHz. Spectral cues, also known as HRTF cues, are a result of the directional-dependent filtering of the external ears, or pinnae. When sound sources are incident from the median plane, the time and level differences between the ears are equal and, as such, spectral cues are the sole mechanism by which localisation is achieved. Studies have shown that accurate median plane localisation is not possible in conditions whereby the pinnae cues are absent. Additionally, accurate vertical localisation necessitates spectral complexity of sound sources as well as the presence of frequencies above around 4 kHz.

With respect to the spectral cues utilised for median plane localisation, there exists a relationship between certain frequency regions and certain directions. The elevation cue, for example, is a one-octave notch between 4 and 10 kHz; increasing the notch centre frequency gives the perception of increased elevation. Additional cues are also provided as a result of head rotations and shoulder and torso reflections, although it should be noted that the latter cue is somewhat weak. Each of the aforementioned cues also have some use in the resolution of front-back confusions, although, as with elevation perception, shoulder and torso reflections

only provide weak cues. With respect to spectral content the region around 2 kHz is important, as are high frequencies. In addition, head rotations have been shown to all but eliminate front-back confusions.

It has been shown in numerous studies that there the median plane localisation of band-limited stimuli is frequency dependent. The directional band phenomenon, for example, describes a condition in which band-limited stimuli (1/3 and 1/6-octave) are localised purely on the basis of frequency, irrelevant of the position of the emitting loudspeaker. This has been shown to be closely related to the aforementioned spectral cues provided by the pinnae. Additionally, the pitch-height effect states that tonal and octave band stimuli presented from vertically arranged loudspeakers in front of the subject are localised on the basis of frequency, with high frequency stimuli being localised physically higher in space than low frequency stimuli. As with directional bands, this may be related to spectral cues. Each of the aforementioned effects provides evidence to suggest that localisation thresholds may be different across the frequency spectrum. This in turn justifies the analysis of a band reduction method.

The phantom image elevation effect describes a localisation phenomenon observed in horizontal stereophony, whereby the perceived elevation of a phantom image is dictated by the base angle between the loudspeaker pair; an increase in base angle corresponds to an increase in perceived elevation. It has been shown in the literature that this effect is source dependent, with more natural and spectrally flat sources being more affected than sources with a low frequency weighting or a lack of high frequency energy. In the literature, it has also been hypothesised that this effect is the result of the brain misinterpreting acoustic crosstalk delay as a shoulder reflection cue. This effect has numerous implications for localisation thresholds. For example, if the effect is source dependent then different localisation thresholds may need to be applied to different sources, which is almost impossible in a practical recording environment. There may also be an impact on the threshold values themselves depending on loudspeaker base angles.

2 THE PERCEPTUAL EFFECTS OF SECONDARY VERTICAL SOURCES

The present thesis aims in part to illicit the most salient perceptual effects of vertical interchannel crosstalk and, further, to establish how they are affected by differences in the localisation threshold method applied to the height layer. This issue can be simplified by considering the direct sound present in the height layer as being a reflection of the main channel signal. When thought of in this way, a simple question arises: how does the perception of the main channel signal vary when a vertical reflection is present?

With respect to the above question, it is clear that one perceptual effect would be an increase in perceived loudness, with this being observed in the literature both for lateral [Haas 1972] and ceiling [Barron 1971] reflections. A range of other potential perceptual effects as a result of vertical reflections are explored in the present chapter. The issue at hand is broken down into three broad sections: the effect on perceived location, the effect on perceived spatial impression and the effect on perceived timbre.

2.1 THE EFFECT ON PERCEIVED LOCATION

This section discusses how a secondary vertical reflection might influence the perceived location of the main channel signal. The effects of both ICLD and ICTD between the direct sound and the vertical reflection are considered.

2.1.1 The Effect of ICLD

Blauert [1997] stated that when two sound sources radiate coherent signals on the horizontal plane, the resultant sound source is formed as a phantom image at a position intermediate between the two loudspeakers. If the amplitude of one of the loudspeakers is subsequently attenuated, then the position of the resultant image shifts in a direction biased towards the louder loudspeaker. A full image shift towards the loudspeaker of greater amplitude, in the case of loudspeakers arranged in conventional stereo, is perceived when the amplitude of the other loudspeaker is attenuated by a minimum of 15 dB [Williams 1987].

Experiments conducted by Somerville et al. [1965] analysed the above effect in relation to vertically arranged stereophonic loudspeakers located in front of the listening position. The upper loudspeaker was positioned 1.5 m above the lower. The experiment considered the effect of ICLD on the perceived elevation of clicks, with the range of ICLDs tested being ±16 dB (2 dB steps). The results of the study revealed the following. When the ICLD was 0 dB, the resultant phantom image was perceived to be roughly at the mid point between the two loudspeakers. Further, the perceived image position became biased towards the louder loudspeaker as the ICLD increased; a full image shift towards the position of the upper loudspeaker, for example, was perceived when the ICLD was 12 dB (lower loudspeaker attenuated). Based on these results it was concluded that the 'fusion of two sources in a vertical plane behaves somewhat similarly to that in a horizontal plane'. However, it should be noted that, within the range of ICLDs tested, there was no condition whereby the phantom image position matched that of the lower loudspeaker.

More recently, a somewhat similar study to that undertaken by Somerville et al. [1965] was conducted by Barbour [2003]. Speech and pink noise sources were presented to subjects as vertically oriented phantom images from stereophonic loudspeakers. With respect to the listening position, the lower loudspeaker was not elevated, whilst the elevation of the upper loudspeaker varied from 45°-90°. The ICLD varied by ±15 dB in 3 dB steps. The results of the study (Fig. 2.1) showed that, for all upper loudspeaker elevation angles, when the ICLD was 0 dB the perceived elevation of the resultant phantom image was between 20° and 25° (i.e. biased

towards the physical position of the lower loudspeaker). In addition, as the upper loudspeaker was attenuated with respect to the lower by 6 dB (-6 dB ICLD) the perceived source elevation fell to below 15°. Moreover, an ICLD of -9 dB was sufficient for the resultant position of each source to match the physical position of the lower loudspeaker in the case that the upper loudspeaker was elevated by 45°. This latter result is interesting with respect to localisation thresholds, as it suggests that the threshold ICLD at which the resultant phantom image is localised at the position of the lower loudspeaker lies between -6 and -9 dB. Equally, Lee [2011] found that the localisation threshold for musical sources was between -6 and -7 dB for ICTDs up to 5 ms. It would appear then that results of the two experiments show good agreement with one another, although it should be noted that each study utilised differing stimuli and upper loudspeaker elevation angles. The two studies were also undertaken with different contexts.

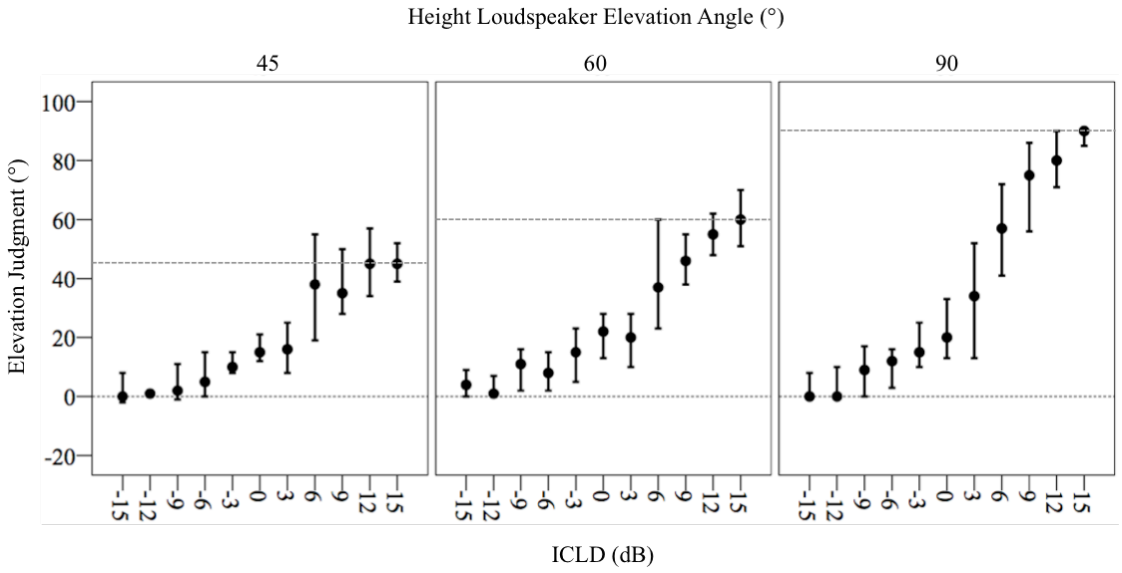

Fig 2.1: The experimental data reported by Barbour [2003], showing the effect of ICLD on median elevation judgments. The upper and lower dotted lines for each graph represent the physical position of the upper and lower loudspeakers respectively [after Barbour 2003].

Despite generally showing a similar trend, there is not entire agreement between the data reported by Somerville et al. [1965] and Barbour [2003]. In the latter study, a full image shift towards the lower loudspeaker was perceived when the ICLD was around -9 dB. However, Somerville et al. [1965] tested with

ICLDs up to -16 dB and found no such effect. In addition Somerville et al. [1965] showed that localisation judgments for the 0 dB ICLD condition were in a position equidistant between the two loudspeakers; Barbour's [2003] data suggested the resultant image was biased towards the lower loudspeaker. There are a number of possible reasons for these differences. Firstly each study used different sound sources, with Somerville et al. [1965] using clicks, whilst Barbour [2003] used pink noise and speech. This indicates that there may be a source dependent effect of ICLD on perceived source elevation. In addition, the two studies used different loudspeaker setup. However, it should be noted that Somervile et al. [1965] were particularly vague about the specific setup they used, which makes a direct comparison difficult. It is not clear, for example, whether or not the lower loudspeaker was elevated with respect to the horizontal plane. Neither is it clear what the elevation angle of the upper loudspeaker was with respect to the listening position. Nevertheless, from the two studies it can be concluded both that the phantom imaging of sound sources for horizontal stereophony is maintained for vertically arranged loudspeakers emitting a coherent signal and that the perceived position of the resultant sound source is influenced by the ICLD.

The principle that phantom images can be formed in vertical space based on amplitude differences between each loudspeaker has found a number of uses practically. Pulkki [1997], for example, developed an amplitude-based 3D panning algorithm known as 'Vector Base Amplitude Panning' (VBAP), which is able to virtually pan sources on the surface of a three-dimensional sphere based on the relative level differences between loudspeakers. Following this, vertical localisation experiments conducted by Wendt et al. [2014] showed that amplitude panning techniques, such as VBAP, were possible, however the stability of such techniques is largely dependent on the subject, with notable intra-subject differences in perceived source elevation being reported. In addition, Sundaram and Kyriakakis [2005] addressed the issue of a centre loudspeaker in a conventional 5.1 loudspeaker array obscuring a television display. A method was developed in this study whereby the centre signal was recreated as a vertically oriented phantom image presented from a pair of coherent speakers located above and below the horizontal plane respectively.

That the perceived elevation of a sound source presented from vertically arranged stereophonic loudspeakers is dependent on the relative amplitude between the two loudspeakers has important implications for the analysis of vertical interchannel crosstalk effects. It can be deduced from the literature that the presence of direct sound in the height layer of sufficient amplitude can affect the perceived elevation of the main channel signal, causing it to be elevated with respect to the horizontal plane. This in turn validates the analysis of localisation thresholds as an attempt to prevent this from happening. However, it should also be noted that Barbour's [2003] results in particular suggested that the 'panning' between the two loudspeakers as a result of changes in ICLD is not linear. Also, as was previously mentioned, Wendt et al. [2014] suggested that the perception of elevation as a result of amplitude panning is highly dependent on the subject. Therefore, depending on the ICLD between the direct sound in the respective layers, vertical interchannel crosstalk might have an unpredictable effect on the perceived elevation of the main channel signal, which in turn might influence the localisation thresholds.

2.1.2 The Effect of ICTD

2.1.2.1 The Precedence Effect

Variations in ICLD are not the only way in which the perceived location of a signal is influenced in horizontal stereophony. In a study conducted by Wallach et al. [1949] clicks were presented to subjects from stereophonic loudspeakers, which were located 3 m from the listening position. The distance between one of the loudspeakers and the listening position was systematically varied. The location of the resultant phantom image was perceived as being biased towards the physical position of the closer loudspeaker, with a full image shift observed when the difference in distance between each loudspeaker and the listening position was 0.3 m. The experiment was subsequently repeated with musical sources and speech, with each source revealing the same effect. Further testing on this phenomenon showed that this effect was a function of ICTD and not ICLD.

Blauert [1997] offered a full review on the underlying mechanisms governing the results that were observed in the Wallach et al. [1949] study. When horizontally arranged stereophonic loudspeakers emit a coherent sound, the resultant phantom image is formed at a position equidistant between the loudspeaker pair. Applying a delay to one of the loudspeakers, in the range of 0-1.1 ms, will cause the phantom image to shift to the earlier loudspeaker, which is known as 'summing localisation'. For delays exceeding 1.1 ms (corresponding to the 0.3 m reported by Wallach et al. [1949]), that are less than an upper, source dependent, limit, a full image shift is perceived, with the source appearing to emanate solely from the earlier loudspeaker. This full image shift is known as the 'precedence effect' [Wallach et al. 1949, Rumsey and McCormick 2009] or 'law of the first wavefront' [Blauert 1997, Rumsey 2005] and is demonstrated in Fig. 2.2. Should the delay between sources exceed the aforementioned upper limit then the sound from the delayed source is perceived as the echo of the earlier sound. This upper limit is called the 'echo threshold' and ranges from between 2 and 10 ms for clicks [Rozenzweig and Rosenblith 1950, Thurlow and Parks 1961] up to 50 ms for continuous speech signals [Haas 1972].

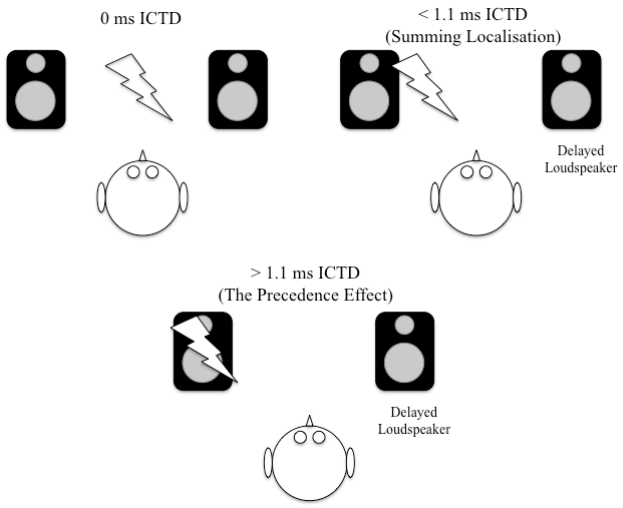

Fig. 2.2: The effects of ICTD on perceived location in horizontal stereophony.

It was supposed by Wallach et al. [1949] that the precedence effect was the mechanism by which the accurate localisation of sound is possible in a reverberant space. When considered in this way, the effect can be thought of as being an echo suppression mechanism. Hass [1972] analysed the degree to which this mechanism functioned. Stereophonic loudspeakers presented speech sources to subjects at a distance of 3 m. The angle between each loudspeaker and the listening position was 45°. In an initial experiment, the precedence effect was demonstrated by introducing ICTDs beyond 1.1 ms; the resultant images were localised at the position of the earlier loudspeaker. Subjects were subsequently required to perceptually match the amplitude of the loudspeakers by attenuating the leading loudspeaker for ICTDs between 1 and 40 ms. Haas [1972] demonstrated that, for delays of up to 20 ms, the lagging source could be increased in amplitude by as much as 10 dB with respect to the primary source before the two sources were perceived as being of equal amplitude. As a result of this study, the magnitude of the echo suppression effect has become known as the 'Haas effect' (Fig. 2.3).

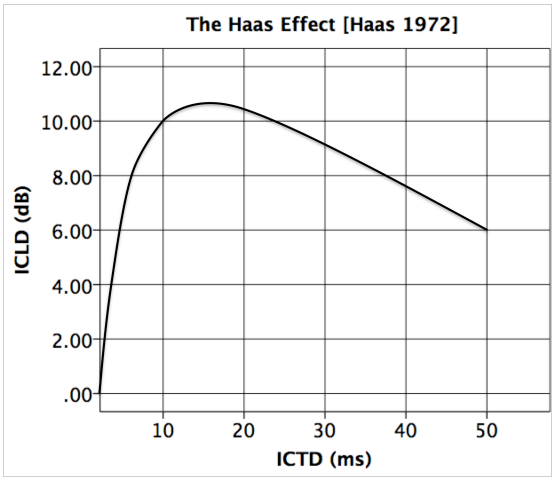

Fig. 2.3: The effect of ICTD on the echo suppression mechanism of the precedence effect, also known as the 'Haas Effect' [Haas 1972].

It should be noted that the operation of the precedence effect is not consistent for all sound sources. According to the literature, the effect generally operates more strongly for sources that are transient in nature [Wallach et al. 1949, Yost et al. 1971, Rumsey 2005], such as clicks, piano music or speech. Rakerd and

Hartmann [1986] expanded on this point following experiments in which the effect of onset time on the operation of the precedence effect was analysed in various sized rooms, each with single reflecting surfaces. The test stimuli were 500 Hz sine tones, whose onset time was varied between 0 and 5000 ms. According to the results of the study, the effect operates most strongly when the onset of the signal is instantaneous and progressively weakens with increases in onset time. The maximum onset time for the operation of the effect was found to relate to the size of the room (i.e. the delay time between the direct sound and reflection), with longer delays resulting in larger onsets before the precedence effect broke down and localisation accuracy decreased (up to 500 ms for 9 ms delay). Somewhat related to this point, Wallach et al. [1949] hypothesised that the precedence effect would operate less strongly for continuous noise sources, as there was no way to define precedence. This however disagrees with Hartmann [1983], who suggested that continuous noise features 'random amplitude fluctuations, which serve as transients'. This would seemingly indicate that the localisation of such sources in horizontal stereophony would be influenced by the precedence effect.

2.1.2.2 Evidence for the Precedence Effect in the Median Plane

The precedence effect is primarily related to localisation in the horizontal plane. Despite this, it has been shown in the literature that the effect also operates for sources incident from the median plane, although it should be noted that there is not total agreement on this subject. Whether or not the precedence effect operates in the median plane has implications for the application of localisation thresholds. Studies conducted by Barbour [2003], Lee [2011] and Stenzl et al. [2014] each showed that sufficient ICLD could prevent the vertical migration of the main channel signal in the presence of a secondary elevated source. Nevertheless, if the precedence effect was found to operate in the median plane then ICLD may not be necessary at all, as sufficient delay between the main and height layers would prevent the direct sound signal in the height layer from having any effect on the perceived location of the main channel signal. Despite this, it is likely that the main channel signal would be affected in other ways, Freyman et al. [1991], for example,

noted that the presence of a lagging sound can alter the 'loudness, pitch, quality and spatial extent of the auditory image', which is discussed in more detail later.

One of the earliest studies that considered the precedence effect in the median plane was conducted by Blauert [1971]. Two loudspeakers were positioned 3.5 m from the heads of subjects, one directly in front and the other behind, in a room that was 'nearly echo free'. The distance of the back loudspeaker could be varied by ±0.3 m with respect to the listening position. The test stimuli were white noise pulses, 1.7 s in duration. Subjects were required to identify if each stimulus was perceived to be in front of, behind or above them. When the rear loudspeaker was closer to the listening position than the front loudspeaker, localisation judgments were biased towards the rear. As the rear loudspeaker moved further away the frequency of rear responses lessened, with responses becoming more biased to the front. The crossover for the change began at around 0.5 ms. It was therefore concluded that the precedence effect operated in the median plane for ICTDs above 0.5 ms. Below this threshold, location judgments were biased towards the earlier loudspeaker, which was considered as being in line with summing localisation.

Experiments conducted by Litovsky et al. [1997] produced results that supported Blauert's [1971] conclusions regarding the operation of the precedence effect in the median plane. Five loudspeakers were arranged in an anechoic chamber directly in front of, behind, above and to the left and right of the listening position. Leading and lagging clicks were presented to subjects separately from the horizontal and median planes. ICTDs ranging from 0-10 ms were tested. For each experiment, subjects were required to identify the loudspeaker closet to where the sound image was perceived. The experimental data suggested primarily that the precedence effect operated in the median plane for ICTDs between 1 and 5 ms, albeit at a much weaker strength with respect to the same effect in the horizontal plane. The effect was found to be at its strongest for ICTDs between 1 and 2 ms, whilst above 5 ms the lagging click was perceived by subjects as being an echo. From their experiments, Litovsky et al. [1997] hypothesised that, although the precedence effect operated similarly in both the median and horizontal planes, the cues used to trigger the effect in each plane were different. They supposed that the horizontal plane precedence effect was a function of binaural cues, whilst

the effect in the median plane was a result of 'spectrally based localisation cues'. Further, they argued that, as a result of this hypothesis, any models of the precedence effect based solely on binaural cues would be insufficient.

A more recent study conducted by Tregonning and Martin [2015] also concluded that the precedence effect operates in the median plane. For their experiments, female speech and conga stimuli were presented at 76 dB SPL from the frontal surround and height layers of an Auro 3D configuration [Auro Technologies 2016]. The distance between each loudspeaker and the listening position was 2.44 m. Stimuli were presented from both the main and height layers simultaneously and were panned horizontally by 0°, ±15° and ±30°. An ICTD was then applied to the height layer. The ICTDs used ranged from 5-25 ms for the speech and 1-5 ms for the conga, which were based on each stimulus' echo threshold. For the experiment, subjects were required to identify the perceived elevation of the stimuli. The results of the experiment showed that localisation judgments were generally in a position biased towards the earlier (lower) loudspeaker layer. As a result of this Tregonning and Martin [2015] concluded that the precedence effect operated vertically, although it should be noted that a full image shift towards the earlier loudspeaker was not observed for any condition. An additional result of interest in the study related to the effect of ICTD. When the ICTD was less than 5 ms, its primary effect was to influence perceived source elevation. However, as the ICTD increased beyond this its effect on perceived elevation lessened, with there being a greater influence on the vertical image spread of each source.

Although there is an apparent similarity in the conclusions of the aforementioned studies, not all experimental data reported in the literature agrees that the precedence effect operates in the median plane. During his analysis of localisation thresholds for musical sources, Lee [2011] argued that, if the precedence effect were indeed a feature of median plane localisation, sufficient ICTD alone would have been appropriate for the localisation thresholds to be met. This in turn would render changes in ICLD entirely unnecessary. However, this was not found to be the case from his experimental data, with ICLDs always being required, even as the ICTD approached 50 ms. Lee [2011] therefore concluded that the precedence effect does not

operate in the median plane. A similar result was subsequently obtained by Stenzl et al. [2014]. It should be noted, however, that, as direct localisation experiments were not conducted, the results of each study were only able to suggest a lack of precedence effect indirectly.

The reasons for the differing conclusions regarding the precedence effect in the median plane are perhaps related to differing interpretations of the effect, as was discussed by Lee [2011]. Blauert [1997] referred to the precedence effect in horizontal stereophony as operating when the perceived location of the sound source corresponds to the location of the loudspeaker whose signal arrives first. Arguably this therefore refers to a 'full image shift' towards the earlier loudspeaker. There is evidence in the studies of Blauert [1971], Litovsky et al. [1997] and Tregonning and Martin [2015] that they did not adhere to this definition when analysing the effect in the median plane:

"While the position of the loudspeakers was changed, the test person was instructed to state the respective direction of his sound sensation. He was allowed the following answers: "v" (front), "o" (above) or "h" (back)." [Blauert 1971]

"…after the stimulus was presented, subjects had to decide which of the three loudspeakers on the plane was closest to the location of the sound image." [Litovsky et al. 1997]

"The images never reached the speaker layers…It was found that vertical ICTDs had a statistically significant impact... This is in agreement with the previous research showing the existence of a vertical precedence effect." [Tregonning and Martin 2015]

Each of the above quotes demonstrates that the authors were concerned more with the localisation dominance of the earlier loudspeaker, rather than the precedence effect as defined for horizontal stereophony by Blauert [1997]. Such a localisation dominance effect was also demonstrated in the experiment conducted by Somerville et al. [1965]. Alongside testing the effects of ICLD on the localisation of clicks presented from vertically arranged stereophonic loudspeakers, the study analysed the effect of applying a 20 ms delay to the lower loudspeaker. The effect of the delay was to bias localisation towards the upper loudspeaker position in the case that there was 0 dB ICLD between the loudspeaker pair. As a result of this, less level reduction was required for a full image shift towards the upper loudspeaker (8-10 dB) compared to when no delay was applied (13-15 dB). For no subject tested was the 20 ms delay alone sufficient to cause a full image shift towards the earlier loudspeaker, although for one of the subjects the difference in perceived elevation and the physical position of the upper loudspeaker was small. This result closely matches those of Lee [2011] and Stenzl et al. [2014] with respect to ICLD always being necessary for a full image shift in the case of vertically arranged stereophonic loudspeakers.

The above discussion arguably shows that the precedence effect has not yet been shown to operate in the median plane. There is, however, a potential localisation dominance of the earlier loudspeaker somewhat akin to the summing localisation discussed by Blauert [1997]. The existence of a vertically oriented localisation dominance effect naturally has implications for localisation thresholds. If a delay in the height layer results in the main channel image being localised in a position biased towards the main channel layer, as was demonstrated by Somerville et al. [1965] for clicks, then this would suggest that the resultant localisation threshold would be lower (i.e. less level reduction would be required). An illustrative example of this concept is shown in Fig. 2.4. However it should be noted that this is not consistent with the results of either Lee [2011] or Stenzl et al. [2014], who found no significant effect on localisation thresholds for upper loudspeaker delays of between 0 and 5 ms. In other words, their results failed to demonstrate the idea that a localisation dominance effect has a significant effect on localisation thresholds when using the blanket reduction method. Further to this point, as the ICTD was increased beyond 10 ms the localisation thresholds also increased in both studies. It therefore remains to be seen whether or not either the precedence or

localisation dominance effects can be demonstrated for vertically arranged loudspeakers and further the implications that this would have for localisation thresholds.

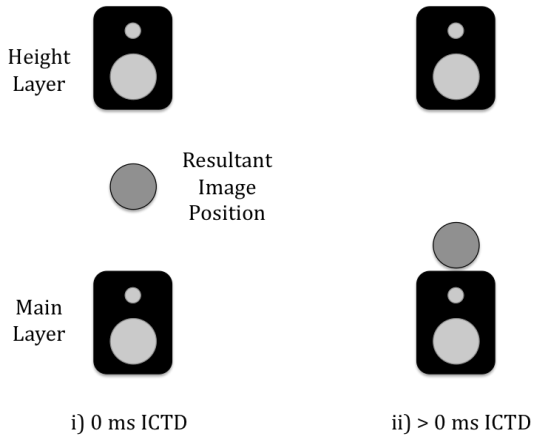

Fig. 2.4: Illustrative example of the potential effects of height channel delay on perceived image elevation for i) 0 ms ICTD and ii) > 0 ms ICTD (i.e. localisation dominance effect).

2.2 THE EFFECT ON PERCEIVED TIMBRE

So far, the effects of vertical reflections on the main channel signal have been considered from the perspective of localisation. However, localisation-based effects are likely not the only changes that would be perceived by subjects. Halmrast [2001] stated that the addition of a reflection would affect the frequency response of the signal, which Seki and Ito [2003] defined as 'colouration'. This colouration is predominantly realised as a result of an interference effect known as 'comb filtering', which is discussed in detail in this section.

2.2.1 Physical Parameters for Comb Filtering

Comb filtering occurs when a signal is added to a delayed version of itself [Toole 2009]. Fundamentally the effect is due to the interaction between sound waves (Fig. 2.5), which was described by Farnell [2010] as follows:

'The amplitude of a new wave created by adding together two others at some moment in time, or point in space, is the sum of their individual amplitudes. Two waves of the same frequency, which have the same phase, match perfectly. They will reinforce each other when superposed, whereas two waves of opposite phase will cancel one another out. If two waves are travelling in opposite directions and meet at some point, they interfere with each other'. [Farnell 2010]

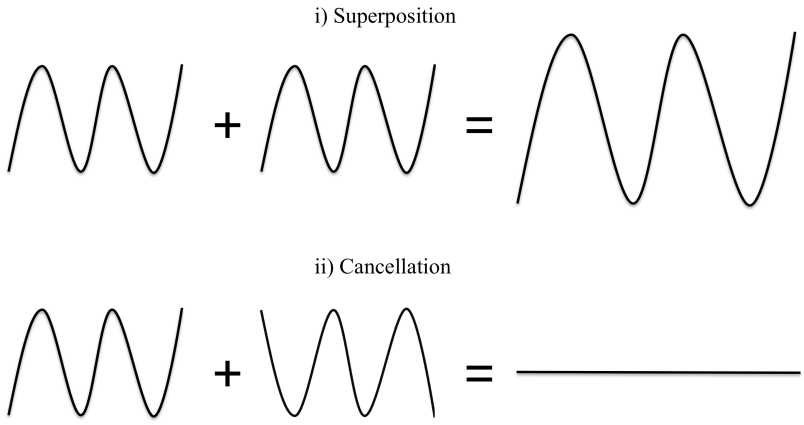

Fig. 2.5: Superposition and cancellation effects from interfering sine waves.

Howard and Angus [2009] discussed phase cancellation in relation to coherent sound sources emitted from two loudspeakers located on the horizontal plane. When there is no ICTD between tonal stimuli, the two

sources will add constructively because they are in phase. However, applying a delay to one of the loudspeakers will result in phase cancellation (i.e. no net amplitude) should the path difference between each source be equal to half a wavelength (1/2 λ). For a complex sound, the pattern of superposition and cancellation results in a frequency spectrum that resembles the teeth of a comb, hence the term 'comb filtering'.

Toole [2009] provided a series of equations in order to calculate the frequencies that will be attenuated and those that will be superposed as a result of comb filtering. The first frequency at which destructive interference will occur can be calculated as follows:

$$f = 1/p \tag{2.1}$$

Where f is the frequency and p is the period of the wave, which can be calculated by doubling the delay between the respective sources. The cancellations at higher frequencies will occur at those frequencies with an odd number of half-wavelengths in the delay interval. They can be calculated as follows:

$$f_n = n(odd)/p \tag{2.2}$$

Where $n(odd)$ is the number of odd integers. With respect to those frequencies that will be superposed (i.e. double in amplitude) the formula is as follows:

$$fPn = n(all)/d \tag{2.3}$$

Where fP is the peak frequency and d is the delay between the two sources.

Toole [2009] stated that the magnitude of the comb filtering effect is dependent on the amount of delay between the reflection and direct sound and also their relative amplitudes. The point was expanded on by

Halmrast [2001], who suggested that a short reflection will give very broad comb filter, whilst late reflections will result only in small ripples in the spectra. The effects of both ICTD and ICLD on the severity of comb filtering are demonstrated in Fig. 2.6. Here it can be seen that the bandwidth of the notches caused as a result of comb filtering is determined by the ICTD; an increase in ICTD decreases the bandwidth of the notches and increases their overall number. In addition, the effect of ICLD is to influence the degree of attenuation for each notch, with decreases in ICLD resulting in less overall reduction in the signal. From a vertical interchannel crosstalk perspective, this analysis demonstrates that timbral colouration of the main channel signal would be the most severe when the crosstalk signal is similar in amplitude and arrives with a short delay with respect to the main channel signal.

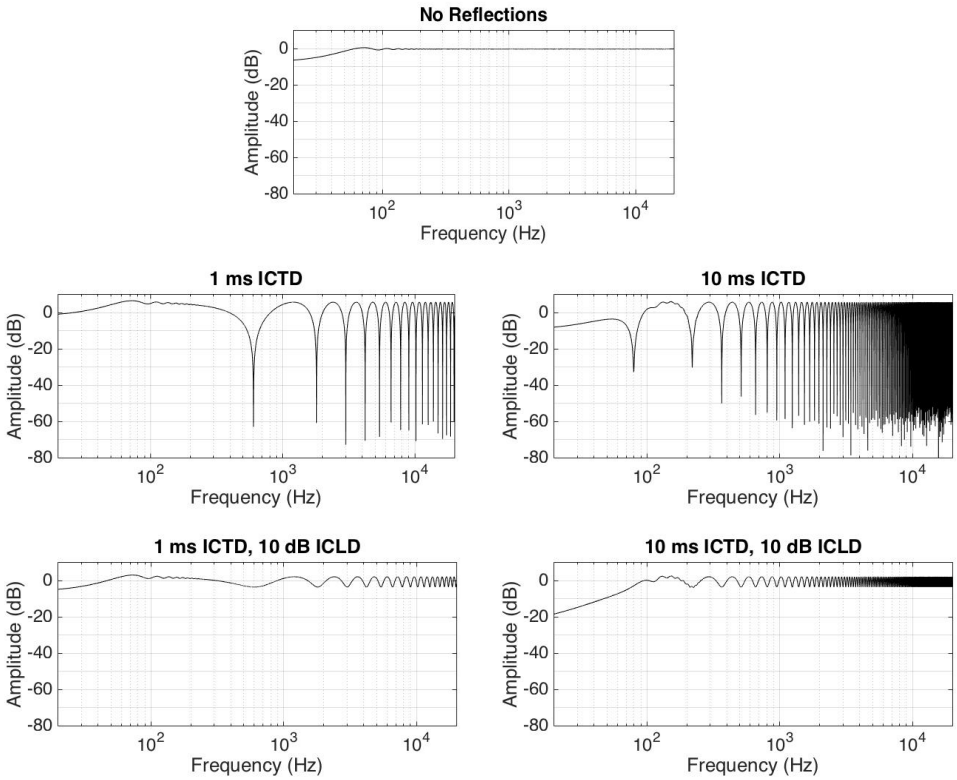

Fig 2.6: The effects of ICTD and ICLD on comb filtering for sine sweeps.

2.2.2 The Perceptual Effects of Comb Filtering

According to Everest and Pohlman [2009], comb filtering is a steady state phenomenon and therefore the implications for music and speech are limited. Despite this, numerous studies have considered the perceptual effects that comb filtering has on musical sources. The majority of these have considered conditions whereby lateral reflections are present, however some studies have also given attention to the perceptual effects of vertical reflections.

2.2.2.1 The Effect of Lateral Reflections

According to Toole [2009], comb-filtering effects can be considered as having both positive and negative influences on musical sources. Indeed, the literature provides a plethora of evidence to support both views in the case that the secondary source is a lateral reflection. Experiments conducted by Somerville et al. [1965] supported the notion that comb filtering is a negative effect. In that study, the perceptual effects of lateral reflections on a direct sound were considered for ICTDs up to 80 ms. The test stimuli were orchestral excerpts and male speech. Sound source presentation was either through headphones or loudspeakers. For the former condition, it was reported that delays in the range of 10-40 ms resulted in a notable decrease in quality, particularly for the string section of the orchestra. A similar effect was observed for loudspeaker presentation, whilst the speech source was also coloured with both low and high frequency distortions.

Other studies reporting on the negative effects of lateral reflections on perceived timbre are as follows. Muller [1968, cited in Barron 1971] suggested that prominent colouration effects are perceived for instrumentation and speech sources when reflections with short delays are present. Also, Barron [1971] reported that lateral delays between 10 and 50 ms sharpen the tone of orchestral sources, with the effect being the strongest at around 20 ms ICTD. More recently, Rumsey [2005] has described comb-filtering effects for ICTDs less than 20 ms as being ‘severe’.

Despite the results of the aforementioned studies, numerous experiments have reported that lateral reflections have beneficial effects on the timbre of music. Barron and Marshall [1981], for example, reported that lateral reflections cause orchestral music to gain body and fullness, which they regarded as being a positive attribute. Additionally, in a more recent study, Halmrast [2000] investigated colouration effects for live orchestral performances. For part of the experiment, the orchestra was surrounded with reflective sidewalls, whose distance with respect to the orchestra was systematically varied. The audience in the study noted that a pleasing effect was realised as a result of moving the sidewalls closer to the ensemble (i.e. decreasing the ICTD between the direct sound and reflection). It is apparent then that lateral reflections have an audible influence on the timbre of musical sources and speech, however there is generally no agreement as to whether or not such effects are preferable.

2.2.2.2 The Effect of Vertical Reflections

The effect of vertical reflections on perceived timbre is a topic that has garnered little attention in the literature. In the aforementioned study conducted by Halmrast [2000], a condition was tested whereby suspended reflectors were positioned above the orchestra. When the distance between the orchestra and reflectors was in the range of 6-7 m, the audience described the resultant sound as if the orchestra were playing inside a box (box-klangfarbe). This effect was less noticeable when the elevation of the reflectors was increased to 9 m. Impulse response measurements subsequently demonstrated that box-klangfarbe was related to reflections between 5 and 20 ms, with the effect lessening as the delay time increased; this being in line with comb-filtering effects decreasing with increases in ICLD (i.e. the amplitude of the reflection will decrease with increases in distance).

Another key result with respect to the perceptual effects of comb filtering from vertical reflections was Barron and Marshall's [1981] finding that colouration effects as a result of a ceiling reflection (0° azimuth, 40° elevation) were more audible compared to those for lateral reflections. However, they did not indicate

specifically what effects were perceived. This result is interesting with respect to the analysis of vertical interchannel crosstalk. In experiments conducted by Lee [2006] the perceptual effects of horizontal interchannel crosstalk for the frontal array of 3-2 microphone techniques (i.e. microphone techniques used for recording in 5.1 surround) were analysed. In an initial experiment anechoic musical sources were presented to subjects using both a crosstalk off (C and R loudspeakers only) and crosstalk on (L, C and R loudspeakers) conditions (Fig. 2.7). The signal from the 'L' loudspeaker in this case can be considered as being a lateral reflection, with the phantom image formed between C and R being the direct sound. As well as eliciting the key perceptual attributes of the effect, Lee [2006] also analysed the 'audibility index' of each attribute (i.e. how audible each attribute is). The results showed that source width and locatedness were the attributes most affected by horizontal interchannel crosstalk, with audibility indexes of 6.5 and 4.7 respectively. However, with respect to timbral effects the audibility indexes were much lower. Fullness, for example, rated at 3.5, whilst brightness scored 1.4 and phasiness 0.5, among other results. This indicates that, although timbral changes are audible for musical sources as a result of an interfering horizontal reflection, they are not as audible as spatial and location changes. However, based on Barron and Marshall's [1981] conclusions, it could be argued that timbral variations would be more audible as a result of vertical interchannel crosstalk compared to horizontal interchannel crosstalk. This is examined in more detail in Chapter 5 of the present thesis.

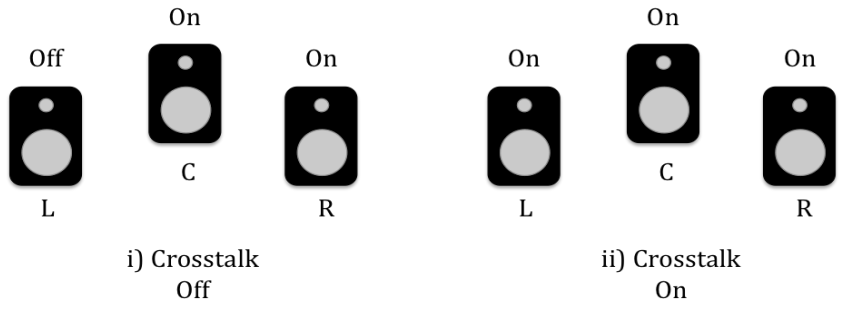

Fig. 2.7: Crosstalk on (i) and crosstalk off (ii) conditions used by Lee [2006] in attribute elicitation experiments for the effects of horizontal interchannel crosstalk in 3-2 microphone technique.

2.2.3 Thresholds for the Audibility of Comb Filtering on Perceived Timbre

A number of studies have considered the threshold at which reflected sound has an audible influence on the timbre of a direct sound. Olive and Toole [1989] presented music, speech, pink noise and noise pulses from a loudspeaker located directly in front of the listening position. The reflecting loudspeakers were located as follows with respect to the listening position: lateral (65° azimuth, 0° elevation), vertical (0°, 65°) and rear (115°, 0°). All loudspeakers were 2 m from the listening position. Subjects completed a series of method of adjustment tasks in which the amplitude of the reflection was adjusted until it was just audible in the presence of the direct sound. Key findings from their experiments in relation to the present thesis are as follows:

1. Delayed sounds arriving from the same direction as the direct sound are less audible than for other reflection locations. For pink noise the masking effect of the direct sound increased the threshold by between 5 and 10 dB.
2. For lateral reflections, the threshold obtained is dependent on the sound source. Complex orchestral sources had thresholds that were affected little by changes in ICTD, whilst speech and individual instruments had a decrease in threshold with increases in ICTD. The latter sources will therefore be affected less by timbral colouration if the delay between the direct and reflected sounds is increased (most notably beyond 40 ms).
3. The threshold was not consistent between individual musical sources. This further suggests that the severity of timbral colouration is source dependent.
4. For pink noise, lateral reflections influenced the sensation of spaciousness at levels close to the threshold. For the vertical reflections, the effect was a change in timbre. This agrees with Barron [1971] in that timbral effects are more severe for vertical reflections than for lateral reflections.

A subsequent study conducted by Bech [1995] analysed both the directions at which reflected sound has the most influence on the overall timbre of a source, as well as the amount of level changes necessary in a reflected sound for there to be an audible change in the overall sound field. The direct sound was emitted from a loudspeaker located to the right-of-centre of a listening room (-22° azimuth, 0° elevation). 17 loudspeakers were positioned in varying locations around the room in order to simulate reflected sound coming from a wide range of azimuths and elevations with respect to the direct sound. A further five loudspeakers were added to provide reverberation. All loudspeakers were positioned 3.1 m from the listening position. Subjects completed an adaptive two-alternative forced choice procedure, during which they were asked to identify the threshold for the detection of the reflected sound for pink noise and speech stimuli. The individual thresholds were subsequently compared to the measured values of a natural reflection in a standard listening room. The results of the experiment showed that only the first-order ceiling (-25° azimuth, 48.2° elevation) and floor (-25°, -28°) reflections contributed individually to the timbre of the speech signal. In addition the thresholds were higher for the speech compared to the pink noise, of which the timbre was also strongly influenced by lateral reflections (65°, 0° and 65°, 30°). This experiment therefore shows the same source dependency for coloration as was reported by Olive and Toole [1989]. Of further interest is the result that demonstrates the importance of vertical reflections in determining the overall timbre of the sound source, which is directly applicable to vertical interchannel crosstalk effects and shows agreement with Barron [1971].

Brunner et al. [2007] considered the just noticeable difference in timbre for snare, speech and piano sources, which were presented monophonically to subjects through headphones. The reflected sounds were delayed by between 0.1 and 15 ms. Subjects completed a 3-down-1-up adaptive tracking procedure to detect the influence of a comb filter on the timbre of each source. Step sizes of 0.5 dB were used for the first reversal, before falling to 0.25 dB for subsequent judgments. The thresholds were found to differ considerably depending on the type of sound source. A minimum mean threshold was observed for ICTDs of 0.8 ms for the piano (13.2 dB) and snare (18.2 dB). For the speech, the mean threshold was 8 dB for 0.1 ms ICTD, which decreased steadily to around 16 dB at 15 ms. Individual subjects also reported thresholds beyond 20

dB for all stimuli. From this study, it was concluded that irregularities in the frequency response (coloration) are more audible than was originally considered. Perhaps this result is related to the use of headphones to present the test stimuli, with Olive and Toole [1989] and Bech [1995] each reporting higher thresholds for musical sources when presented from loudspeakers. Nevertheless, the result is valuable in showing that coloration effects are source dependent, which agrees with Olive and Toole [1989] and Bech [1995].

In relation to vertical interchannel crosstalk, the aforementioned studies are valuable in numerous ways, Firstly, that coloration effects are influenced by sound source means that certain sources may undergo a greater amount of timbral coloration at the localisation threshold than do others. In addition the results presented by Olive and Toole [1989] and Brunner et al. [2007] suggest that an increase in the ICTD may increase the amount of colouration perceived by subjects. This would therefore indicate that, if the effect was deemed negative, the ICTD between the height and main layers of either loudspeakers (in the case of image rendering) or microphones (in practical recording environments) should be kept to a minimum. Moreover, in relation to the work of Lee [2011], the aforementioned studies demonstrate that at the localisation threshold (-6 to -7 dB for ICTDs up to 5 ms) the ICLD between the main channel and crosstalk signals is not sufficient to prevent colouration, with Olive and Toole [1989] showing that the maximum threshold for the audibility of a reflection in the presence of a direct sound is -10 dB ICLD. This therefore demonstrates that the timbre of the main channel signal would be affected by the presence of a reflection even at the localisation threshold. The specific ways in which the timbre is affected is explored in Chapter 5.

2.2.4 Other Factors Affecting the Audibility of Comb Filtering

Alongside being somewhat source dependent, the audibility of comb filtering also has a dependency on the number of sound sources and the presence of reverb. Halmrast [2001], for example, noted that the overall effect of comb filtering was more severe for loudspeaker presentation compared to the same piece played by

a live orchestra. It was supposed that this related to the number of reflections and interfering sources, which were found to be inversely proportional to the amount of audible colouration.

The literature supports the idea that an increase in the number of reflections lessens the audibility of comb filtering effects. Olive and Toole [1989] reported that the thresholds of detection for speech, music and noise signals increases in the presence of reverberation delayed by more than 30 ms with respect to the direct sound. However, for delays below this the effect is minimal. Additionally, Bech [1995] found that the threshold of detection of speech and pink noise reflections in the presence of a direct sound increases by between 2 and 5 dB if reverberation is removed. This result was interpreted in terms of the reverberation acting as a masker, which affects the audibility of the colouration as a result of comb filtering. Furthermore, it was concluded that the contribution of individual reflections to the timbre of the sound field would increase with decreasing levels of the reverberant field. Toole [2009] suggested that comb filtering is the most audible in an echo free environment in the presence of a single, strong reflection, particularly from the median plane, however ceases to be an issue in a normal reflective environment.

The effect of reverberation on coloration has applications in light of developing 3D audio systems. One of the benefits of the addition of the height channels is to increase the spatial attributes of the sound field through the addition of reflected sounds. It is therefore plausible that comb filtering effects caused by the presence of vertical interchannel crosstalk signals could be masked at the localisation threshold provided sufficient late reflections were present at the reproduction stage. It is unfortunate that this was not analysed by Lee [2006] for horizontal interchannel crosstalk. In that study the timbral effects elicited were regarded as having too low an audibility index to be considered as being salient. In subsequent experiments Lee [2006] analysed the effects of reverberation on the grading of the most salient attributes (source width and locatedness) and it would have been interesting to see how timbral changes were affected by the different conditions. Another point of note, however, is that timbral coloration is reportedly more severe for vertical reflections than for horizontal [Barron 1971], which means that the timbral changes caused as a result of vertical interchannel crosstalk may be stronger than those obtained for horizontal interchannel crosstalk by

Lee [2006]. If the effect were more severe vertically then perhaps the masking effect of reverberation would be less strong compared to if the crosstalk signal were a lateral source. This would have to be studied further.

2.3 THE EFFECT ON PERCEIVED SPATIAL IMPRESSION

The present section undertakes a review of spatial impression and how it relates to vertical interchannel crosstalk. Morimoto [2002] described spatial impression as being a multidimensional characteristic of human auditory perception that is associated with the acoustics of space. In addition, Mason and Rumsey [2000] referred to it as being 'the auditory perception of the locations, dimensions, and other physical parameters of a sound source and the acoustic environment in which the source is located'.

It should be noted that a number of early studies examined spatial impression in the context of lateral reflections. Therefore, although the present section is concerned with analysing the perceptual effects of vertical reflections on spatial impression, it is considered as being necessary to also review spatial impression in the horizontal plane in order that the issue at hand may be more fully understood. In this regard spatial impression is generally divided into two broad sections: 'apparent source width' (ASW) and 'listener envelopment' (LEV) [Bradley and Soloudre 1995].

2.3.1 ASW

The concept of ASW is shown in diagrammatical form in Fig. 2.8. Everest and Pohlmann [2010] described the effect as relating to the 'perceived breadth' of a sound source. Additionally, Toole [2009] considered ASW to be the 'broadening' of a source, whilst Sato and Ando [2002] related it to a source's 'acoustical width'. The perception of ASW is predominantly the result of reflections that arrive within 80 ms of the direct sound, which are also known as 'early reflections' [Rumsey 2005].

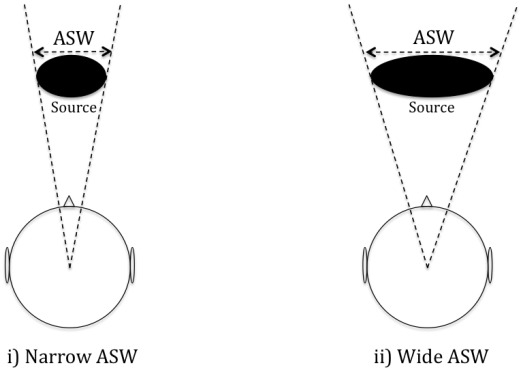

Fig. 2.8: Diagram of the concept of ASW.

2.3.1.1 Characteristics of ASW

In a series of early studies, a number of the characteristics of ASW were derived. It should be noted that these studies did not describe ASW directly, as the term had not yet been developed. Instead, terms like 'spatial impression' were often used. However, it was noted by Bradley and Soloudre [1995] that the effects that were being described in those studies were indeed synonymous with ASW.

Barron [1971] presented musical sources to subjects from a loudspeaker positioned directly in front of the listening position at a distance of 3 m. The same source was presented from a second loudspeaker (40° azimuth) as a lateral reflection, with delays between 0 and 120 ms being applied to the reflection loudspeaker with respect to the main. The reported effects of the lateral reflection included timbral and localisation changes, which were regarded as being negative. However, also reported for the majority of lateral reflections was an increase in 'spatial impression' (ASW), which was produced for delays between 10 and 80 ms and resulted in an apparent broadening of the sound source. From his experiments, Barron [1971] noted that the perceived magnitude of ASW increased with increases in the amplitude of the reflection relative to the direct sound. Additionally, reflection delays less than 5 ms were found to contribute little to ASW. The effect was also found to be independent of delay time for delays up to 80 ms.

Barron and Marshall [1981] built upon the experiments of Barron [1971] and were able to address the effect of early reflections on ASW in more depth. A number of the experiments reported in the study were repetitions of those conducted by Barron [1971]. However, novel conclusions regarding ASW were nevertheless made for orchestral music. During one of the experiments, it was identified that ASW was at a maximum when the reflections were presented from two loudspeakers located at ±90° on the horizontal plane. Subsequent measurements showed that, when reflections of different azimuths were compared, a different result was obtained with a single reflection compared to a reflection pair. It was considered that this was due to image shifts caused by the lateral reflections. Following subsequent verification experiments, the following hypothesis was made:

"A lateral reflection from an azimuth α contributes a fraction of its energy, λ, to the subjective lateral sound and the remaining proportion (1-λ) to the subjective frontal sound…The spatial impression is assumed to be a monotonic function of the ratio of lateral to frontal energy". [Barron and Marshall 1981]

Barron and Marshall [1981] used the term 'lateral energy fraction' (L_f) to describe their hypothesis. Noting that ASW linearly relates to L_f within the first 80 ms after the arrival of the direct sound. Their equation for lateral energy fraction is as follows:

$$L_f = \sum_{t=5\text{ ms}}^{80\text{ ms}} r\cos\phi \quad / \quad \sum_{t=0\text{ ms}}^{80\text{ ms}} r \tag{2.4}$$

Where *r* is the energy of the reflection. Additionally, *r* in the divisor also includes the energy of the direct sound. L_f calculations made by Barron and Marshall [1981] based on the above formula were shown to match closely to their experimental data.

Blauert and Lindemann [1986] also studied ASW, which they called 'auditory spaciousness'. They conducted a series of experiments in which orchestral sources were presented to subjects from a loudspeaker located directly in front of the listening position. Early lateral reflections and reverberation were produced from loudspeakers located at azimuths of either ±45° or ±90° on the horizontal plane. All loudspeakers were 2 m from the listening position. With respect to the direct sound, the early lateral reflections were delayed by 20 and 30 ms. A number of characteristics of ASW were derived, which included perceptual judgments on preference. Firstly, it was identified that ASW is most influenced by early lateral reflections, this agreeing with both Barron [1971] and Barron and Marshall [1981]. Furthermore, sources that were perceived as being spacious were generally more preferred, indicating that ASW is a positive attribute of a sound source. A further finding of note was that when the lateral reflections only featured spectral energy below 3 kHz a sense of depth was perceived; ASW was only influenced when energy above 3 kHz was present. This demonstrates both that ASW is a multidimensional attribute and that the image widening reported by Barron [1971] is a result of high frequency energy present in early lateral reflections.

2.3.1.2 ASW in relation to IACC

Although Barron and Marshall [1981] related the degree of perceived ASW to the lateral energy fraction, the literature has shown that this is not the only attribute with which ASW correlates. Blauert [1997] related ASW to the similarity of the ear input signals, which is also known as interaural cross correlation (IACC) [Rumsey 2005]:

"(In the case of two sources radiating a coherent signal) When the ear input signals are fully coherent, a single auditory event of relatively limited extent appears. Its "centre of gravity" is on the median plane. As the degree of coherence decreases…the area over which components of the auditory events are found becomes greater." [Blauert 1997]

A number of studies have demonstrated the above point experimentally. Licklider [1948] and Chernyak and Dubrovsky [1968, cited in Kurozumi and Ohgushi 1983] each demonstrated the relationship between IACC and ASW using both speech and white noise sources presented through headphones. Keet [1968, cited in Barron 1971] measured IACC by comparing differences in impulse responses recorded by stereo microphones and identified a linear and inversely proportional relationship between ASW and IACC. Kurozumi and Ohgushi [1983] investigated the effect of cross-correlation on the sound image quality of white and band-limited noise. Cross-correlation differs from IACC in that it relates to differences between the inputs of a source at different loudspeakers instead of at the ears. They found both that the sound image appeared wider as the cross-correlation decreased and that cross-correlation and IACC were closely related.

2.3.1.3 ASW as a Result of Vertical Reflections

According to Furuya et al. [1995] the perception of ASW is affected little by vertical reflections. This was shown experimentally by Barron [1971], who considered the perception of ASW when single lateral (40° azimuth) and ceiling (40° elevation, 0° azimuth) reflections were presented simultaneously alongside a direct sound. The effect of the ceiling reflection on the direct sound was to influence its perceived timbre and elevation, however the overall effect on ASW was small. This seems somewhat logical when considering the direction from which the reflection is incident with respect to the direct sound. For a lateral reflection, the direct sound and reflection are both on the horizontal plane. It therefore seems reasonable that the presence of the reflection would make the source broader (i.e. more spread horizontally), which would increase the perception of ASW. Conversely, for a vertical reflection it is more likely that the source would extend vertically than it would horizontally. This concept is known as ‘vertical image spread’ and is considered in Section 2.3.3.

The apparent minimal effect of vertical reflections on the perception of ASW is interesting with respect to vertical interchannel crosstalk. When investigating horizontal interchannel crosstalk in the context of the

frontal array of microphones in a 3-2 configuration, Lee [2006] noted that one of the more salient effects was an increase in horizontal source width. Additionally, it was earlier mentioned how the vertical interchannel crosstalk signal was somewhat similar to a vertical reflection. Therefore, based on the above discussion, it is somewhat apparent that the source width increases reported by Lee [2006] for horizontal interchannel crosstalk would not be maintained for vertical interchannel crosstalk. This would seemingly indicate that the most salient effects of the two interchannel crosstalk effects would be different, at least from the perspective of spatial impression. Arguably then, this warrants an analysis of the perceptual effects of vertical interchannel crosstalk, which is considered in Chapter 5 of the present thesis.

2.3.2 LEV

LEV was described by Beranek [2010] as being 'the degree to which reverberant sound seems to surround the listener'. Toole [2009] considered it as being a sense of being surrounded by a diffuse array of sounds that are not associated with any localisable sound images. LEV has been shown to relate to energy arriving after 80 ms [Bradley and Soloudre 1995, Rumsey 2005], with increases in the amplitude of such reflections contributing to a greater sense of perceived LEV [Bradley et al. 2000, Hanyu and Kimura 2001]. In much the same way as ASW, LEV is regarded as being a positive quality of sound [Nyberg and Berg 2008].

2.3.2.1 Characteristics of LEV and Objective Measurement Methods

Bradley and Soloudre [1995] demonstrated that perceived listener envelopment is affected by the angular displacement, level and temporal distribution of the late arriving energy and further that it correlated closely to the degree of late arriving lateral energy after 80 ms (Fig. 2.9). This is somewhat similar to the lateral fraction developed by Barron and Marshall [1981] for measuring ASW, albeit the delay times vary between the two. A later study by Soloudre et al. [2003] analysed this further. Listening tests were first conducted in which the perceived LEV was measured in listening tests for gated energy bursts. Following this, objective

measurements of LEV were conducted, with the threshold delay time for the late arriving sound being both 80 and 105 ms. The results showed that, for spectral energy less than 1 kHz, measurements in which only the delays after 105 ms were considered resulted in better correlation to the results of their listening tests than did those that included delayed energy between 80 and 105 ms. It was however noted that the 105 ms measure was not consistently better than the 80 ms measure with respect to the listening tests conducted.

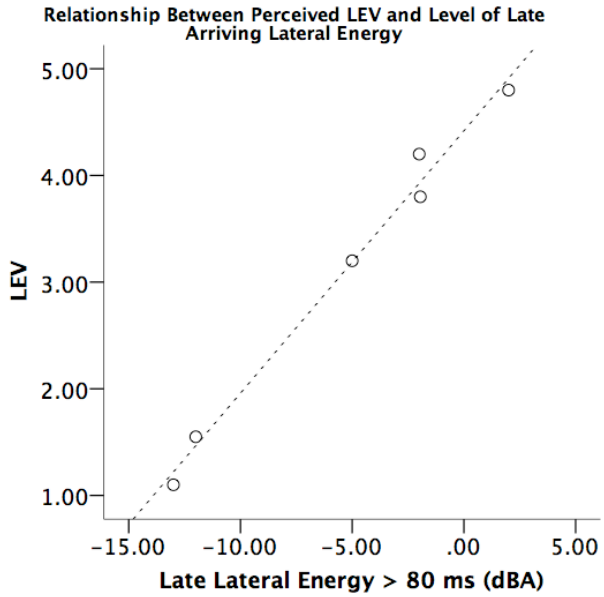

Fig. 2.9: The relationship between perceived LEV and the level of late arriving lateral energy [after Bradley and Soloudre 1995].

Perhaps the inconsistencies identified between the perceived LEV in the listening tests conducted by Soloudre et al. [2003] and those measured using late lateral energy arriving after either 80 or 105 ms is related to their use of lateral energy only. Furuya et al. [2001] considered how the perception of LEV is influenced by the direction from which the late reflections arrive. Orchestral music was presented from a series of loudspeakers arranged on the horizontal plane. The direct sound was presented from a loudspeaker in front of the listening position, with early reflections (< 80 ms) presented from L and R loudspeakers with azimuth angles of ±45°. The late arriving reflections were presented from loudspeakers positioned at 0°, ±90° and 180° on the horizontal plane. Two further late reflection loudspeakers were located at ±90° and

were elevated by 80°. All loudspeakers were 1.5 m from the listening position. Paired comparison tests were completed in which subjects were required to judge the difference in perceived LEV between each stimulus. The results of the study showed that LEV is affected not only by the arrival of late lateral energy but also by late energy arriving from both behind and above the subject. Late frontal reflections were found to have little influence. The experimental data led Furuya et al. [2001] to suggest that the exclusion of late reflections that were not lateral was 'risky' in the context of acoustical design. It was further regarded that the exaggeration of energy from a partial direction might cause unnaturalness in the room dimension. This may explain the inefficiencies in the measurements made by Soloudre et al. [2003], who only considered late lateral energy as a measure of LEV.

Hanyu and Kimura [2001] also noted some key characteristics of LEV, whilst simultaneously developing an alternative method by which it could be objectively measured. In a series of listening tests orchestral sources were presented in an anechoic chamber from loudspeakers positioned in 22.5° intervals spanning the entire horizontal plane, 2 m from the listening position. The results of their listening tests revealed the following:

1. LEV increases with increasing levels of lateral reflections.
2. Reflections arriving from the front have some influence on LEV.
3. The contribution of individual reflections to perceived LEV depends on the direction of arrival of other reflections.
4. LEV increases with adequate spatial balance in the energy of arriving reflections.

Based on their experimental data Hanyu and Kimura [2001] developed an alternative method for the objective measurement of LEV, which they called SBT_s (spatially balanced T_s, where T_s is the centre time for each arrival). Rather than consider only the effect of lateral reflections, SBT_s is based on the integration of the mutual effect of a pair of reflections. The method was found to correlate well with the results of their subjective listening tests (Fig. 2.10).

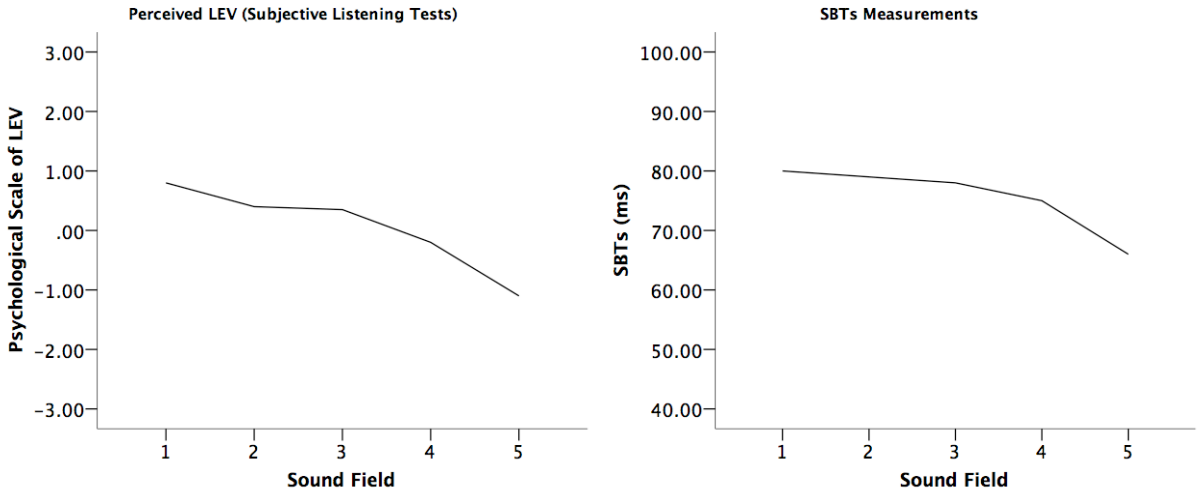

Fig. 2.10: Perceived (left) and measured (right) LEV for different sound fields from Hanyu and Kimura [2001]. Objective measurements were made using the SBTs method [after Hanyu and Kimura 2001].

2.3.2.2 The Effect of Vertical Reflections on the Perception of LEV

As was previously noted, a study conducted by Furuya et al. [2001] found that vertical reflections could contribute to the perception of LEV. In a previous study conducted by the same authors [Furuya et al. 1995] this had been examined in greater detail. Musical sources were presented to subjects from a single loudspeaker located in front of the listening position. Alongside this were presented 20 lateral and 20 vertical reflections, with ICTDs up to 200 ms, from loudspeakers located all around the listening position. The height layer of loudspeakers was elevated by 50° with respect to the horizontal plane. The amplitude of the reflected sound with respect to the direct sound varied between -3 and -12 dB, whilst the amplitude ratio of lateral to vertical reflections varied between 0.3 and 0.5. The delay time between the first vertical reflection and the direct sound was 68 ms, being 50 ms longer than that between the direct sound and the first lateral reflection. Subject's perceived sense of envelopment was found to increase as the amplitude of the vertical reflections increased relative to that of the lateral reflections. With respect to the influence of vertical reflections on the perception of LEV, it was therefore concluded by Furuya et al. [1995] that, although lateral

reflections are the primary cause of the sense of envelopment, reflections from other directions, including from vertical positions, should not be ignored.

Based on the experimental data reported by in both the Furuya et al. [1995, 2001] studies it would appear that vertical reflections do have some influence on the perception of LEV. By extension this would indicate that perceived LEV would be affected to some extent by vertical interchannel crosstalk. Despite this, it is important to consider that Lee [2006] found that the effect of horizontal interchannel crosstalk on the perception of changes in LEV was small, with the audibility index obtained from the study being 0.7/10. Therefore, as lateral reflections contribute more to perceived LEV than do vertical reflections [Furuya et al. 1995, Bradley et al. 2000], it can be argued that if the effect of horizontal interchannel crosstalk on the perception of changes in LEV is small then the same could be hypothesised for vertical interchannel crosstalk. This is considered further in Experiment Six of the present thesis.

2.3.3 Vertical Image Spread

Vertical image spread (VIS) is a spatial attribute that is in many ways conceptually similar to ASW. However, where ASW relates to an increase in the perceived horizontal width of a sound source, VIS describes a source's *vertical* width. The concept of VIS is shown in Fig. 2.11. In general, VIS has garnered little attention in the literature and, even then, the studies have not necessarily used the term directly. For example, Furuya et al. [1995] noted that while vertical reflections contribute little to the perception of ASW, they do have an influence on the perceived 'auditory size' of a sound image. They also discussed how the use of the term 'auditory size' likely related to a source's vertical spread. This therefore demonstrates some of the similarities between ASW and VIS. In either case, the perceived width of a direct sound is influenced by the presence of a reflection. In addition, the amplitude of the reflection relative to the direct sound determines the magnitude of the change in width.

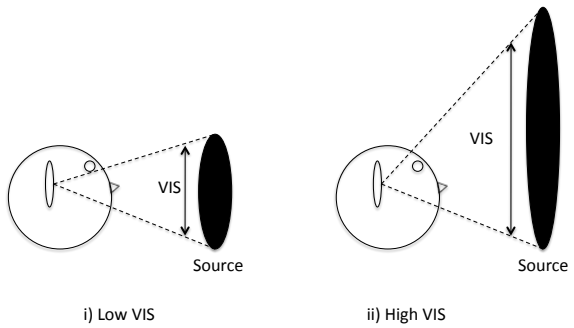

Fig. 2.11: The concept of vertical image spread (VIS).

Alongside the relative amplitudes between the direct sound and the reflection, a series of other attributes have been shown to influence the perception of VIS. One such attribute is ICTD. Furuya et al. [1995], for example, reported that degree of perceived VIS progressively increased as the ICTD increased from 10-80 ms. In relation to this, Tregonning and Martin [2015] found that, for vertically arranged loudspeakers, the primary function of ICTDs above 5 ms was to influence the perceived VIS of sound sources. It is interesting to note that Barron [1971] found that perceived ASW is only influenced by lateral delays that arrive after 5 ms of the direct sound, which arguably shows further how VIS and ASW are conceptually similar. Moreover, experiments conducted by Cabrera and Tiley [2003] demonstrated that increases in the amplitude of the direct sound could affect the perception of VIS, as could the spectrum of the signal. In that study, octave bands of noise, with centre frequencies ranging from 125 Hz – 8 kHz, were presented from a series of five vertically arranged loudspeakers, with elevation angles of ±0°, ±7.9° and ±15.6°. The test stimuli also included broadband and low pass (3 kHz) noise. Stimuli were presented to subjects at amplitudes of both 64 and 84 Phon. The study showed both that perceived VIS was greater when the stimuli were presented at 84 Phon, compared with 64 Phon, and further that the high frequency stimuli had a greater perceived VIS than did the low frequency stimuli, which tended to have a greater horizontal spread. It should be noted that Cabrera and Tiley [2003] suggested that the effect of loudness on perceived VIS was the more dominant of the two influences reported in their study.

In a recent experiment, Lee [2016] developed a method of rendering VIS based on the pitch-height effect, which he called 'Perceptual Band Allocation' (PBA). This method involves decomposing broadband signals into octave bands and routing them individually to either the main or height layers depending on the desired amount of VIS. In experiments designed to test this method, Lee [2016] noted that the perceived degree of VIS afforded by a given PBA method was dependent on the routing of certain octave bands. For example, conditions whereby the 8 kHz octave band was routed to the height layer yielded a higher degree of perceived VIS compared to when the same band was routed to the main layer. An important effect of lower frequencies was also identified, with the experimental data showing that, for conditions whereby the 8 kHz band was routed to the height layer, the perceived degree of VIS was less when the 250 and 500 Hz bands were routed to the main layer compared to the height. This result was interpreted based on the results of listening tests conducted as part of the study, which showed that these stimuli were localised notably higher for height layer only presentation compared to main layer only presentation. It should also be noted that the greatest degree of perceived VIS was for conditions whereby all bands were routed to the height layer alone. These results would seemingly indicate a dominance of certain frequencies in the perception of VIS, although this was not considered in the Lee [2016] study.

Based on the literature, it can be hypothesised that perceived VIS is one of the attributes that will be most affected by vertical interchannel crosstalk. This hypothesis is based on the following. Firstly, as has previously been mentioned the vertical interchannel crosstalk signal is essentially a vertical reflection, which has been shown to influence the perception of VIS [Furuya et al. 1995]. Furthermore, the presence of additional direct sound would naturally increase the perceived amplitude of the main channel signal, which in turn would increase its apparent vertical size [Cabrera and Tiley 2003]. In addition, unless the main and height microphones were coincident in a practical recording situation, the crosstalk signal would also be delayed with respect to the main channel signal, with increases in ICTD being shown to increase the perception of VIS [Furuya et al. 1995, Tregonning and Martin 2015]. This latter point, however, would only be relevant for spacings between the main and height layers of microphones that result in a path difference for the direct sound of at least 1.7 m (corresponding to an ICTD of 5 ms). Further to this, based on the results

in the Cabrera and Tiley [2003] study, it could be suggested that the increase would be more apparent for high frequency sources compared to low frequency sources. This however remains to be seen. It is also not clear which of the aforementioned factors would be the most dominant in causing the change in perceived VIS, with the effects of ICLD, ICTD and loudness not being compared in the literature to the knowledge of the author.

2.4 SUMMARY

The present chapter has explored the perceptual effects that vertical interchannel crosstalk might have on the main channel signal. This has been largely considered from the perspective of the effects of delayed vertical reflections. Firstly, the issue at hand was explored from the perspective of how the perceived location of the main channel signal might be affected. Then, attention was given to how the timbre might change as a result of comb filtering. Finally the effects on perceived spatial impression were considered.

In summary, studies have shown that vertical reflections can influence the perceived elevation of a sound source. The degree to which the source appears elevated is determined by the relative level between the direct sound of the source and its reflection. However, it has also been demonstrated that introducing a time delay between the direct sound and reflection will bias the perceived location in the direction of the earlier loudspeaker. The studies that reported this phenomenon concluded that this was due to the precedence effect operating for vertically arranged sound sources, however the evidence suggests that it is more of a localisation dominance effect of the earlier loudspeaker. No study to date has demonstrated that the precedence effect operates vertically in the strict sense of the term.

Time delays between the direct sound and reflection can also lead to the timbre of the source being affected by an interference effect called ‘comb filtering’, which is considered in the literature as having both positive and negative effects. Comb filtering has been shown to be dependent on both the delay time between the

direct sound and reflection, as well as their relative levels. Increased time delay will reduce the bandwidth of the comb filtering notches, whilst a reduction in the reflection level will lessen the amount of attenuation. Timbral coloration as a result of comb filtering has been shown to be more severe when the reflections are elevated, whilst reverberation has been shown to somewhat mask the effect.

Reflected sources will also influence perceived spatial impression, which comprises both ASW and LEV. ASW is due to early arriving lateral reflections (< 80 ms) and results in a broadening of the sound source. The effect relates both to IACC and the lateral fraction, both of which have been used as objective measures. LEV is the sense of being surrounded by a sound source and relates to late arriving lateral energy (> 80 ms). Among other methods, LEV has been measured from the perspective of late arriving lateral energy. An increase in the levels of early and late lateral energy with respect to the direct sound will respectively increase ASW and LEV and both are considered as being positive attributes. However, from the perspective of vertical interchannel crosstalk it is thought that no variations in ASW would be perceived, whilst changes in LEV are expected to be minimal. Instead, the primary effect is thought to be variations in perceived VIS, which is conceptually similar to ASW and relates to the perceived vertical width of a sound source. The perception of VIS has been shown to be affected by the ICLD and ICTD between the direct sound and a vertical reflection, as well as by the amplitude of the direct sound itself.

3 THE FREQUENCY DEPENDENCY OF LOCALISATION THRESHOLDS[1,2]

Two experiments are presented in this chapter. In the first (Experiment One), the frequency dependency of localisation thresholds was analysed in anechoic conditions. The second experiment (Experiment Two) considered the localisation of band-limited stimuli presented from vertically arranged stereophonic loudspeakers. The purpose of Experiment Two was to further analyse the underlying mechanisms relating to the results of Experiment One in order to try to explain them more fully.

3.1 EXPERIMENT ONE: THE ANALYSIS OF FREQUENCY-DEPENDENT LOCALISATION THRESHOLDS

As discussed in Chapter One, it has been reported in the literature that localisation judgments for tonal and band-limited stimuli are frequency-dependent in the case that they are incident from the median plane. In Chapter Zero, it was proposed that if a frequency dependency for vertical localisation existed then this might also mean that localisation thresholds would vary across the frequency spectrum. As a result of this, a method of applying the threshold whereby the direct sound in the height layer undergoes frequency-

1 Wallis, R. and Lee, H. [2015]: 'The Effect of Interchannel Time Difference on Localisation in Vertical Stereophony', *Journal of the Audio Engineering Society*, 63(10), pp. 767-776.

2 Wallis, R. and Lee, H. [2016]: 'Vertical Stereophonic Localisation in the Presence of Interchannel Crosstalk: The Analysis of Frequency-Dependent Localisation Thresholds', *Journal of the Audio Engineering Society*, 64(10), pp. 762-770.

dependent attenuation was proposed (the band reduction method). In order that such a method can be derived and verified, it is first necessary to determine whether or not there exists a frequency dependency of localisation thresholds, which is the primary aim of the present experiment.

Of further interest is the effect of ICTD on the frequency dependency of localisation thresholds. It is important to analyse this for a number of reasons. One of these is related to the existence of the precedence effect in vertical stereophony. As was discussed in Chapter Two, should the effect be shown to operate then this might indicate that vertical interchannel crosstalk would not affect the perceived location of the main channel signal, provided there was sufficient ICTD between the direct sounds present in the respective layers. Alongside this is the desire to ensure that the results are applicable to practical situations. With respect to image rendering techniques for 3D audio, which would arguably be the primary application for band reduction, the ICTD between the direct sounds in the respective layers could be varied in order to achieve a specific effect, such as the enhancement of perceived VIS as stated by Tregonning and Martin [2015]. Under such circumstances, it would be necessary to determine how the use of different ICTDs would affect the localisation threshold in order that changes could be made without the perceived location of the main channel signal being affected. This reasoning is also applicable to practical recordings for 3D audio formats, in which the ICTD would vary as a result of differing distances between the main and height layers of microphones.

From the above background, the following research questions were derived:

- Do localisation thresholds have a frequency dependency?
- How are localisation thresholds for band-limited stimuli affected by ICTD?
- Can evidence be found to support the hypothesis that the precedence effect operates in vertical stereophony?

This experiment is organised as follows. The first section describes the experimental method used in the study. Following this, the results of the experiment are presented, along with statistical analysis. Finally, the results are discussed, with a particular focus on the effects of both frequency and ICTD on the localisation threshold.

3.1.1 EXPERIMENTAL HYPOTHESIS

The first null hypothesis for this experiment is that localisation thresholds are not frequency dependent. Whether or not this is the case would arguably depend on how the perceived location of band limited stimuli changes for vertical phantom image presentation compared to main layer only presentation. This is a topic that, to the knowledge of the author, has not been considered in the literature. In Fig. 3.1, two different ways in which the perceived location of the test stimuli might be affected are shown. The implications for each of these conditions with respect to localisation thresholds are as follows. For 'Condition 1' the change in elevation for vertical phantom image presentation follows that reported for complex stimuli by Somerville et al. [1965], Barron [1971] and Barbour [2003]. In this case, the perceived elevation of the phantom image has a dependency on the ICLD. It should also be noted that there might also be frequency-dependent differences in perceived elevation between the main layer only and phantom image conditions based on the aforementioned discussions on the pitch-height effect and directional bands. This can be seen in Fig. 3.1 with the perceived difference in elevation between main layer only and phantom image presentation being greater for stimulus 1 than for stimulus 2. This would therefore indicate that less ICLD would be necessary for stimulus 2 to meet the localisation threshold compared to stimulus 1, which would result in frequency dependent localisation thresholds. On the other hand, according to Pratt [1930], Blauert [1969], Cabrera and Tiley [2003], the localisation of band-limited stimuli is a function of frequency only. If this is the case then, hypothetically, how the stimuli are presented to subjects should not have an effect on the perceived location of the test stimuli. This can be seen in Condition 2 of Fig 3.1. In this example, even though stimulus 1 is perceived as being elevated with respect to stimulus 2, the difference in perceived elevation between the

main layer only and phantom image conditions is the same. Therefore, if localisation phenomena such as the pitch-height effect are maintained when band-limited stimuli are presented as vertically oriented phantom images then localisation thresholds would not be expected to vary across the spectrum.

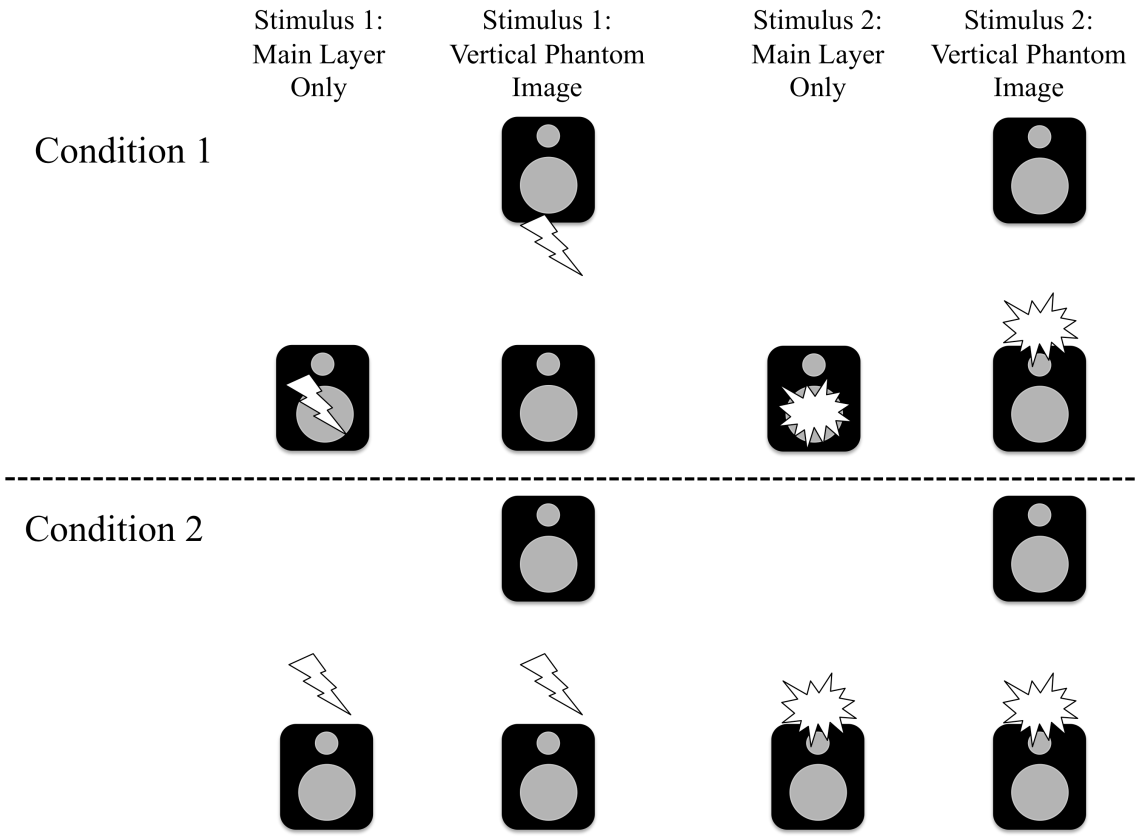

Fig. 3.1: Demonstration of how the frequency-dependency of localisation thresholds might depend on the effect of presentation method on perceived source elevation.

In addition to the above, a further null hypothesis for the present experiment is that ICTD will not have an effect on the localisation thresholds. With respect to the literature, the studies conducted by both Lee [2011] and Stenzl et al. [2014] reported that delays in the range of 0-10 ms had little effect on the localisation thresholds for musical sources. Further, neither study found any evidence to support the operation of the

precedence effect for vertical stereophony. Based on this result, and the discussion in Chapter Two that no study has of yet fully demonstrated the effect in the median plane, it is hypothesised that a similar observation will be made in the present study. What is interesting, however, is that neither of the aforementioned studies showed any localisation dominance effect, which has been heavily reported in the literature [Somerville et al. 1965, Blauert 1971, Litovsky et al. 1997, Tregonning and Martin 2015]. A potential reason for this with respect to the Lee [2011] study may be due to the decision to not time align the loudspeaker layers. Based on the dimensions of the experimental setup in that study this means that, for the 0 ms ICTD condition, the direct sound from the main layer would have arrived at the listening position 0.7 ms before that from the height layer. Had time alignment been done in that study then it is arguable that a localisation dominance effect might have been observed. Based on this discussion, it is hypothesised at the very least that less level reduction would be necessary in the case of a delayed height channel, compared to when no delay is present. It should be noted however that Stenzl et al. [2014] did time align their loudspeakers layers and equally showed no localisation dominance effect.

3.1.2 EXPERIMENTAL DESIGN

3.1.2.1 Physical Setup

Fig. 3.2 shows the physical setup used for the experiment, which was conducted in the anechoic chamber at the University of Huddersfield. The experiments utilised two Genelec 8040A loudspeakers, which were positioned as follows. The lower loudspeaker (the main layer) was positioned 1.2 m above the ground, 1.8 m away from, and directly in front of, the listening position. The upper loudspeaker (the height layer) was located directly above the main layer, at a distance of 1 m, forming a 30° elevation angle to the listening position. Appropriate time and level alignment was applied to the main layer, with respect to the height layer, in order to compensate for the differences in distance between each loudspeaker layer and the listening position. An acoustically transparent curtain was positioned between the listening position and the

loudspeakers in order to obscure the nature of the test setup from subjects. The ear height of subjects was aligned to the centre point between the woofer and tweeter on the main layer loudspeaker using a height-adjustable chair.

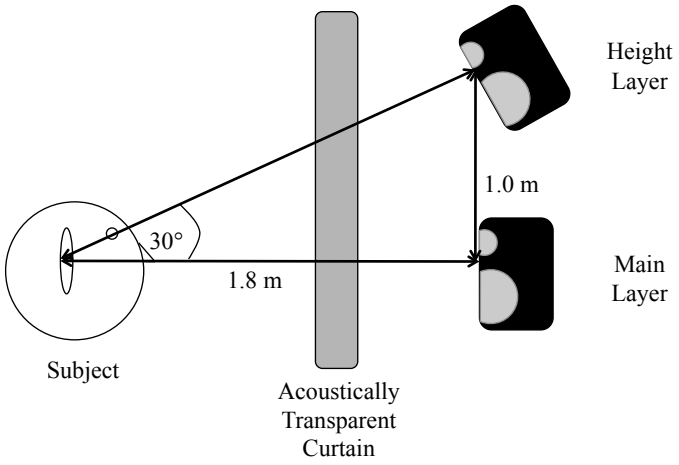

Fig. 3.2: Physical setup used for Experiment One.

3.1.2.2 Test Stimuli

The test stimuli used for the experiment were continuous octave bands of pink noise. Octave bands were chosen based on the results of the Cabrera and Tiley [2003] study, which showed that localisation judgments for octave bands of noise were frequency dependent. Octave bands were therefore considered as being ideal in that they were narrowband enough for frequency-dependent localisation, however were also broadband enough that a wide frequency range could be analysed without the need for excessive numbers of test stimuli. The centre frequencies of the octave bands ranged from 125 Hz - 8 kHz. They were created by brick wall filtering broadband pink noise using an FFT filter. An additional broadband pink noise source was also tested.

Each stimulus was ten seconds in duration, which included a one second fade in/out. Stimuli were presented to subjects as vertically oriented phantom images, with the height layer delayed with respect to the main by 0, 0.5, 1, 5 and 10 ms. The delay times were chosen for a number of reasons. One of these related to the aim of analysing the operation of the precedence effect in the median plane. Blauert [1997] stated that the threshold for the precedence effect in horizontal stereophony was 1.1 ms. Therefore, the ICTDs above this threshold were chosen in order to analyse the operation of the effect, whilst those below were chosen to determine if any effects of summing localisation (e.g. localisation dominance) could be observed. Additionally, in the context of practical recordings for 3D audio formats, the delay times represented varying distances between the main and height microphone layers. 0 ms, for example, is representative of a coincident technique, whilst 10 ms corresponds to a mic spacing that results in a path difference between the direct sound arriving at each layer of 3.4 m, which was considered as being a likely maximum in practical recording situations.

In total there were 56 stimuli (eight frequencies with five ICTDs). Each stimulus was calibrated to 75 dB LAeq at the listening position when presented from the main layer only. The amplitude of the stimulus when presented as a phantom image was dependent on the amplitude of the height layer relative to the main, which was to be varied by subjects as described in Section 3.1.2.4.

3.1.2.3 Subjects

12 subjects, comprising staff and both postgraduate and final year undergraduate students from the University of Huddersfield's Music Technology courses, participated in the listening tests. These subjects were chosen because of their critical listening experience in spatial audio, making them better suited than more naïve subjects to determine the subtle localisation differences caused by vertical interchannel crosstalk. They all reported normal hearing.

3.1.2.4 Test Method

The basic methodology for the experiment was similar to that used by Lee [2011]. Subjects were presented with a reference stimulus, being a given source presented from the main layer only, and a series of test stimuli, being the same stimulus presented as a vertically-oriented phantom image with a test ICTD applied to the height layer. For each test stimulus, subjects were required to identify the minimum necessary attenuation of the signal in the height layer for the resultant phantom image position to match that of the reference (i.e. the localisation threshold).

The threshold detection method used for the experiment was the method of adjustment (MOA). This is an indirect scaling method that requires subjects to reduce the amplitude of a stimulus until it is equivalent to that of a reference [Bech and Zacharov 2006]. Cardozo [1965] asserted that the principal application of MOA is in situations whereby stimuli differ from one another by more than one attribute. Bech [1998] further suggested that subjects are better able to focus on a single attribute in the face of others that vary when using MOA. This is directly applicable to the present experiment. Although subjects were tasked with identifying the differences in perceived elevation between the test and reference stimuli, the use of ICTD would invariably result in timbral differences between the two (see Chapter Two for a full review). This is perhaps why Lee [2011] chose to use the MOA method in the analysis of localisation thresholds for musical sources, although the reasons for his decision are not discussed in his paper. It was further considered that the timbral changes alongside the location shifts rendered other methods, such as the two alternative forced choice method [Yogan and Stocker 2014], inadequate. Bech [1998], for example, reported subject's difficulty in distinguishing between test and reference stimuli that differed in loudness, timbre and spaciousness when utilising an adaptive form of this method.

The graphical user interface for the experiment was created using Max/MSP (Fig. 3.3). The interface split the experiment into eight subtests, with each subtest focusing on a single frequency band. Within each subtest was the reference and five test stimuli (labeled 'A', 'B', 'C', 'D', and 'E'). For each of the test

stimuli, subjects were presented with a slider, with values ranging from 0 to 100 in increments of 1. The slider controlled the amplitude of the height layer as follows. Slider values were first divided by 100 to give 'x', which lay between 0 and 1. The amplitude of the height layer was then multiplied by x and therefore decreased with decreases in the slider value. A slider value of '100' resulted in 0 dB ICLD between the main and height layers. A value of '0' resulted in the height layer having zero amplitude (-∞ dB ICLD). Slider values were converted into decibel values internally. The decibel values were not shown to subjects during any part of the test. Subjects were also unaware that they were controlling the amplitude of a loudspeaker. The amplitude of the main layer was kept constant throughout each test.

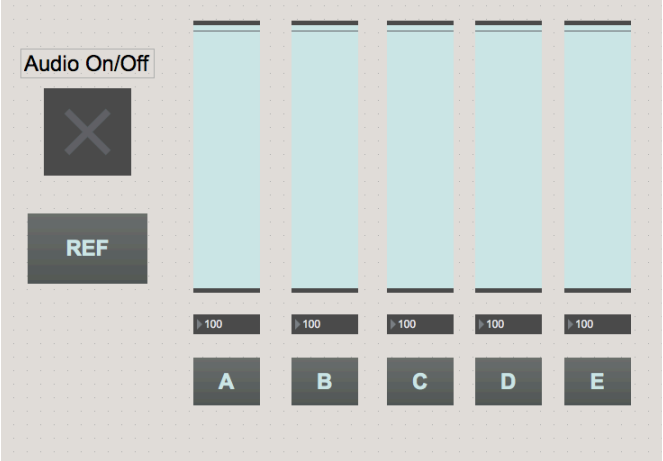

Fig. 3.3: Max/MSP interface used for Experiment One.

During the tests, the heads of subjects were not fixed, however they were strictly instructed to face forwards, keeping their head still and using only their eyes to look at the test interface. A guide point for the ear height and distance was placed on the right-hand side of each subject to help maintain the correct listening position throughout. Prior to the start of each test, all subjects sat a supervised practice, which utilised a speech source, in order to ensure that the instructions were understood. The order of subtests and the stimuli within each subtest were randomised for each subject. The whole test was completed in a single sitting, which lasted around 20 minutes.

3.1.3 DATA ANALYSIS AND RESULTS

Levene and Shapiro-Wilk tests were first conducted, using the SPSS software, in order to determine the suitability of the collected data for parametric statistical analysis. The results of the Levene test showed homogeneity of variance for all frequencies, whilst the Shapiro-Wilk test showed that not all scores in each condition featured normal distribution. This therefore meant that the assumptions of Analysis of Variance (ANOVA) were violated. For these reasons, non-parametric tests were chosen for the statistical analysis.

3.1.3.1 The Effect of ICTD

Fig. 3.4 shows the median localisation thresholds for each frequency at each ICTD. The medians have been plotted with notch edges. According to both McGill et al. [1965] and Kirchner [2001], notch edges represent a way of determining confidence intervals for non-parametric data, with an overlap between notches indicating that pairs of stimuli are not significantly different from one another with 95% confidence. Notch edges can be calculated as follows [Kirchner 2001]:

$$\text{notch edges} = \text{median} \pm 1.57(\text{interquartile range}/\sqrt{n}) \quad (3.1)$$

Where n is the number of subjects.

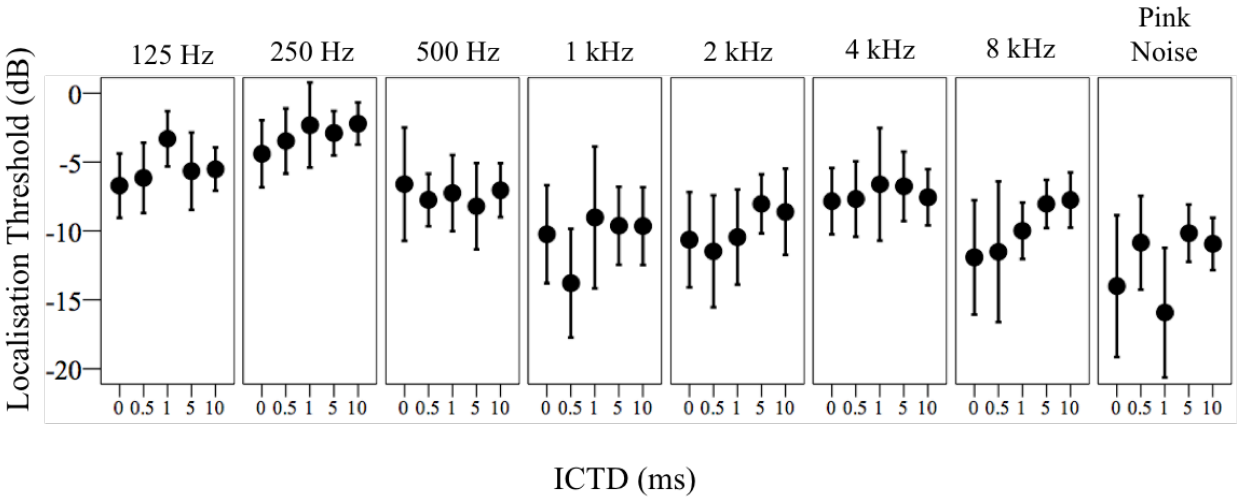

Fig. 3.4: Medians and associated notch edges of the experimental data arranged to compare the localisation thresholds for each octave band at each ICTD.

Based on the overlapping notch edges in Fig. 3.4, it would appear that the effect of ICTD was not significant on the localisation thresholds obtained in the study. In order to analyse this further, a Friedman test was conducted; the statistical power was judged based on the critical p value of 0.05. The results of this analysis showed that the effect of ICTD was not significant for any of the octave bands, with the exception of 8 kHz ($p = 0.001$). In addition to this, the effect size (Kendall's W) was less than 0.5 for all octave bands (including for 8 kHz), which indicates that the effect of ICTD was not large.

In order to identify which pairs of ICTD were significantly different from one another for the 8 kHz band, a Wilcoxon test was conducted. As such analysis necessitated the performance of multiple pair-wise tests, it was decided to use the Bonferroni correction in a bid to reduce any type-I errors (detecting an effect that is not present) [Simner 1986]. It should be noted, however, that Perneger [1998] urged caution when using this method, as the reduction of type-I errors can result in an increase in type-II errors (failure to detect an effect that is present [Lieberman and Cunningham 2009]). According to the results of the Wilcoxon test, the threshold obtained for 10 ms was significantly higher than that for 0 ms ($p = 0.03$). Despite this, it is clear from Fig. 3.4 that there is heavy overlap between the notch edges for this pair of stimuli. When considering

this, along with the effect size (0.413) it seems reasonable to deduce that differences among the different ICTDs for the 8 kHz band are negligible. Overall it can therefore be concluded that the effect of ICTD on localisation threshold was not significant for any stimulus within the present experiment.

3.1.3.2 The Effect of Frequency

In the previous section, it was shown that the effect of ICTD on the localisation thresholds obtained in the experiment was not significant. It is therefore possible to combine all the data for each of the frequency bands, rather than consider each ICTD individually. The median localisation thresholds for each frequency, with ICTDs amalgamated, are plotted with notch edges in Fig. 3.5. From this, a frequency dependency of localisation thresholds is apparent. It can be seen that the threshold was the highest at low frequencies (-5.3 dB at 125 Hz and -3.03 dB at 250 Hz), falling gradually to between -9 and -10.5 dB as the frequency increased beyond 1 kHz. The threshold was the lowest for the broadband source (-11.56 dB), whilst there was also a small 'peak' for 4 kHz (-6.96 dB).

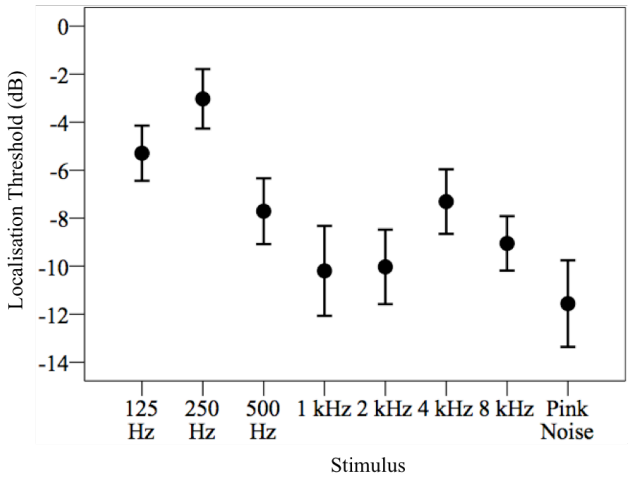

Fig. 3.5: Median localisation thresholds for each frequency band, with results for individual ICTDs amalgamated, plotted with notch edges.

Consideration of the notch edges alone suggests that the effect of frequency on localisation threshold was significant. A Friedman test was conducted in order to analyse this further (critical p value = 0.05). The results of this analysis showed a significant effect of frequency ($p < 0.001$). The results of a Wilcoxon test, which was conducted to analyse which pairs of frequencies were significant from one another, can be seen in Table 3.1. From this analysis the following can be concluded. Firstly, the threshold for the 250 Hz octave band was significantly higher than those for all the other stimuli tested. Further, the threshold for 125 Hz was significantly higher than those for 1, 2 and 8 kHz, as well as for the broadband pink noise. Also, the thresholds for 1, 2 and 8 kHz were not significantly different from one another, whilst 4 kHz was significantly higher than 1 and 2 kHz. It should be noted that, although the Wilcoxon test results suggest other significantly different pairs, including between 4 and 8 kHz and between 8 kHz and the pink noise, the overlapping notch edges suggest that these are likely to be type-II errors as a result of the use of the Bonferroni correction. Overall, the results of the Friedman and Wilcoxon tests, along with the lack of overlap of notch edges shown in Fig. 3.5, shows that localisation thresholds vary across the frequency spectrum, with the low frequencies needing significantly less level reduction compared to the mid-high frequencies.

Table 3.1: Wilcoxon Test Results for The Effect of Frequency (Bonferroni Correction Applied).

	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	Broadband
125 Hz	X	1.000	0.112	0.000	0.000	0.056	0.000	0.000
250 Hz	1.000	X	0.000	0.000	0.000	0.000	0.000	0.000
500 Hz	0.112	0.000	X	0.000	0.000	1.000	0.000	0.000
1 kHz	0.000	0.000	0.000	X	1.000	0.000	1.000	0.056
2 kHz	0.000	0.000	0.000	1.000	X	0.000	1.000	0.028
4 kHz	0.056	0.000	1.000	0.000	0.000	X	0.000	0.000
8 kHz	0.000	0.000	0.000	1.000	1.000	0.000	X	0.028
Broadband	0.000	0.000	0.000	0.056	0.028	0.000	0.028	X

3.1.4 DISCUSSION

The results of the present experiment have shown primarily that localisation thresholds are frequency dependent. For the low frequency stimuli, the thresholds were found to be reasonably high, with less than 6 dB ICLD necessary for both the 125 and 250 Hz bands. In addition to this, the thresholds for the mid-high frequencies (being 1, 2 and 8 kHz) were significantly lower than those for the low frequencies, with ICLDs between -9 and -11 dB being necessary. Also, the thresholds for the 500 Hz and 4 kHz bands were similar to one another, being in the range of -6 to -7 dB, whilst the broadband source required the greatest amount of ICLD (-11.5 dB). Therefore, the null hypothesis raised in Section 3.1.1, that localisation thresholds do not have a frequency dependency, can be rejected. On the other hand, because there was no significant effect of ICTD on any of the localisation thresholds obtained in the study, the null hypothesis that ICTD does not affect localisation thresholds can be accepted. This result is in positive agreement with the localisation thresholds obtained by Lee [2011] for musical sources with ICTDs up to 10 ms. Overall, these results might suggest that, in terms of localisation, vertical interchannel crosstalk would be more disturbing to the main channel signal for the mid-high frequencies than they would be for the low frequencies. Moreover, in the context of microphone array configuration, the amount of attenuation of direct sound necessary in the height microphone layer would appear to be consistent irrelevant of the spacing between the main and height layers, at least for path differences between the direct sound arriving at each layer up to about 3.4 m (i.e. the ICTD of 10 ms corresponds to a path difference of 3.4 m).

3.1.4.1 Explanations For The Frequency Dependency of Localisation Thresholds

In Section 3.1.1 it was hypothesised that the frequency dependency of localisation thresholds would be related to frequency-dependent differences in perceived elevation between the main layer only and vertical stereophonic conditions. Following the experiment, the author conducted informal listening exercises in which this hypothesis was scrutinized. For the broadband pink noise source, a notable increase in perceived

elevation was observed when source presentation shifted from main layer only to vertical phantom image; this being in line with the data reported by Barbour [2003]. Therefore, vertical interchannel crosstalk can be said to have affected the perceived location of the main channel signal for this stimulus.

Attempts to explain the localisation threshold results for the broadband pink noise source were made as follows. Since the primary cue for elevation perception is known to be the spectral filtering of the pinnae in the 4-10 kHz range [Hebrank and Wright 1974a], it was first considered how the ear input spectra are affected by the presence of the height layer. In Fig. 3.6, the difference in spectral energy (i.e. the delta spectrum) between the main layer only and i) the height layer only, ii) the vertical phantom image condition with 0 dB ICLD and iii) the vertical phantom image condition with -11.5 dB ICLD (i.e. the localisation threshold) has been shown. The HRIRs used for this exercise were taken from the MIT's KEMAR dummy head database [Gardner and Martin 2000]. The creation process for the delta spectra was as follows. Firstly, the main and height layer only conditions used the HRIRs taken at 0° azimuth for 0° (main layer) and 30° (height layer) elevation respectively. Additionally, the 0 dB ICLD condition was created by summing together the HRIRs for the main and height layers, whilst the process for the localisation threshold conditon was the same, albeit that the height layer HRIR was reduced by 11.5 dB before being combined with that for the main layer. FFTs were then conducted on each of the four conditions individually (main layer only, height layer only 0 dB ICLD and localisation threshold) with the resultant spectrum for the main layer only condition being subtracted from those for the other three conditions to give the three delta spectra shown in Fig. 3.6. It is important to note that the subtraction of the main layer was only undertaken after the signals had been converted into spectra. Had this been done before the FFTs were conducted then the resultant spectrum for the 0 dB condition would have been identical to that for the height layer only (i.e. the main layer would have been added to the height layer and then subtracted again). With respect to Fig. 3.6, any point where the line falls below 0 dB represents dominance of the spectral energy for the main layer only and vice versa.

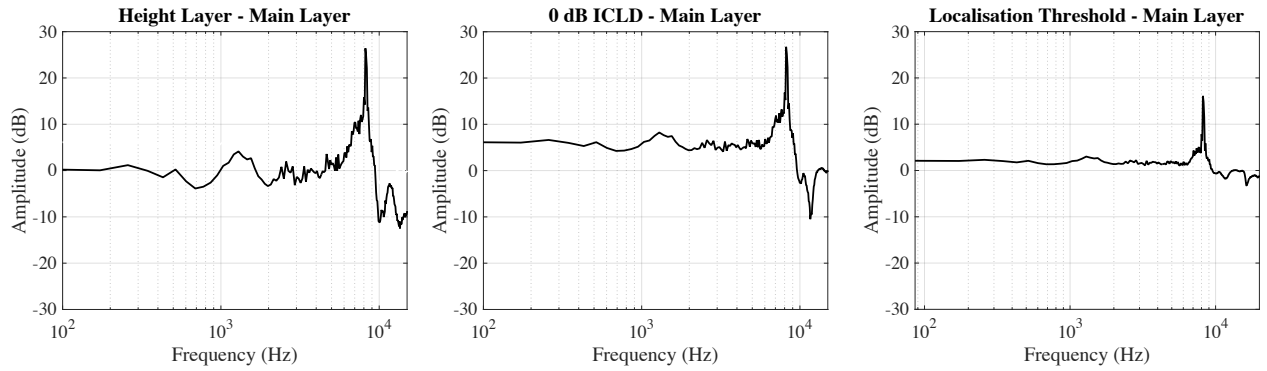

Fig. 3.6: Difference between the HRTFs of (i) height layer only, (ii) height and main layers with 0 dB ICLD and (iii) height and main layers combined with the height layer level reduced by 11.5 dB (localisation threshold), to that of the main layer only.

From the delta spectrum for the main layer subtracted from the height layer, it can be seen that the predominant difference in energy between the two layers is a peak (height layer dominant) in the range of 7-9 kHz, a region that is associated with localisation above the subject [Blauert 1969, Hebrank and Wright 1974a]. Given that Chun et al. [2011] reported that amplitude increases in the 4.5-9 kHz region resulted in increases in perceived source elevation, it is reasonable to conclude that the height layer dominance between 7 and 9 kHz would result in sources presented from the height layer being elevated with respect to those presented from the main layer. Based on these discussions, the threshold results for the broadband pink noise can be explained as follows. When a source is presented as a vertically oriented phantom image, the resultant ear input spectrum will depend on the amplitude of the height layer relative to that of the main. For phantom image conditions whereby the ICLD is small, the contribution of each layer to the resultant spectrum will be somewhat similar. As a consequence of this, the phantom image condition will feature more energy in the 7-9 kHz region compared to the main layer only condition. This is shown in the '0 dB ICLD – main layer' delta spectrum in Fig. 3.6, where the difference in energy is around 26 dB, and will manifest in the phantom image condition being elevated with respect to the main layer only condition. However, as the ICLD increases, the main layer becomes more dominant in determining the resultant ear input spectrum, which

means that the differences between the phantom image and main layer only conditions in the 7-9 kHz region would decrease. This can be seen in the delta spectrum for 'localisation threshold – main layer', where it is interesting to note that there remains around 15 dB more energy in this region for the phantom image condition compared to the main layer only. This analysis would therefore indicate that the localisation threshold for the broadband pink noise was perceived to have been met at the point where the difference in energy in the 7-9 kHz region between the main layer only and phantom image conditions was not sufficient to be interpreted as a difference in perceived elevation. The reason that an exact energy match was not necessary is open to further study. It should be noted, however, that the ear input spectra are dependent on the subject and that it is therefore difficult to generalise HRTF characteristics. Based on this, it is apparent that further study is necessary, using measured HRTFs of different subjects, before the importance of the 7-9 kHz region in particular in determining the localisation threshold can be confirmed.

The balance of spectral cues provided by the main and height layers respectively demonstrates clearly how the localisation threshold for the broadband pink noise was derived. However, this arguably does not explain the frequency-dependency of localisation thresholds observed in the present experiment. Indeed there are two issues with the hypothesis. Firstly, of all the octave bands tested, only the 8 kHz octave band contained spectral energy in the 7-9 kHz region. Furthermore, the informal listening conducted by the author did not identify a noticeable increase in perceived elevation for the octave band stimuli when source presentation changed from main layer only to vertical phantom image. Instead, the most salient difference was an increase in perceived VIS, which was particularly noticeable when the ICLD was 0 dB. It could therefore be the case that subjects interpreted differences in perceived VIS between the test and reference conditions as being differences in perceived elevation. This seems reasonable given that the test conditions essentially required subjects to directly compare the perceived elevation of a very wide image to that of one that was much narrower. In an experiment conducted by Furuya et al. [1995], it was identified that the perception of VIS was related to the amplitude of the height layer relative to the main. Therefore, in the present study, increases in ICLD would have decreased the difference in VIS between the main layer only and phantom image

conditions, which ultimately would have resulted in the localisation threshold being met for the band-limited stimuli.

Based on the above observations, the significant effect of frequency on localisation threshold might be explained by the differences in perceived VIS between the test and reference stimuli with changes in frequency. This hypothesis is illustrated in Fig. 3.7. For 'Condition 1' the influence of the height channel on the increase in perceived VIS is small since the 'reference' inherently has a large spread, necessitating a small amount of reduction in the height channel level (high localisation threshold). For 'Condition 2', however, the change in VIS is considerably larger, requiring an increased amount of level reduction (low localisation threshold). From the results of the current study, the following might be inferred. Firstly, based on its non-significant effect on localisation threshold, ICTD has little effect on the perceived VIS of octave bands presented from vertically arranged stereophonic loudspeakers in front of the listening position; this would explain the non-significant effect of ICTD. Additionally, the increase in VIS from single loudspeaker to vertical phantom image presentation is significantly greater for the 1, 2 and 8 kHz octave bands than for the 125 and 250 Hz bands. It should be noted, however, that this hypothesis would require further study.

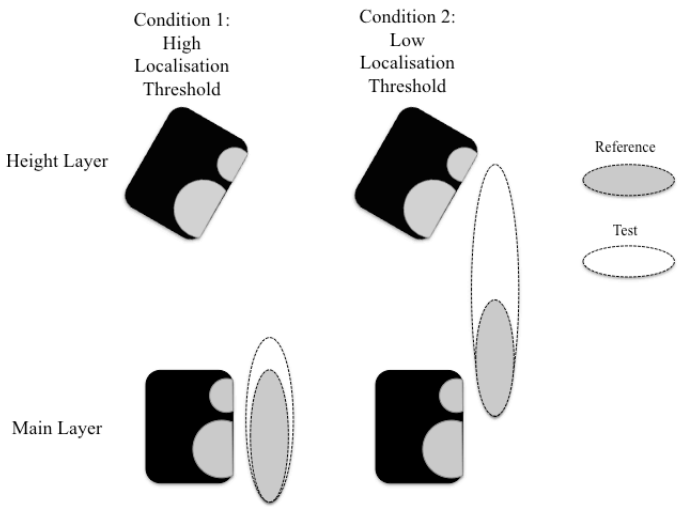

Fig. 3.7: Illustrations to explain how variations in vertical image spread might affect localisation thresholds.

3.1.4.2 The Precedence and Localisation Dominance Effects

A key aim of the present experiment was to determine whether any evidence could be found to support the existence of the precedence effect in the median plane. Further, it was hypothesised that there would at least be some localisation dominance effect of the earlier loudspeaker, which would mean that less level reduction would be necessary in the height layer when a delay was present. With respect to the octave band stimuli, such an analysis is thought to not be relevant. This is due to the informal listening exercises, which identified that localisation judgments for these stimuli are generally consistent irrelevant of presentation method, with time and level differences between the loudspeakers having little perceptual effect (this is examined further in Experiment Two). However, both the localisation dominance and precedence effects can be analysed for the broadband pink noise source, whose perceived elevation was affected by vertical interchannel crosstalk.

With respect to the broadband pink noise, the experimental data showed that each test stimulus required ICLD before its position perceptually matched that of the reference; there was no condition whereby ICTD alone was sufficient. This suggests that the precedence effect does not operate for vertically arranged stereophonic loudspeakers. If it had, then ICLD would not have been required with sufficient ICTD. It could be argued that the range of ICTDs tested was not adequate to test for the effect, as the threshold in the median plane could perhaps be higher. This seems unlikely however, with Lee [2011] testing with ICTDs up to 50 ms and equally finding no evidence for the effect. It should also be noted that, as this was not a pure localisation study, it is only suggested indirectly that the precedence effect did not operate.

In terms of the localisation dominance effect, a series of interesting results were obtained. Firstly, the localisation thresholds for the broadband pink noise were not significantly different from one another for ICTDs up to 10 ms, which agrees with Lee's [2011] result for musical sources. This indicates that increased spacing between the main and height layer of microphones would have no effect on the localisation threshold. However, an unexpected result was observed when the ICTD was 0 ms. For this condition, the threshold obtained was not significantly different compared to those conditions in which an ICTD was present. This

result would seemingly indicate one of the following. Firstly, dominance of the earlier loudspeaker has no effect on the perceived location of vertically arranged phantom images; this seems unlikely given the plethora of research that suggests the opposite [Somerville et al. 1967, Blauert 1971, Litovsky et al. 1997, Tregonning and Martin 2015]. Alternatively, it may be that the perceived location of the stimulus is not the only factor that determines the localisation threshold. This somewhat agrees with the aforementioned hypothesis regarding VIS for the octave band stimuli. Incidentally, notable increases in VIS for the broadband pink noise source were also observed for the phantom image conditions compared to main layer only presentation.

An example of the above hypothesis is as follows. The presence of an ICTD might have resulted in the physical location of the broadband pink noise moving closer to the earlier loudspeaker due to the localisation dominance effect. Despite this, the source may still have appeared to be in a different location to the reference due to differences in perceived VIS. The degree of perceived VIS is primarily related to the ICLD [Furuya et al. 1995], although ICTD has some small effects [Tregonning and Martin 2015]. It therefore stands to reason that, irrelevant of any localisation dominance effects, the localisation thresholds would be consistent whether an ICTD was present or not. Should this be the case, then it might render differences in VIS between the test and reference sounds as being the primary factor in determining whether or not the localisation threshold has been met for complex sources, as opposed to the balance of spectral cues hypothesis postulated in Section 3.1.4.1. This, however, would need to be studied further, as a reduction in the height layer amplitude would simultaneously reduce the VIS and increase the influence of the main layer on the resultant ear input spectra as shown in Fig. 3.6. In addition, it is not yet known whether a localisation dominance effect did not operate for definite in the present experiment.

3.1.4.4 Practical Implications and Future Works

It was previously mentioned that localisation thresholds might be as much related to differences in perceived VIS between the test and reference stimuli as they are to differences in perceived elevation. Additionally, in Section 3.1.1 it was mentioned that no study had considered the localisation of band-limited stimuli in the median plane when sources were presented as vertically arranged phantom images, as opposed to from real sources (although such an analysis was performed informally following the present experiment). Conducting such an experiment formally would make it possible to further understand the results of the present study and may help to establish the most salient factors in the determination of localisation thresholds. This is explored in Experiment Two.

The results also demonstrated that localisation thresholds have a frequency dependency. It is therefore possible to suggest, at least in theory, that a band reduction method would be suitable in the application of localisation thresholds. This would involve applying the individual localisation thresholds obtained for each band to a complex signal, such as music, rather than applying level reduction across the whole frequency spectrum (blanket reduction). Although this approach may be difficult to execute within a practical recording situation, there would certainly be implications for 3D mixing using discrete sound sources. However, the present study has been conducted in anechoic conditions and, as a result, does not represent a 'natural' listening environment. In order that a band reduction method can be suitably developed it is first necessary to see how localisation thresholds vary in such conditions, which is considered in Experiment Four.

In addition to the above, it would be worth examining if the localisation shift effect of vertical interchannel crosstalk could be eliminated through the manipulation of selected frequency bands that are perceptually dominant. The analysis in Section 3.1.4.1, for example, indicates that differences in spectral energy in the 7-9 kHz region between the main layer only and phantom image conditions might be a key mechanism for determining whether or not the localisation threshold has been met for broadband pink noise. In addition, as mentioned previously, an experiment conducted by Chun et al. [2011] demonstrated that musical sources

presented from stereophonic loudspeakers on the horizontal plane could appear perceptually elevated by up to 20° when the signal underwent directional band boosting in the 4.5-9 kHz range. Based on this, directional band reduction could hypothetically be applied to decrease perceived elevation. This could be an alternative method for preventing the height channel signal from affecting the perceived location of the main channel signal and would have implications for the rendering of 3D images. This is considered in Experiment Five.

The non-significant effect of ICTD on the localisation threshold, as well as the absence of the precedence effect in vertical localisation, has implications for the design of microphone configurations for recording in 3D audio formats. In the context of preventing vertical interchannel crosstalk from affecting the localisation of the main channel signal, it is clear that there should be a focus on the attenuation of direct sounds in the height microphone layer, with the spacing between layers being less of an issue. This would make cardioid microphones the ideal choice for the height layer, as they would be able to provide the necessary attenuation of direct sounds to limit the location-based effects of vertical interchannel crosstalk. Omnidirectional microphones would be less appropriate in this context.

3.1.5 CONCLUSION

The present experiment carried out an analysis of how vertical interchannel crosstalk varies across the frequency spectrum. Seven octave bands of pink noise with centre frequencies ranging from 125 Hz to 8 kHz, as well as broadband pink noise, were presented to subjects as phantom images from vertically arranged stereophonic loudspeakers. The height layer was delayed with respect to the main by 0, 0.5, 1, 5 and 10 ms. Subjects were required to identify the minimum amount of attenuation necessary in the height layer for the resultant phantom image position to match that of the same stimulus played from the main layer only (the localisation threshold).

The results of the study showed that the effect of frequency on the localisation thresholds obtained was significant. The thresholds were the highest at low frequencies (-5.3 dB at 125 Hz and -3.03 dB at 250 Hz),

falling to between -9 and -10.5 dB as the frequency increased beyond 1 kHz. It was hypothesised that the primary reason for this was frequency-dependent variations in perceived VIS between the vertical phantom image and main layer conditions. In addition, the threshold for the broadband pink noise source was the lowest of all thresholds (-11.56 dB). This result was interpreted in terms of the dominance of the height layer on spectral cues, particularly in the 7-9 kHz region, with it being demonstrated that increases in ICLD resulted in a spectrum more similar to that for main layer only presentation.

The effect of ICTD was not significant on the localisation thresholds obtained for any of the test stimuli. Additionally, ICLD was always necessary for the broadband pink noise source; there was no condition whereby ICTD alone was sufficient. This result indirectly indicates that the precedence effect is not a feature of vertical stereophony, which agrees with the literature. However, localisation dominance effects were also not observed for the broadband pink noise source. This suggests that, rather than there being no localisation dominance effect, that it is not the most important factor in determining the localisation threshold, with differences in VIS arguably being more salient.

3.2 EXPERIMENT TWO: THE EFFECT OF INTERCHANNEL TIME DIFFERENCE ON LOCALISATION IN VERTICAL STEREOPHONY

In Experiment One, it was established that localisation thresholds have a frequency dependency in anechoic conditions. However, informal listening exercises indicated that the perceived elevation of the octave band stimuli appeared to be little affected when source presentation shifted from main layer only to vertical phantom image. This indicates that differences in perceived elevation were not the most salient factor in determining the localisation thresholds for the octave band stimuli. Instead, it was suggested that frequency-

dependent differences in perceived VIS between the main layer only and phantom image condition was the primary mechanism. The aim of the present experiment is to conduct formal localisation experiments on the test stimuli used in Experiment One in order to validate the aforementioned observations.

In addition, with respect to the broadband pink noise source, the data in Experiment One showed no evidence to support the existence of the precedence effect in vertical stereophony. However, the experiment was not a pure localisation study and, as a result, the effect was only considered indirectly. Therefore, it should not be concluded that there is no evidence to support the existence of the precedence effect in vertical stereophony until a pure localisation study has been undertaken. This is also true of the localisation dominance effect, which was also not observed in Experiment One however was presumed to have featured based on the literature.

The analysis undertaken in the present experiment is also considered as being particularly important with respect to the aims of the thesis. For both the broadband pink noise and octave band sources, an analysis of the correlation between the perceived elevation of the test stimuli and the localisation thresholds obtained will help to establish the most salient attributes in the determination of localisation thresholds. If it can be shown that there is no correlation then this would arguably help to validate the VIS hypothesis. At the very least, the hypothesis relating to the balance of spectral cues for the broadband pink noise source could arguably be rejected. Conversely, if there is some correlation then a new hypothesis may need to be developed to explain the frequency-dependency of localisation thresholds.

Alongside being able to further explain the results of Experiment One, the present experiment also provides further fundamental understanding about the localisation of band-limited stimuli in the median plane. To the knowledge of the author, no study has considered the median plane localisation of band-limited stimuli when presented as vertically arranged phantom images, with previous studies, including those of Pratt [1930] and Blauert [1969], considering localisation when sources were presented from single loudspeakers only. It is

therefore unclear if such phenomena as the pitch-height effect are maintained for phantom source presentation and if not then how localisation judgments are affected.

From the above background the following research questions were derived:

- Is there any correlation between the perceived elevation of the test stimuli and the localisation thresholds obtained in Experiment One?
- What is the effect on perceived elevation when delays are applied to the height layer and can any evidence be found to support the operation of the precedence effect or localisation dominance in the median plane?
- Is the frequency dependency of median plane localisation maintained when octave band stimuli are presented as phantom images from vertically arranged stereophonic loudspeakers?

3.2.1 EXPERIMENTAL HYPOTHESES

The first null hypothesis for this experiment is that the pitch-height effect is not maintained when band-limited stimuli are presented as vertically oriented phantom images. Informal listening undertaken by the author following Experiment One suggested that there was not a noticeable increase in the perceived elevation of the octave band stimuli when source presentation shifted from main layer only to vertical phantom image. This would seemingly indicate that the frequency dependency of median plane localisation is not influenced by whether the source is presented as a monophonic or stereophonic image. Additionally, if vertical localisation judgments for octave bands are dependent solely on frequency, as was reported by Cabrera and Tiley [2003], then how said octave bands are presented to subjects should arguably not have an influence on perceived elevation. It is therefore hypothesised that localisation judgments for the octave band

sources presented using the vertical stereophonic condition will be in line with the pitch-height effect and will not differ significantly from the results for single loudspeaker presentation.

In addition to the above, the second null hypothesis for this experiment is that localisation judgments for pink noise are not affected either by localisation dominance of the earlier loudspeaker or by the precedence effect when source presentation is from vertically arranged stereophonic loudspeakers. The existence of a localisation dominance effect in vertical stereophony has been demonstrated in a number of studies, including those of Somerville et al. [1965] and Tregonning and Martin [2015], and manifests in increases in ICTD causing the perceived phantom image position to move closer to the earlier loudspeaker. It is thought that the same effect will be observed in the present study for broadband pink noise. Despite this, a full image shift is not expected in line with the lack of evidence supporting the operation of the precedence effect in the median plane. Should it be the case that such effects are observed, this would indicate that there is no correlation between perceived source location and the localisation thresholds derived for the pink noise in Experiment One.

3.2.2 EXPERIMENTAL DESIGN

The present experiment was almost identical to Experiment One with respect to the location, setup, test stimuli and subjects used. However, this experiment also featured a numbered scale, which was located directly in front of the listening position and spanned the entire height of the room (Fig 3.8). The scale was numbered from 0 to 100 and had a step size of 10. With respect to the listener, the lower loudspeaker was located at ‘52’ on the scale, with the upper loudspeaker at ‘83’. The presentation methods used for the experiment were as follows: i) main layer only, ii) height layer only, iii-vii) vertically oriented phantom images with 0 dB ICLD and one of the test ICTDs applied to the upper loudspeaker.

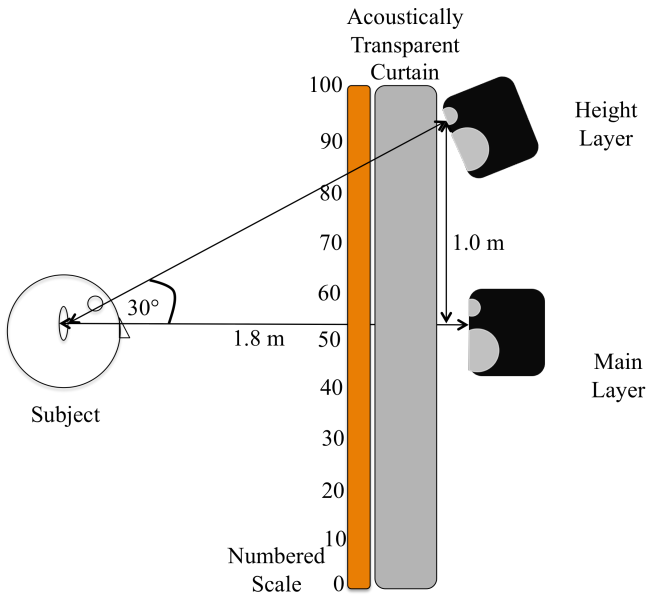

Fig. 3.8: Physical setup used for Experiment Two.

The test methodology was similar to that utilised by Pratt [1930], Roffler and Butler [1968b] and Cabrera and Tiley [2003], in that subjects were required to identify the location of each stimulus using the numbered scale in front of them. The graphical user interface used for the experiment was created using Max/MSP (Fig 3.9). For each test stimulus, subjects were presented with a slider, which had values ranging from 0 to 100 in increments of 1. This slider was to be adjusted until its value matched the perceived location of the stimulus on the scale in front. Each subject's sitting position was adjusted so that the ear height matched the height of the main layer loudspeaker (52 on the scale). The distance between the subject's ear and the main layer loudspeaker was set to 1.8 m. Although the subjects' heads were not fixed and their movements were not monitored, they were strictly instructed to maintain the set position, facing forward, keeping their head still and using only their eyes if they needed to look at the scale or the test interface. A guide point for the ear height and distance was placed on the right hand side of the subject to help maintain the correct listening position throughout the test. All subjects sat a supervised practice, which used a speech source, to ensure that they fully understood the test instructions.

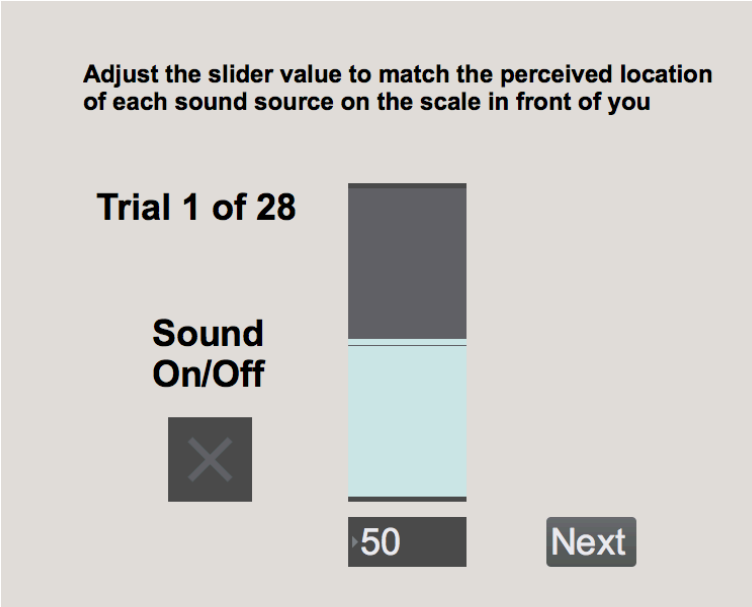

Fig 3.9: Max/MSP interface used for Experiment Two.

Due to the number of stimuli, the test was conducted in two parts, containing 28 stimuli each. Subjects were required to wait a minimum of three hours between each half of the test to remove the effects of any fatigue. The stimuli were randomized between the two tests and the test order was randomized for each listener to prevent any psychological bias.

It was decided to present the test results as elevation angles, as opposed to simply showing the gradings given on the scale. To achieve this it was necessary to calculate how many degrees of elevation a step increase of 1 on the scale corresponded to. Given that the height loudspeaker was located at '83' on the scale and was elevated by 30° with respect to the lower loudspeaker, which was located at '52', it was calculated that a step increase of 31 corresponded to an elevation increase of 30°. Therefore, a step increase of 1.03 on the scale was equivalent to 1° of elevation. Consequently, upon completion of each test, all the subjective gradings on the scale were divided by 1.03 in order to present them as elevation angles.

3.2.3 DATA ANALYSIS AND RESULTS

Levene and Shapiro-Wilk tests were first conducted, using the SPSS software, in order to determine the suitability of the collected data for parametric statistical analysis. The results of the Levene's test showed homogeneity of variance for all frequencies, whilst the Shapiro-Wilks test showed that not all scores in each condition featured normal distribution. This therefore meant that the assumptions of Analysis of Variance (ANOVA) were violated. For this reason, non-parametric tests were chosen for the statistical analysis.

3.2.3.1 The Effect of Frequency

The median perceived elevation angles for each frequency from each presentation method are plotted with notch edges in Fig. 3.10. From the lack of overlap between the notch edges, it would appear that, generally, the high frequency stimuli (4 and 8 kHz) were localised in a physically higher position than were the low frequency stimuli (125-500 Hz) for all presentation methods. However, there is also considerable overlap between the notch edges for the high and low frequencies for 0.5 ms ICTD presentation, as well as between the 500 Hz and 4 kHz stimuli for height layer only presentation. The significant effect of frequency for all presentation methods was confirmed with a Friedman test ($p < 0.001$ for each).

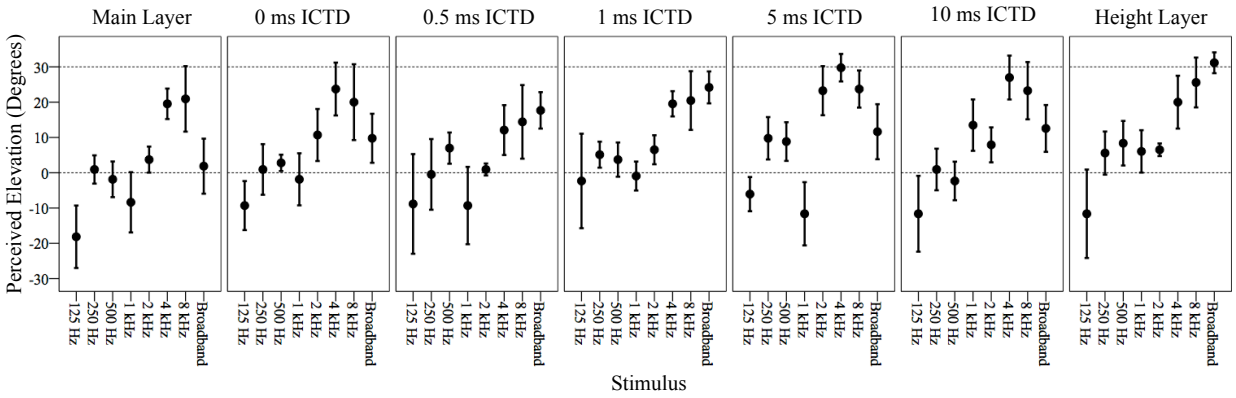

Fig. 3.10: Medians and associated notch edges showing the results of the localisation experiment. The dashed lines at 0° and 30° represent the physical positions of the main and height layers respectively.

Despite the aforementioned result, it cannot be said that there was a linear relationship between pitch and height for any presentation method. Within the low frequency range, there were no significant differences between localisation judgments for the 250 and 500 Hz bands, neither was there any between the high frequency stimuli. In addition to this, the mid frequencies (1 and 2 kHz) had variable localisation depending on the presentation method. The 1 kHz band, for example, was sometimes localised beneath the position of the main layer (main layer only, 0.5 ms ICTD, 5 ms ICTD), whilst at other times it was perceived as being above it (10 ms ICTD, height layer only). Likewise, for the 2 kHz band localisation judgments were sometimes similar to those for the 250 and 500 Hz bands (height layer, 1 ms ICTD), whereas for other presentation methods it was perceived as being significantly higher than these stimuli (5 ms ICTD).

3.2.3.2 The Effect of Presentation Method

A Friedman test was conducted to analyse the effect of presentation method on perceived elevation. The results showed significance for the 125 ($p < 0.05$) Hz and the 1 ($p < 0.01$), 2 ($p < 0.01$) and 4 kHz ($p < 0.05$)

octave bands as well as for the broadband ($p < 0.001$) source. In Fig. 3.11, the perceived elevation of each stimulus has been grouped by frequency. The data has been plotted with notch edges.

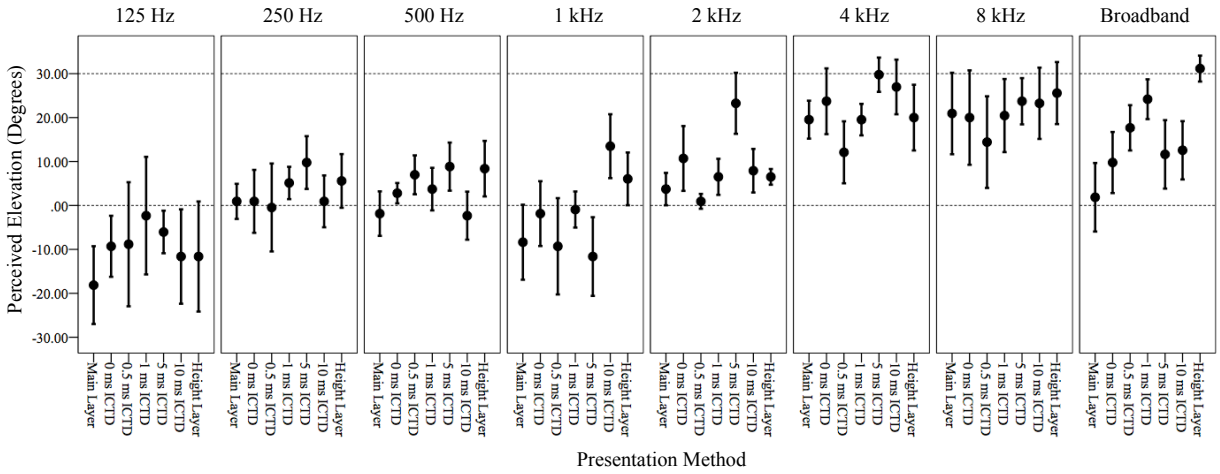

Fig. 3.11: Medians and notch edges showing the results of the localisation experiment. The data has been arranged to show the effect of presentation method.

From Fig. 3.11 it can be seen that presentation method had little or no effect on the localisation judgments for the low frequencies. Although the Friedman test results suggested that the effect of presentation method was significant for the 125 Hz band, in Fig. 3.11 it can be seen that the notch edges for all presentation methods do overlap. It should be noted however that the overlap between main layer only presentation and the 5 ms ICTD is minimal. Regarding the results for the 250 and 500 Hz bands, there is agreement between the results of the Friedman test and Fig. 3.11 although again the overlap between some pairs of presentation methods is small.

As the frequency increased it is clear that presentation method began to have some influence on localisation judgments. Despite this, the significant pairs are not consistent for all stimuli. For example, the results for 1 kHz show that the 10 ms ICTD was localised in a significantly higher position than all other presentation methods, with the exception of height layer only presentation. However, for 4 kHz the 10 ms ICTD was only localised significantly higher than the 0.5 ms ICTD, whilst for 2 kHz this presentation method was localised significantly lower than the 5 ms ICTD, with localisation being similar to most of the other presentation

methods. Presentation method can therefore be said to have had a significant effect on localisation judgments for stimuli in the range of 1-4 kHz, although the overall effect was somewhat erratic. For the 8 kHz stimuli, presentation method had no significant effect.

Localisation of the broadband stimulus was also influenced by presentation method. There are clear significant differences between localisation judgments for main and height layer only presentation, where the stimulus was accurately localised at the physical position of the emitting loudspeaker. Overall, judgments for height layer only presentation were significantly higher than those for the other presentation methods, with the exception of the 1 ms ICTD, although notch overlap in this case is minimal. There were no significant differences between localisation judgments for the 0, 0.5, 5 and 10 ms ICTDs. The 1 ms ICTD was however localised significantly higher than the 0, 5 and 10 ms ICTDs.

3.2.4 DISCUSSION

The experimental data obtained in the present experiment shows that the pitch-height effect governs the median plane localisation of octave band stimuli. Moreover, the effect is maintained when the stimuli are presented either monophonically or as ICTD-panned phantom images from vertically arranged loudspeakers. Therefore, the null hypothesis that the pitch-height effect is not maintained when band-limited stimuli are presented as vertically oriented phantom images can be rejected. However, despite the fact that for the majority of conditions the high frequencies were localised in a significantly higher position than the low frequencies, the correlation between pitch and height was not linear for any presentation method. This is demonstrated in the lack of significant differences between localisation judgments for the 4 and 8 kHz, as well as for the 250 and 500 Hz, octave bands for all presentation methods. Additionally, localisation judgments for the mid frequency stimuli were highly erratic, appearing somewhat random at times, with perceptual elevation certainly not in line with the pitch-height effect.

3.2.4.1 The Relationship with Localisation Thresholds (Experiment One)

The primary reason for conducting the present experiment was to assist in explaining the experimental data reported in Experiment One. In that experiment, it was identified that the localisation thresholds obtained were not significantly affected by the ICTD for any of the stimuli tested. This result might be explained for the 125-500 Hz and 8 kHz octave bands with the result in the present experiment, that the perceived elevation of each of these sources was not significantly affected by ICTD. This would seemingly indicate then that the perceived differences between the main layer only and phantom image conditions would be consistent with changes in ICTD, which would explain why ICTD did not have a significant effect. However, this hypothesis is insufficient in explaining the non-significant effect of ICTD on the localisation thresholds obtained for the broadband pink noise, as well as for the 1-4 kHz octave bands, whose perceived elevation was found to be significantly affected by changes in ICTD in the present experiment. A further point of note is that no consistent increases in perceived elevation were observed for the octave band sources when presentation shifted from main layer only to vertical phantom image. Instead, perceived elevation was generally similar for both conditions, in a manner comparable to that reported from the informal listening conducted following Experiment One. This is interesting because thresholds in excess of -9 dB were reported for the 1, 2 and 8 kHz octave bands, even though the results of the present study indicate that the main layer only and phantom image conditions resulted in similar perceived elevation.

From the results of the present experiment, it is apparent that that there was no correlation between the perceived elevation of each of the octave band stimuli and the localisation thresholds that were obtained in Experiment One. Consequently, the results obtained in Experiment One for these stimuli should not be interpreted from the perspective of localisation. Instead, the results arguably help to validate the VIS hypothesis proposed in Section 3.1.4.1. To reiterate, when an image with a large VIS (vertical phantom image) is directly compared to one with a low VIS (main layer only) the differences in perceived VIS could be interpreted as location differences, particularly as identifying the specific location of a wide image is generally difficult. The degree of VIS largely depends on the relative amplitudes between the loudspeaker

layers [Furuya et al. 1995]. Therefore, a reduction of the height layer's amplitude would reduce the difference in perceived VIS between the main layer only and phantom image conditions and give the perception that the two sources are more closely located. This would explain the large amounts of level reduction necessary for the 8 kHz octave band, for example, even though perceived elevation for this frequency was independent of presentation method.

In addition, it is arguable that the aforementioned VIS hypothesis also explains why the effect of ICTD on the localisation thresholds obtained in Experiment One was not significant, even though there were ICTD-based fluctuations in perceived elevation for 1-4 kHz identified in the present experiment. As was discussed in Chapter Two, the influence of an interfering vertical reflection on the main channel signal is predominantly dependent on the relative amplitudes of the two sources, although other factors such as ICTD do have some effect [Tregonning and Martin 2015]. Reducing the difference in perceived VIS by attenuating the height layer would also have the effect of lessening the interfering effects of ICTD the perceived elevation of a given stimulus. It therefore stands to reason that the amount of necessary attenuation is consistent irrelevant of the ICTD, as the degree of interference depends primarily on the relative amplitudes of the two layers as opposed to the delay between them.

With respect to the broadband source, the experimental data shows that increases in source elevation are perceived when source presentation changes from main layer only to vertical phantom image. This result therefore agrees with the informal observations of the author discussed in Section 3.1.4.1. Arguably, this result supports the suggestion that the balance of spectral energy provided by the main and height layers respectively in the 7-9 kHz range, as was suggested in Section 3.1.4.1, is an important factor in determining the localisation threshold for more complex sources. Furthermore, even if the 7-9 kHz region specifically was not as dominant as suggested by this hypothesis, the results of the experiment do at least imply that consideration must be given to the cues that are used for elevation perception when trying to explain the results.

3.2.4.2 The Effect of ICTD

A further aim of the present experiment was to determine whether or not evidence could be found for either the precedence effect or localisation dominance in median plane stereophony. In the range of ICTDs tested there was no condition whereby the broadband pink noise was localised at the position of the main layer. The present experimental data therefore suggests that the precedence effect is not a feature of vertical stereophony, at least in the current experimental setup using the 30° elevation angle. This supports the results presented in the studies conducted by both Lee [2011] and Stenzl et al. [2014], as well as those reported in Experiment One of the present thesis.

Additionally, the experimental data also showed no evidence of localisation dominance. When elevation judgments for the broadband pink noise were influenced by ICTD, the resultant effect was an increase in perceived elevation, which is the opposite of what was expected. As a result of these findings, the null hypothesis that localisation judgments for broadband pink noise are not affected either by localisation dominance or the precedence effect when source presentation is from vertically arranged stereophonic loudspeakers can be accepted. In order to gain objective insights into this result, the ICTDs used for the experiment were applied to head related impulse responses (HRIRs) measured for 0° and 30° elevation angles at 0° azimuth, taken from the MIT's KEMAR dummy head database [Gardner and Martin, 2000] (Fig. 3.12). For each graph, the black lines show the original signals, whilst the red lines represent the same signals smoothed using a moving average filter. Consideration of the original signals alone shows that comb filtering had a large influence on the frequency content of all broadband pink noise stimuli presented stereophonically with a delay in the height layer. The effects of comb filtering arguably make it possible to explain the observed results. It is well established that spectral cues are required for a sound to be localised in a specific region in vertical space. Previous research from Hebrank and Wright [1974a] and Asano et al. [1990] has shown that the spectral cues for elevation correspond to a notch in the frequency spectrum in the range between 4 and 10 kHz, with an increase in the centre frequency of this notch leading to the perception of increased elevation. From Fig. 3.12 it can be seen that the effect of comb filtering is to introduce

additional notches in the frequency spectrum. It is possible that the human auditory system can interpret such notches as spectral cues for vertical localisation, therefore affecting perceived elevation.

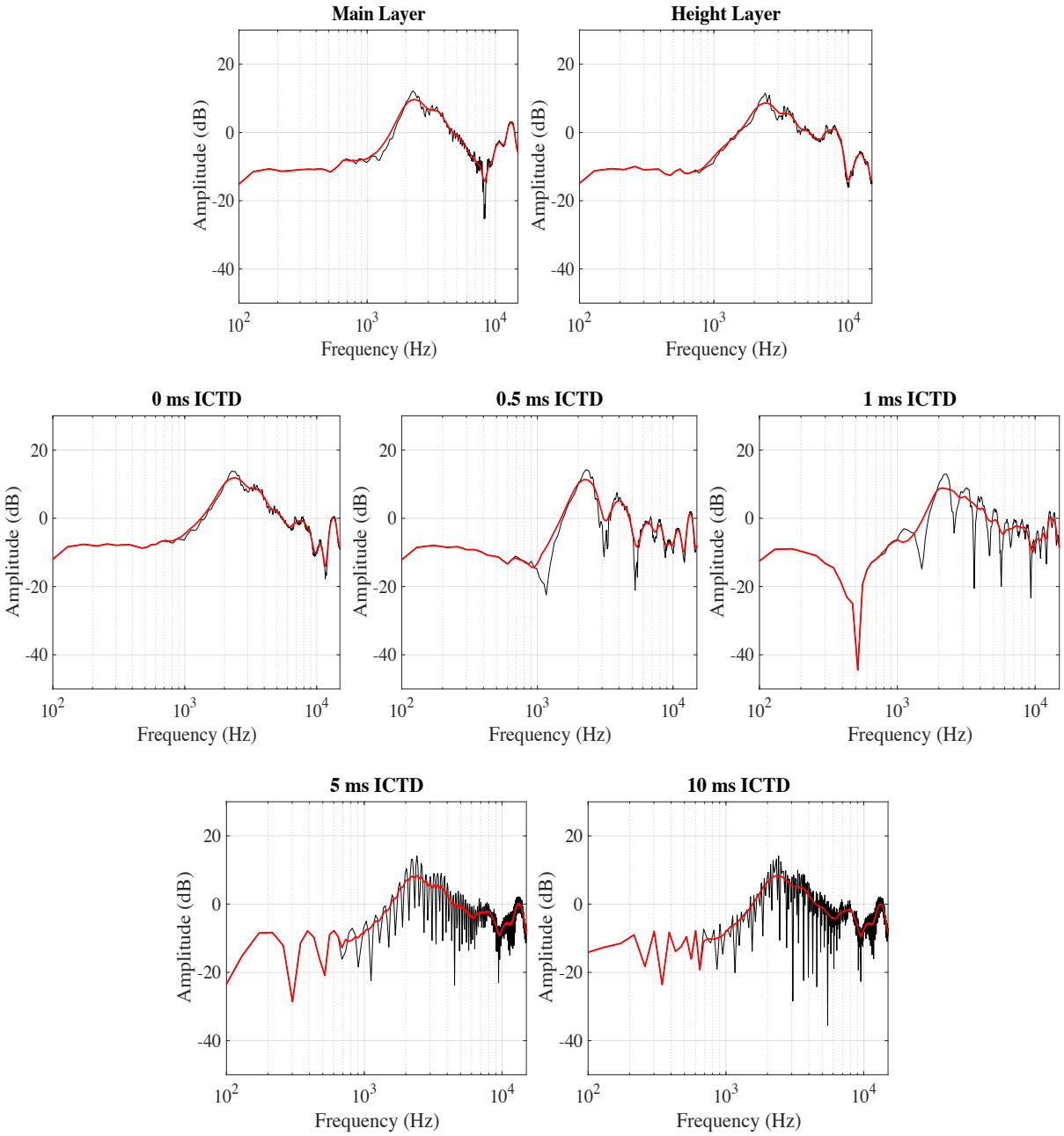

Fig 3.12: HRIRs taken from the MIT Database [Gardner and Martin 2000] for i) main layer only, ii) height layer only and iii-vii) 0 dB ICLD with one of the test ICTDs applied to the height layer. The red lines are a smoothed version (moving average filter) of the original signals (black lines).

When the broadband pink noise source was presented with ICTDs between 0 and 1 ms, an increase in ICTD led to the perception of increased elevation. From Fig. 3.12, it is noticeable that the 0.5 ms ICTD features notches in the region of 3 and 5 kHz, which are not present for the 0 ms ICTD and that were also maintained when the signal had been smoothed. These notches may have been interpreted as elevation cues, leading to the pink noise presented with 0.5 ms ICTD being perceptually elevated with respect to that presented with 0 ms ICTD. Additionally, the 1 ms ICTD features notches in the region of 5, 6 and 9 kHz. This stimulus was perceived to be more elevated than the 0.5 ms ICTD, which may be due to the increased centre frequencies of the notches. If this were the case then this would verify Hebrank and Wright's results [1974a]; that increased notch centre frequency between 4 and 10 kHz leads to the perception of increased elevation. It should be noted however that the differences in perceived elevation between 0 and 0.5 ms and between 0.5 and 1 ms were not significant. In addition to this, the spectral enevelope in the 4-10 kHz range for the 1 ms ICTD is similar to that for the 0 ms ICTD when the signals have been smoothed. This might therefore suggest that the comb filtering notches had less of an influence on perceived elevation than had been suggested previously. This being said, it is also clear that there is a prominent, wide notch around 500 Hz that is for the most part absent from the responses from the other ICTDs. Based on this observation, the following could be suggested. The 500 Hz notch at 1 ms would have resulted in the 1 ms ICTD condition having greater perceived high frequency weighting compared to each of the other test stimuli. It might therefore be the case that this weighting caused an increase in the perceived elevation of the stimuli based on the pitch-height effect. This would have to be studied further but does at least provide a further reason why the 1 ms ICTD resulted in the highest perceived elevation for the broadband pink noise source. Further, is should be noted that, based on these discussions, it is not entirely clear how the distinct notches affect perceived elevation and this also requires further study.

Interestingly, the frequency content for the 5 and 10 ms ICTDs features considerably more notches between 4 and 10 kHz than for the other stimuli and yet they were perceptually no more elevated than the stereophonic broadband stimulus with no ICTD applied. Again, this result might be interpreted based on Hebrank and Wright's conclusions [1974a], which suggest that the elevation cue between 4 and 10 kHz is a

1-octave notch. In Chapter Two, it was demonstrated that as the ICTD increases the bandwidth of the notches due to comb filtering decreases. It may therefore be that the comb filtering notches for the 5 and 10 ms ICTDs are not of sufficient bandwidth to be interpreted as elevation cues. Additionally, the overall spectral envelope of these stimuli in the 4-10 kHz region is close to that for the 0 ms ICTD and therefore the similarities in perceptual elevation seem plausible. Further, it is also true that, although the spectrum below 1 kHz has been modified compared to the 0 ms condition, the notches are not as severe as for the 1 ms ICTD. This would therefore mean that the high frequency weighting of the source would not be as high for the 5 and 10 ms conditions compared to the 1 ms condition, which would further explain why the sources were perceived as being less elevated.

Perhaps then the reason that localisation dominance effects were not observed in the present study lies in the choice of stimuli. In the studies in which a localisation dominance effect was demonstrated, the stimuli that were used would have been less influenced by the comb filtering compared to pink noise. Litovsky et al.'s study [1997], for example, used clicks, whilst Tregonning and Martin [2015] used conga and female speech excerpts. It could be argued then that the coloration of the broadband pink noise source in the present study was the reason that localisation dominance effects were not observed. This would also indicate that spectral cues are more dominant than are time-based cues with respect to localisation in vertical stereophony; this would need to be studied further.

The aforementioned hypothesis is, however, not sufficient to explain the erratic effects of ICTD observed for the octave band stimuli. The experimental data showed that ICTD had little effect on the perceived elevation of the 125-500 Hz and 8 kHz octave bands. On the other hand, the 1-4 kHz bands were heavily affected by changes in ICTD. For these stimuli, even though the effect of ICTD was significant the spread of results did not follow the same pattern as for the broadband pink noise source. This being said, it is clear that comb filtering would have affected these stimuli and it remains arguable that in some way this would have influenced their perceived elevation. Additionally, despite the presence of comb filtering the 8 kHz octave band was almost entirely unaffected by changes in presentation method. This might be related to the strength

of the relationship between 8 kHz and above localisation as described by Blauert [1969], although this would not explain the wide error bars for the 8 kHz stimuli observed in the present test. It is clear spectral cues in relation to the localisation of octave band stimuli requires further study.

3.2.4.3 Comparison with Cabrera and Tiley [2003]

The result that the pitch-height effect governs the localisation of octave bands presented from vertically arranged loudspeakers has been previously demonstrated by Cabrera and Tiley [2003]. In their experiment, pink noise, filtered pink noise and octave bands were presented to subjects from either 1, 3 or 5 contiguous loudspeakers arranged vertically in the median plane, with elevation angles of 0°, ±7.9° and ±15.6°. The centre frequencies of the octave band stimuli were 0.125, 0.5, 2 and 8 kHz. Stimuli were presented to listeners in ten 200 ms bursts at loudness levels of 64 and 84 phon. There are a number of similarities between the results obtained in the respective studies. Firstly, when stimulus presentation was from the non-elevated loudspeaker, both studies found that the 125 Hz band was localised beneath the position of the emitting loudspeaker, whilst 2 and 8 kHz were localised above it. It should be noted however that in Cabrera and Tiley's [2003] study 500 Hz was perceived as being beneath the lower loudspeaker, whilst in the present study 500 Hz coincided more with the physical position of the main layer loudspeaker. Additionally, both studies showed that the broadband (pink noise) stimuli were localised accurately at the position of the emitting loudspeaker (this was the case for monophonic presentation in the present study).

However, although for upper loudspeaker presentation both studies showed a relationship between pitch and height, Cabrera and Tiley's [2003] results showed that judgments for the 125 and 500 Hz octave bands were almost identical. Conversely, in the present study the 500 Hz band was localised significantly higher than 125 Hz. In this case, judgments for 500 Hz were more in line for those with 2 kHz. Moreover, when Cabrera and Tiley [2003] presented the 8 kHz octave band from the uppermost loudspeaker, localisation judgments were significantly higher than for when the same stimulus was presented from the non-elevated loudspeaker.

This was the case for both loudness levels. There also appeared to be some correlation between the position of the emitting loudspeaker and the perceived location of the stimulus. No such correlation could be seen in the present study. Additionally, the difference between localisation for the 8 kHz stimuli when presented monophonically from either loudspeaker was not significant. It should be noted that there were a number of differences in the experimental setup that may have contributed to these differences. Firstly, there were differences in upper loudspeaker elevation in the respective studies. In the present study the height layer loudspeaker was elevated by 30°, whereas in Cabrera and Tiley's [2003] study the uppermost loudspeaker was only elevated by 15.6°. Moreover, Cabrera and Tiley [2003] presented their stimuli as 200 ms bursts; in the present study stimulus presentation was continuous. It is possible that the listeners in Cabrera and Tiley's [2003] study were therefore afforded additional localisation cues due to the burst nature of the stimuli, leading to more accurate localisation of the 8 kHz octave band. This would require further study.

Also, Cabrera and Tiley [2003] did not consider the localisation of octave bands with centre frequencies of 0.25, 1 and 4 kHz. This makes it difficult to analyse whether the lack of pitch-height linearity was as much a feature in their study as it was in the present. The experimental data presented here indicates that localisation judgments for the 250 Hz band were similar to those for the 500 Hz band, whilst the 4 kHz band was localised similarly to the 8 kHz band. Localisation judgments for the 1 kHz band were slightly more erratic. Generally this stimulus was localised beneath, or in the position of, the lower loudspeaker.

3.2.5 CONCLUSION

The present experiment was conducted in a bid to understand the relationship between perceived source elevation and the localisation thresholds obtained in Experiment One. The experiments investigated the effect of ICTD on the vertical stereophonic localisation of octave band (125 Hz - 8 kHz) and broadband pink noise stimuli. Stimuli were presented either monophonically or as stereophonic phantom images, with the height layer delayed with respect to the main. The experiment used ICTDs of 0, 0.5, 1, 5 and 10 ms.

The experimental data obtained from the study showed that localisation under the above conditions is governed by the pitch height effect. For the majority of presentation methods, the high frequency stimuli were localised in a significantly higher position than were the low frequency stimuli. Despite this, the relationship between pitch and height was found to be non-linear in all cases. Additionally, localisation for the mid frequency stimuli was found to be somewhat erratic.

ICTD was found to have a random and inconsistent effect on the perceived elevation of the 1-4 kHz octave bands; the other bands were unaffected. Localisation judgments for the broadband pink noise source showed no evidence of either the precedence effect or localisation dominance, although the effect of ICTD remained significant. These results were interpreted based on the effects of comb filtering, which were reasoned to have distorted the spectral cues used in vertical localisation.

Ultimately, the results of the study showed that there was no correlation between the perceived elevation of the sound sources and the localisation thresholds that were obtained in Experiment One. This also helped to validate the hypothesis that the most salient attribute in the determination of the localisation thresholds for the octave band stimuli in Experiment One was the difference in perceived VIS between the main layer only and vertical phantom image conditions. However, the spectral dominance of the main layer should not be ruled out as an explanation for the thresholds obtained for the broadband pink noise source.

3.3 SUMMARY

Two experiments have been reported in this chapter. In the first, listening tests were conducted in order to investigate the frequency dependency of localisation thresholds in relation to vertical interchannel crosstalk. Octave band and broadband pink noise stimuli were presented to subjects as phantom images from vertically arranged stereophonic loudspeakers located directly in front of the listening position. With respect to the listening position, the main layer was not elevated; the height layer was elevated by 30°. Subjects completed a method of adjustment task in which they were required to reduce the amplitude of the height layer until the

resultant phantom image matched the position of the same stimulus presented from the main layer alone. The height layer was delayed with respect to the main by 0, 0.5, 1, 5 and 10 ms.

In the second experiment, the effect of ICTD on the vertical stereophonic localisation of band-limited stimuli was analysed. The study utilised seven octave bands of pink noise, with centre frequencies ranging from 125 Hz to 8 kHz, as well as a broadband pink noise source. Stimuli were presented either monophonically or as stereophonic phantom images, with the height layer delayed with respect to the main by 0, 0.5, 1, 5 and 10 ms. Subjects identified the perceived location of each stimulus using a numbered scale located directly in front of the listening position.

The key findings of the investigations are as follows:

Experiment One:

- Localisation thresholds for octave band stimuli are frequency dependent. Low frequency stimuli required significantly less level reduction than did the mid-high frequencies.
- There was no significant effect of ICTD on localisation threshold for any stimulus.
- ICLD was always necessary for the broadband pink noise source, suggesting that the precedence effect did not operate.
- It might be possible to apply localisation thresholds by manipulating the amplitude of single frequency bands, rather than by reducing direct sound levels as a whole, in the height layer.
- The frequency dependency of localisation thresholds might be related to differences in VIS between the main layer only and phantom image conditions.
- For the broadband pink noise source, the relative strength of the spectral cues from each loudspeaker layer may determine whether the localisation threshold has been met, particularly in the 7-9 kHz region. However, the lack of a localisation dominance effect also indicates the importance of VIS differences.

Experiment Two:

- The pitch-height effect is maintained for octave band stimuli presented as vertically oriented phantom images.
- The 1-4 kHz octave band stimuli were significantly affected by ICTD, however the effect was random and inconsistent; the other octave bands were not significantly affected.
- For the broadband source, increasing the ICTD from 0-1 ms caused the resultant phantom image to perceptually move closer to the later loudspeaker. This arguably showed a lack of localisation dominance and was likely related to the effects of comb filtering.
- The perceived elevation of the test stimuli did not correlate with the localisation thresholds derived in Experiment One. This arguably validates the VIS hypothesis for the band-limited stimuli. For the broadband source, the results support the hypothesis that the balance of spectral energy between the main and height layers in the range at which the pinna cues for elevation exist is important.
- No evidence could be found to support the operation of the precedence effect in median plane stereophony.

4 ANALYSIS OF BAND AND BLANKET REDUCTION LOCALISATION THRESHOLD METHODS[3,4]

Three experiments are presented in this chapter. In the first (Experiment Three), localisation thresholds were obtained using the blanket reduction method for a broad range of natural sound sources. In the second (Experiment Four), the frequency dependency of localisation thresholds was explored in the presence of reflections, building on the findings of Experiment One and also providing the experimental basis for the development of a band reduction method. The final experiment (Experiment Five) was a verification test, in which the blanket and band reduction thresholds obtained in Experiments Three and Four were tested.

4.1 EXPERIMENT THREE: LOCALISATION THRESHOLDS FOR NATURAL SOUND SOURCES (BLANKET REDUCTION)

One of the aims of the present thesis is to analyse how the perceptual effects of vertical interchannel crosstalk vary when using the blanket and band reduction methods. In addition, it is of further interest to determine which method subjects consider as being the most preferred. In order that it is possible to address these aims, it is first necessary to derive localisation thresholds using each method, with the present experiment focusing on blanket reduction.

[3] Wallis, R. and Lee, H. [2017]: 'The Reduction of Vertical Interchannel Crosstalk: The Analysis of Localisation Thresholds for Natural Sound Sources', *Applied Sciences,* 7(3), 278 pp. 1-20.

[4] Wallis, R. and Lee, H. [2016]: 'The Frequency Dependency of Localisation Thresholds in the Presence of Reflections', Presented at the 29th Tonmeistertagung.

It should be noted that thresholds using blanket reduction have already been reported in the literature, with Lee [2011] and Stenzl et al. [2014] finding that the threshold for musical sources lies between -6 and -7 dB for ICTDs up to 10 ms. Despite these findings, it has been decided to conduct a novel analysis into localisation thresholds using the blanket reduction method. The reason for this is as follows. Experiment One of the present thesis identified that there existed a frequency dependency of localisation thresholds; with the mid-high frequency stimuli generally requiring more level reduction than did the low frequency stimuli. These results seem to provide an implication for the analysis of localisation thresholds for natural sound sources with different spectral balances; it might be suggested that the threshold for a high frequency dominant source would be lower than that for a low frequency dominant source. However, it should be noted that this hypothesis is at odds with the findings in the Lee [2011] and Stenzl et al. [2014] studies, with each reporting that the effect of sound source was not significant on the localisation thresholds obtained. However, since the sources used in those studies were somewhat limited (cello and bongo by Lee [2011], cello, bongo and speech by Stenzl et al. [2014]), a wider range of natural sound sources would need to be tested in order to confirm the source dependency of the localisation threshold.

Of further interest in the present experiment is how localisation thresholds are affected by the way in which the test stimuli are presented to subjects (the presentation method). A recent study conducted by Lee [2016] examined localisation judgments for continuous broadband pink noise presented from loudspeakers arranged in two layers in a manner similar what was reported in Experiment Two. However, rather than testing a vertical stereophonic condition (i.e. a single loudspeaker in each layer), Lee [2016] utilised a configuration that will henceforth be referred to a ‘vertical quadraphonic’. In this case, the main layer consisted of stereophonic loudspeakers with a base angle of 60°, whilst the height layer comprised a further pair of stereophonic loudspeakers, each of which was located directly above a loudspeaker in the main layer. With respect to conventional 3D audio reproduction systems, such as Auro 9.1 [Auro Technologies 2016], the loudspeakers used in the vertical quadraphonic condition would correspond to L, R, HL and HR. The results of Lee’s [2016] study showed that the pink noise was perceived as being elevated with respect to the physical position of each layer. This differs from the results of Experiment Two, which showed that

localisation judgments were accurate. This difference in results is indicative of the phantom image elevation effect, in which stimuli are perceived as being more elevated when presented as stereophonic phantom images compared to single source only presentation [De Boer 1947, Lee 2017]. This has notable implications for the reduction of vertical interchannel crosstalk. If main channel images are elevated with respect to the physical position of the main channel layer as a result of the phantom image elevation effect, then it could be argued that the location-based effects of vertical interchannel crosstalk would be less distracting. Consequently, the localisation threshold might be much lower under such circumstances or, alternatively, might not be necessary at all. It is therefore of interest to determine how the localisation thresholds would vary when stimuli are presented as vertically arranged quadraphonic phantom images compared to for vertical stereophonic presentation.

From the above background the following research questions were derived:

- Does there exist a sound source dependency for localisation thresholds?
- How do localisation thresholds vary for vertical quadraphonic stimulus presentation compared to vertical stereophonic?
- Can any evidence be found to support the existence of the precedence effect in the median plane?

This section is organised as follows. An experiment is first described in which the effects of sound source, presentation method and ICTD on the localisation threshold were determined. Following this, the results are presented and analysed. The section concludes with discussions pertaining to the results of the experiment, as well as the implications for image rendering and microphone techniques.

4.1.1 EXPERIMENTAL HYPOTHESIS

The first null hypothesis to be tested in this experiment is that localisation thresholds using the blanket reduction method are not source dependent. It is anticipated that this null hypothesis will be accepted based on the following reasons. Firstly, the blanket reduction studies conducted by Lee [2011] and Stenzl et al. [2014] each showed that localisation thresholds were not source dependent. Alongside this are the results and discussions presented in the present thesis. For instance, the frequency dependency of localisation thresholds has only been observed for octave band stimuli, which are notably more narrowband than natural sound sources. This is a particularly salient point for the following reason. Following Experiment One, it was hypothesised that the frequency dependency of localisation thresholds for octave band sources might be related to frequency-dependent differences in perceived VIS between the main layer only and phantom image conditions. It can be argued then that natural sound sources would be too broadband for such frequency-dependent differences to be perceived.

In addition, a further null hypothesis to be considered is that presentation method does not have a significant effect on localisation thresholds when using the blanket reduction method. It is thought that the acceptance of this null hypothesis will be dependent on the difference in perceived source elevation between the quadraphonic and stereophonic conditions for main layer only presentation being maintained when the stimuli are presented as vertically oriented phantom images. If the difference is maintained, then this would suggest that the perceived difference in elevation between the phantom image and main layer only conditions would be similar for a given ICLD for both methods. This would arguably manifest in a similarity of localisation thresholds. Conversely, if the difference is not maintained then the perceived difference between main layer only and vertical phantom image presentation would be smaller for the quadraphonic method than for stereophonic, which would likely mean that the localisation threshold would be higher (i.e. less level reduction needed) for the former method.

The final null hypothesis for the present experiment is that the precedence effect does not operate for vertically oriented phantom images. It is anticipated that this null hypothesis will be accepted for the following reason. If the precedence effect were to operate then sufficient ICTD alone would result in the localisation threshold being met, without the need for ICLD. Such a result has not been reported either in the present thesis for octave bands and broadband pink noise, or for musical sources in the studies of Lee [2011] or Stenzl et al. [2014]. Despite this, it is possible that localisation dominance effects might be observed. In Chapter Two, it was discussed how Somerville et al. [1965], Blauert [1971], Litovsky et al. [1997] and Tregonning and Martin [2015] had all shown localisation dominance of the earlier loudspeaker for sources incident in the median plane. It seems plausible that the same effect might be observed in the present experiment. This would manifest in less ICLD being required for sources in which an ICTD above 1 ms is present. It should be noted, however, that this was not observed by either Lee [2011] or Stenzl et al. [2014] or in the experimental works presented thus far in the present thesis.

4.1.2 EXPERIMENTAL DESIGN

4.1.2.1 Physical Setup

Fig. 4.1 shows the physical setup used for the experiment, which was conducted in the ITU-R BS.1116-compliant listening room [ITU 1994] at the University of Huddersfield. The experiments utilised six Genelec 8040A loudspeakers, which were arranged in two layers, 'height' and 'main'. The main layer consisted of centre (C), left (L) and right (R) loudspeakers, which were each positioned 1.2 m above the ground and 2 m from the listening position. With respect to the listening position, the centre loudspeaker was located at 0° azimuth, with the left and right loudspeakers at ±30°. The height layer comprised the three remaining loudspeakers, each of which were positioned 1.15 m directly above a loudspeaker in the main layer (HL, HR and HC). With respect to the listening position, the main layer was not elevated, whilst the height layer was elevated by 30°. Appropriate time and level alignment was applied to the main layer with respect to the height layer to accommodate for the difference in distance between the loudspeakers in each layer and the

listening position. An acoustically transparent curtain was positioned between the listening position and the loudspeakers in order to obscure the nature of the test setup from subjects. The ear height of subjects was aligned to the centre point between the woofer and tweeter on the main layer of loudspeakers using a height-adjustable chair.

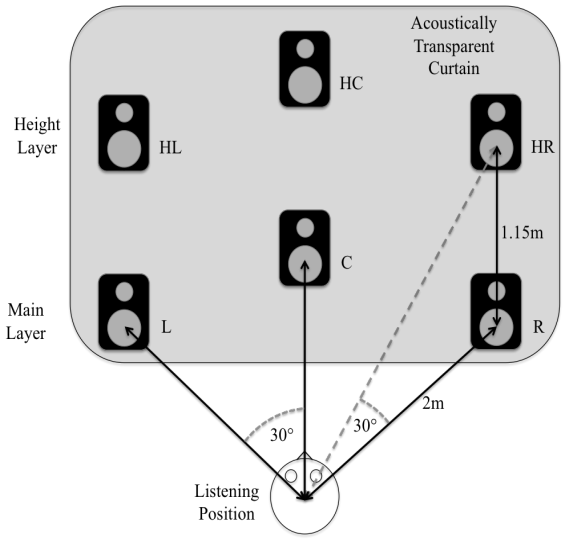

Fig. 4.1: The physical setup used for Experiment Three.

4.1.2.2 Test Stimuli

The test stimuli used for the experiment were anechoically recorded guitar, speech, conga, quartet and oboe excerpts (Fig. 4.2). These stimuli were chosen primarily due to their variations in spectral content. The predominant energy of the oboe source, for example, was in the range of 600 Hz – 2 kHz, with notable peaks around 700 Hz and 1.5 kHz, whilst that for the conga ranged from 150-500 Hz. Given that in Experiment One it was reported that the localisation thresholds for low frequency octave bands were significantly higher compared to those for the mid-high frequency bands, it was thought that these two sources in particular would be beneficial in analysing the source dependency of localisation thresholds for natural sound sources. The speech source was chosen due to its broadband nature. It was reasoned that if a source dependency could be identified based on frequency then the inclusion of a wideband source would potentially make it possible

to identify the frequency region that is more dominant when determining the localisation threshold for broadband sources. The guitar and quartet sources were chosen due to their varying balance between low and high frequency content, which was more even for the quartet compared to the guitar (although both were more dominant in the region below 1 kHz). This again was due to the aim of analysing whether or not the frequency dependency of localisation thresholds could translate to natural sound sources. It should also be noted that the sources contained a varied blend of continuous and transient characteristics, as can be seen in Fig. 4.2, although it has not yet been reported in the literature how this would affect the localisation threshold.

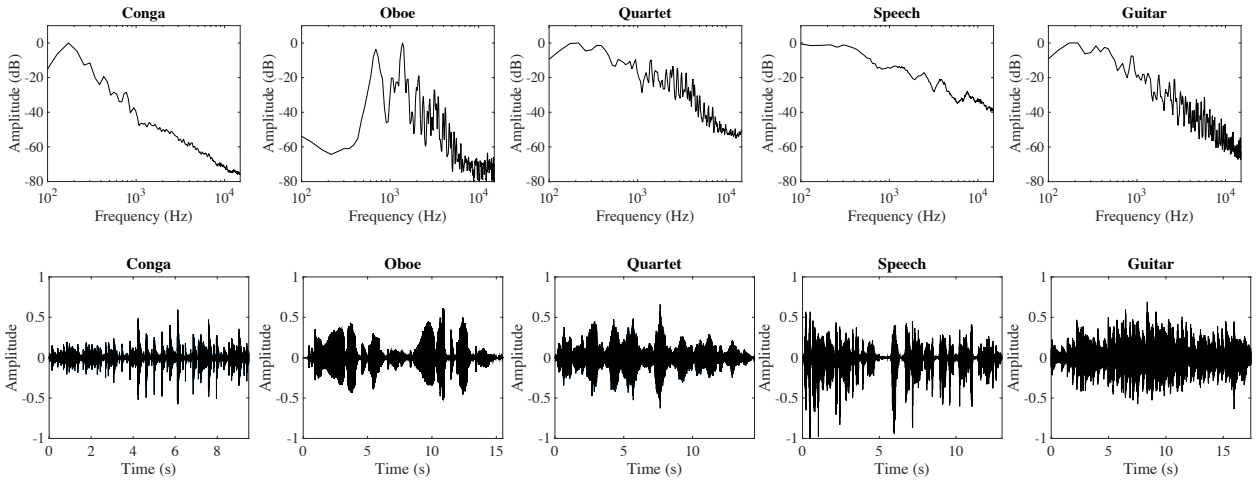

Fig. 4.2: Spectra and waveforms of test stimuli used for Experiment Three.

The test stimuli were presented to subjects as vertically oriented phantom images using the following two conditions (Fig. 4.3):

1. Vertical stereophonic: stimulus presentation from the C (main layer) and HC (height layer) loudspeakers.
2. Vertical quadraphonic: stimulus presentation from the L, R (main layer), HL and HR (height layer) loudspeakers.

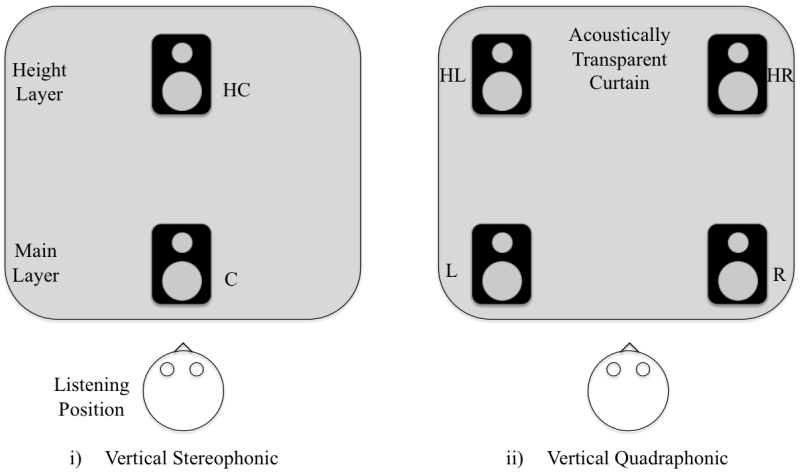

Fig. 4.3: Presentation methods for test stimuli.

For each condition, the resultant phantom image was formed directly in front of the subject (i.e. on the median plane). The height layer was delayed with respect to the main layer for both conditions by 0, 1 and 10 ms. The delay times were chosen to emulate different spacings between the main and height microphone layers in the context of concert hall recording, with 10 ms being a likely maximum spacing (corresponding to a path difference of 3.4 m between the direct sound arriving at each microphone layer). In total, there were 30 stimuli (five sources, three delay times and two presentation methods). The amplitude of each stimulus at the listening position when presented from the main layer only (either C or L and R) was 70 dB LAeq. The amplitude of the stimulus when presented as a phantom image was dependent on the amplitude of the height layer relative to the main, which was to be varied by the subject as described in Section 4.1.2.4.

4.1.2.3 Subjects

Ten subjects, comprising staff and both postgraduate and final year undergraduate students from the University of Huddersfield's Music Technology courses, participated in the listening tests. These subjects were chosen due to their critical listening experience in spatial audio, making them better suited than more

naïve subjects to determine the subtle localisation differences caused by vertical interchannel crosstalk. They all reported normal hearing.

4.1.2.4 Test Method

The basic methodology of for the experiment was similar to that used for Experiment One. For each stimulus, subjects were presented with a 'test' and 'reference' sound. The 'reference' was the stimulus presented from the main layer only. The 'test' sound was the stimulus presented as a vertically oriented phantom image with one of the three test ICTDs applied to the height layer. For each 'test' sound, subjects were required to reduce the amplitude of the height layer until they perceived the location of the resultant phantom image to be matching that of the 'reference'. To ensure the localisation threshold was found in each case, they were asked to set the amplitude of the height layer to the highest possible point at which this condition was met.

The MOA used for Experiment One was limited in that it presented subjects with a large range from which to find the threshold, which made it difficult for subjects to give precise answers [Lawless 2013]. Despite this, MOA was still considered as being the most appropriate method for the present experiment for the reasons discussed in Section 3.1.2.4. This limitation was addressed for the present experiment by requiring subjects to complete a three-stage MOA for each stimulus. Each stage was designed to be a more refined version of the previous stage as follows:

- Stage 1: The amplitude of the height layer could be adjusted from 0 to -25 dB in 5 dB steps. The localisation threshold was therefore found to within 5 dB.
- Stage 2: The amplitude of the height layer could be adjusted in the 5 dB range determined by the previous stage. The step size was 1 dB.
- Stage 3: The amplitude of the height layer could be adjusted in the 1 dB range determined by the previous stage. The step size was 0.25 dB.

This method can be considered as combining the standard MOA with adaptive testing. Adaptive threshold detection methods, which include PEST [Taylor and Creelman 1967] and Up-Down [Levitt 1971], present stimuli to subjects at amplitudes determined by the history of the test run. This allows testing to be made at levels closer to the threshold, increasing efficiency [Levitt 1971]. It was however decided against using an adaptive method outright based on research conducted by Hesse [1986], which indicated that the subsequent duration of the test is between three and five times longer compared to when MOA is used. As the method used for the present experiment used elements of both MOA and adaptive testing, it was named as an 'adaptive method of adjustment' (AMOA). It was considered that this fusion of methods would improve the accuracy of the test, whilst still making it relatively quick and easy for subjects to complete. The graphical user interface for the AMOA task was created using Max/MSP (Fig. 4.4).

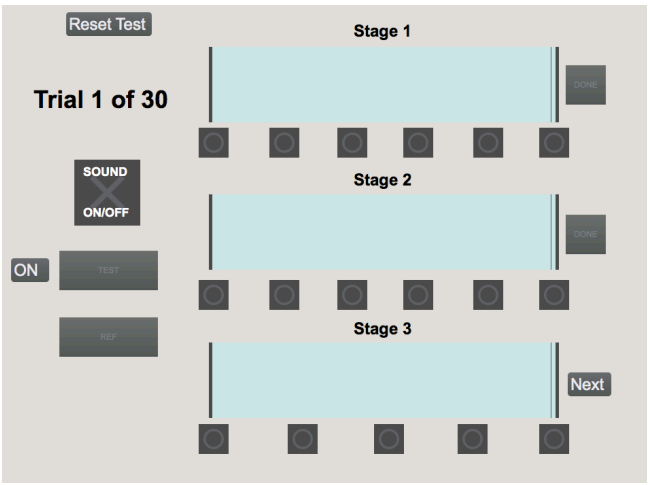

Fig 4.4: Max/MSP interface used for Experiment Three (the AMOA method).

During the test, subjects were strictly instructed to face forwards, keeping their head still and using only their eyes to look at the test interface. The heads of subjects were not fixed, however head movements were monitored using a motion tracker device [Johnson et al. 2016]. The tracker utilised a Kinect [Microsoft 2017] in order to track the movement of subjects against a fixed location (i.e. the 'ideal' listening position). If the subject's head deviated from the listening position beyond a range of natural motion (10 mm in any direction) then the Kinect triggered a warning to a Max/MSP interface that was located in front of the

listening position, just below the main layer of loudspeakers. The interface informed subjects that their listening positon was not correct and instructed them on how to move their head in order to get back to it (i.e. if a subject strayed too far to the left then the interface would tell them to move to the right). In addition to the use of the motion tracker, a guide point for the ear height and distance was placed on the right hand side of the subject to help maintain the correct listening position throughout the test. Prior to the start of each test, all subjects sat a supervised practice, which utilised a speech source, in order to ensure that the instructions were understood. The test was completed in two sittings, each of which contained 15 stimuli and lasted around 20 minutes. The order of the tests, as well as the stimulus order, was randomised for each subject.

4.1.3 DATA ANALYSIS AND RESULTS

Levene and Shapiro-Wilk tests were first conducted, using the SPSS software, in order to determine the suitability of the collected data for parametric statistical analysis. The results of the Levene test showed homogeneity of variance for all sound sources, whilst the Shapiro-Wilk test showed that not all scores in each condition featured normal distribution. This therefore meant that the assumptions of Analysis of Variance (ANOVA) were violated. For these reasons, non-parametric tests were chosen for the statistical analysis.

4.1.3.1 The Effect of Presentation Method

Fig. 4.5 shows the median localisation thresholds for each stimulus at each ICTD for both presentation methods. The medians have been plotted with notch edges. In general, there is considerable overlap between the notch edges for each presentation method, which suggests that the effect of presentation method was not significant. However, it is clear that in some cases the overlap between notches is minimal (e.g. conga at 1 ms, quartet at 10 ms). In order to analyse this further, the results for vertical stereophonic and quadraphonic presentation were compared for each stimulus using Wilcoxon tests. The critical p value was 0.05.

According to this analysis, the effect of stimulus presentation was only significant for the Oboe with 0 ms ICTD (p = 0.036). However, it is clear from Fig. 4.5 that there is a large overlap between the notch edges for this stimulus. In addition to this, the effect size r, calculated based on Cohen [1988], was 0.49, which is not considered as being a large effect [Cohen 1988]. It could be argued then that the significant effect identified in the Wilcoxon test was a type-I error, being a false positive when in fact there is no true effect [Lieberman and Cunningham 2009]. It can therefore be concluded that the effect of presentation method on the localisation thresholds obtained was not significant.

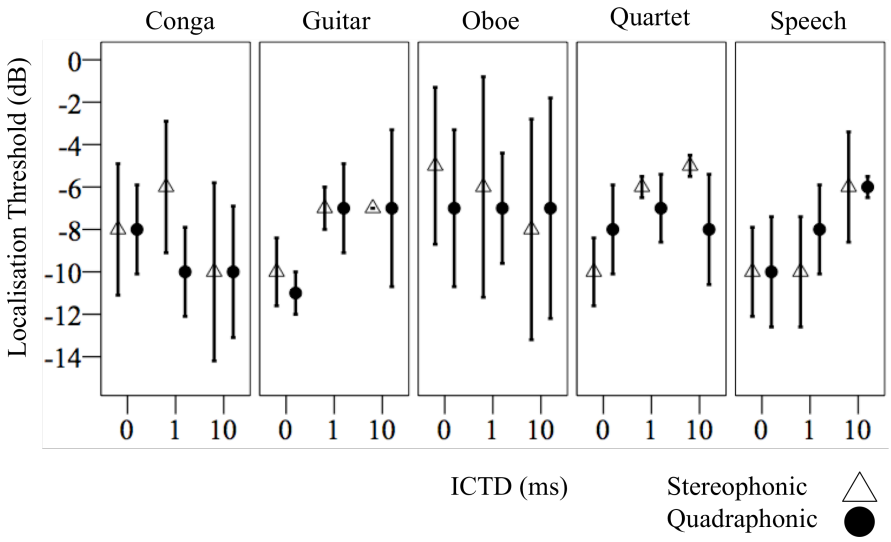

Fig. 4.5: Medians and associated notch edges for each experimental condition.

4.1.3.2 The Effect of ICTD

As it was identified that the effect of stimulus presentation on the localisation threshold was not significant, the results for stereophonic and quadraphonic presentation were combined. Fig. 4.6 shows the effect of ICTD on the localisation thresholds obtained, with combined results for stimulus presentation. As before, the median localisation thresholds have been plotted with notch edges. Consideration of the notch edges suggests that the effect of ICTD on the localisation threshold was significant for at least some of the sound

sources. The median threshold for the guitar, for example, looks significantly lower for 0 ms (-10 dB) than for 1 and 10 ms (-7.5 dB). Equally, the quartet at 0 ms (-9.5 dB) looks significantly lower than for 1 ms (-6 dB). Friedman tests (critical p value = 0.05) showed that the effect of ICTD was significant for the guitar (p = 0.002), speech (p = 0.021) and quartet (p = 0.005). The effect was not significant for the oboe (p = 0.418) and conga (p = 0.788). A Wilcoxon test was subsequently conducted for the guitar, speech and quartet sources, with the Bonferroni Correction being applied to reduce type-I errors. The results showed the following. For the guitar, significant differences were identified between the 0 ms ICTD and both the 1 (p = 0.015) and 10 ms (p = 0.027) ICTDs. For speech, there were significant differences between 10 ms and both 0 ms (p = 0.012) and 1 ms (p = 0.024). For the quartet, the 0 ms and 10 ms (p = 0.015) ICTDs were significantly different from one another. These results generally agree with the notch edges shown in Fig. 4.6, although there are some small differences. It can therefore be concluded that the effect of ICTD on localisation threshold was significant.

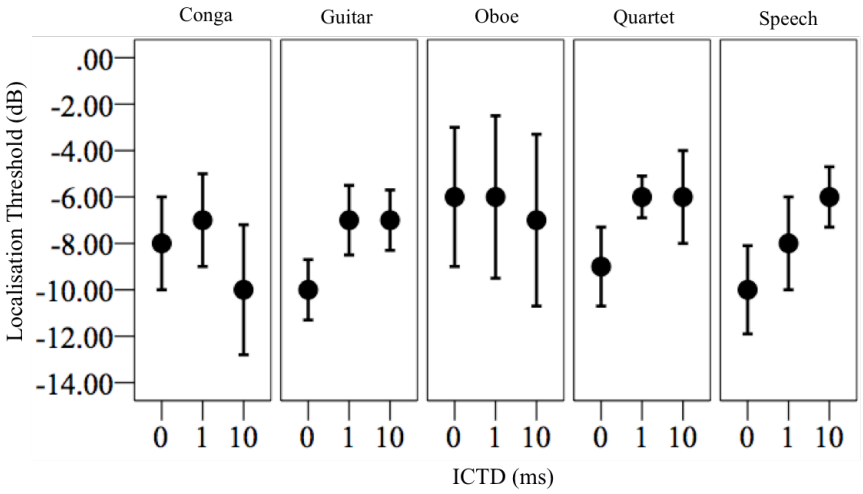

Fig. 4.6: Medians and associated notch edges with the results for both presentation methods combined.

4.1.3.3 The Effect of Sound Source

Fig. 4.7 shows the median localisation thresholds for each stimulus at each ICTD. The medians have been plotted with notch edges. The notch edges alone suggest that the effect of sound source on the localisation threshold was not significant. However, it should be noted that there are a number of notch edges that have minimal overlap (e.g. the guitar and oboe at 0 ms). A Friedman test conducted on the data indicated that the effect of sound source was significant for the 0 (p = 0.001) and 10 ms (p = 0.039) ICTDs. A Wilcoxon test was subsequently conducted to identify which pairs of stimuli were significantly different from one another, again with the Bonferroni Correction being applied. The results of this analysis showed no significantly different pairs for the 10 ms ICTD. This suggests that sound source had no significant effect on the localisation threshold for this ICTD, which agrees with the overlap of notch edges in Fig. 4.7. It should also be noted that, although the overlap between conga and speech is notably minimal, the effect size indicated a small effect (r = 0.28). For the 0 ms ICTD, significant differences were identified between the oboe and both the guitar (p = 0.01) and speech (p = 0.05). However, the effect size was not large in either case (r = 0.49 between the oboe and guitar and r = 0.42 between the oboe and speech). In addition, there is overlap between all notch edges. Further, the effect size (Kendall's W), which was calculated during the Friedman test, was low (0.262). Based on this analysis, it can therefore be concluded that the effect of sound source on localisation threshold was not significant. This would suggest that the same localisation thresholds could be applied to all natural sound sources tested in the present study.

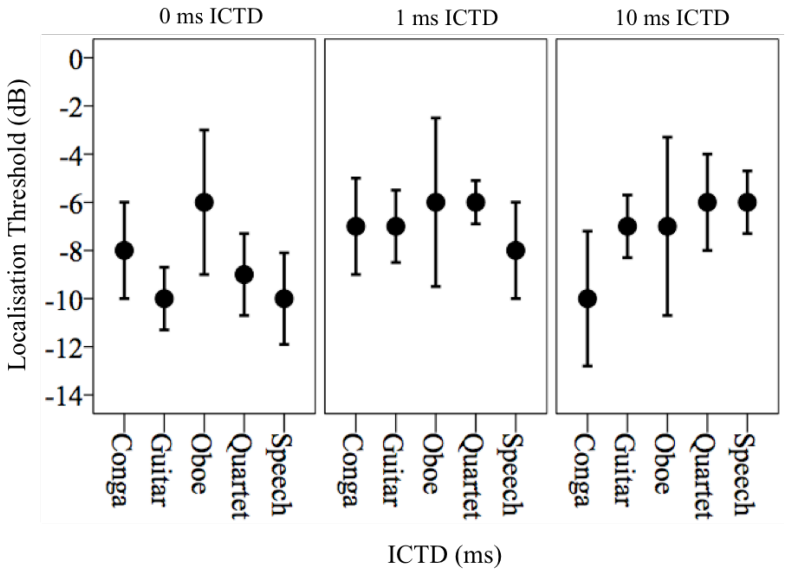

Fig. 4.7: Medians and associated notch edges, with the results for both presentation methods combined, arranged to compare the localisation thresholds for each sound source at each ICTD.

4.1.3.4 Localisation Thresholds for Combined Sources

As it was shown that the effect of sound source on the localisation threshold was not significant, the results for each sound source were combined. This is shown in Fig. 4.8. The median threshold for sources with 0 ms ICTD was -9.5 dB. Based on the notch edges, the threshold for this ICTD appears to be significantly lower than the -7 dB median threshold found for the 1 and 10 ms ICTDs. This significance was confirmed with the results of both Friedman ($p = 0.000$), and Wilcoxon ($p = 0.01$ between 0 ms and 1 ms, $p = 0.00$ between 0 ms and 10 ms) tests. It can therefore be concluded that the only variable whose effect was significant on the localisation thresholds obtained in the present study was ICTD. The effects of sound source and presentation method were not significant.

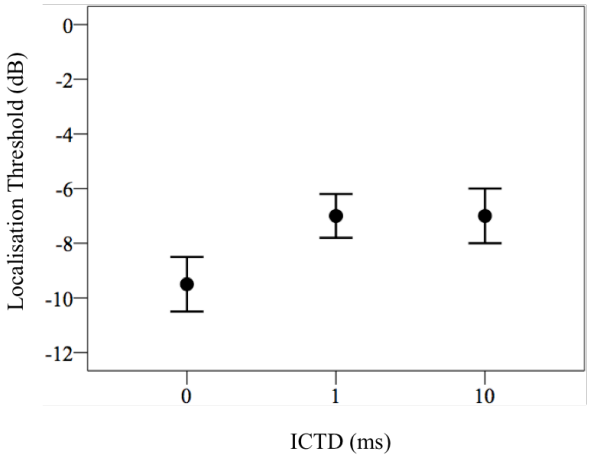

Fig. 4.8: Localisation thresholds for combined sources.

4.1.4 DISCUSSION

The experimental data obtained in the present experiment showed that localisation thresholds using the blanket reduction method are not significantly affected either by sound source or presentation method. Instead, the only variable whose effect was significant was ICTD. When the ICTD was 0 ms, the threshold was found to be -9.5 dB, which was significantly lower than the -7 dB found for ICTDs of both 1 and 10 ms.

4.1.4.1 The Sound Source Dependency of Localisation Thresholds

One of the key aims of the present experiment was to determine whether or not there existed a sound source dependency of localisation thresholds. The results of the experiment showed that, although the median localisation threshold varied for different sound sources, these differences were not significant. This agrees with the results reported in the Lee [2011] and Stenzl et al. [2014] studies, as well as with what was hypothesised in Section 4.1.1. Therefore, the null hypothesis that localisation thresholds using the blanket reduction method are not source dependent can be accepted. In Section 3.1.4.1, it was proposed that that the

mechanism that determines the localisation threshold for broadband pink noise is the relative balance of spectral energy provided by the main and height layers respectively, particularly in the 7-9 kHz range. A brief overview of that hypothesis is offered thus. Broadband pink noise presented as a vertically oriented phantom image is elevated with respect to the same source presented from the main layer only, arguably due to increased energy in the 7-9 kHz region. As the ICLD increases, this difference in energy decreases and the perceived difference in elevation between the two conditions lessens. At the localisation threshold this difference in energy is not sufficient to be interpreted as a difference in perceived elevation, although there is still more 7-9 kHz energy for the phantom image condition compared to the main layer only.

It seems logical that the above hypothesis would be as applicable to natural sound sources as it is to broadband pink noise. This is partly because both types of source are complex and feature a broad frequency spectrum. In addition to this, it is known from localisation studies that natural stimuli presented from the main layer only are perceived as being less elevated than those presented as vertically oriented phantom images [Barbour 2003, Tregonning and Martin 2015], which was demonstrated for pink noise in Experiment Two. It can therefore be argued that the determining factor as to whether or not the localisation threshold has been met for natural sound sources is dependent predominantly on the ICLD (i.e. the relative strengths of the main and height layers) and is independent of the fine spectral details of the sound source. This would have to be studied further, however, as it can be seen from Fig. 4.2 that the oboe and conga sources featured little spectral energy above 4 kHz, making it less clear how differences in energy in the 7-9 kHz range would affect the perceived elevation of these sound sources.

4.1.4.2 The Effect of Presentation Method

A further aim of the present experiment was to analyse how the localisation thresholds would be affected by changing the presentation method from vertical stereophonic to vertical quadraphonic, with the results obtained showing that the effect was not significant. This therefore means that the null hypothesis that the

effect of presentation method on localisation thresholds when using the blanket reduction method is not significant can be accepted. In Section 4.1.4.1, it was discussed how a key mechanism in determining whether or not the localisation threshold had been met might be the balance of spectral cues provided by the main and height layers respectively. The results of the present experiment would seemingly indicate that this balance was not affected when source presentation changed from vertical stereophonic to vertical quadraphonic.

In order to gain further objective insights into this result, the influence of the height layer on the frequency spectrum of the ear-input signal was analysed for each presentation method. For this, the spectral magnitude of the ear-input signal resulting from the main layer only was subtracted from that for both the main and height layers combined using the MIT's KEMAR HRIR database [Gardner and Martin 2000]. Fig. 4.9 shows this analysis when the ICLD was 0 dB (no height layer level reduction applied) and when the localisation threshold (-9.5 dB) was applied. From the plots, the following can be observed. Firstly, the spectral energy in the 7-9 kHz range for the 0 dB ICLD condition was dominant over that for the main layer only condition for both presentation methods, in a manner similar to that demonstrated in Section 3.1.4.1. In addition to this, when the localisation threshold was applied for each method, the difference in energy in this region was decreased between 8 and 10 dB, whereas that below this region was about 5 dB or less. This therefore supports the hypothesis that decreases in spectral energy in the 7-9 kHz region will result in the localisation threshold being met, with similar reductions for both presentation methods for a given ICLD likely being the reason for the non-significant effect of presentation method.

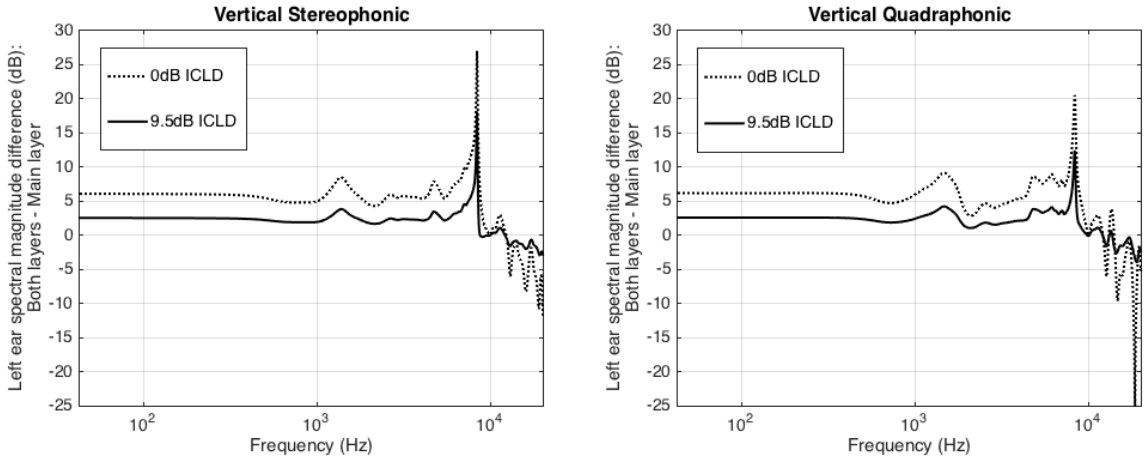

Fig. 4.9: Difference in spectral energy between the main layer only and phantom image conditions for both presentation methods with 0 dB ICLD between the main and height layers and 9.5 dB ICLD (localisation threshold).

The non-significant effect of presentation method is interesting when the results of previous localisation studies are considered. As shown in Experiment Two of the present thesis, as well as in the studies of Roffler and Butler [1968a] and Cabrera and Tiley [2003], for a single loudspeaker placed in front of the listener in the median plane, the perceived image of broadband noise tends to be localised accurately at the physical position of the loudspeaker. Conversely, the phantom centre image of the noise produced from stereophonic loudspeakers at the ear height (i.e., the main layer of the quadraphonic condition in the present experiment) would be elevated with respect to the physical position of the loudspeaker as reported by Lee [2016, 2017] and De Boer [1947]. Furthermore, a similar degree of difference in perceived elevation between the real and phantom image conditions would be observed for elevated loudspeakers (i.e. the height layers of the present study), based on data presented in Experiment Two of the present thesis and Lee [2016]. From the above, it can be inferred that, for 0 dB ICLD, sound sources presented using the vertical quadraphonic condition would be elevated with respect to those presented using vertical stereophonic, with the difference in perceived elevation being similar to that for the same sources presented from the main layer only. This would therefore imply that the perceived differences in elevation between the main layer only and phantom image conditions for a given ICLD would be similar for both presentation methods, as is demonstrated in Fig. 4.10, which might further explain why the effect of presentation method was not significant.

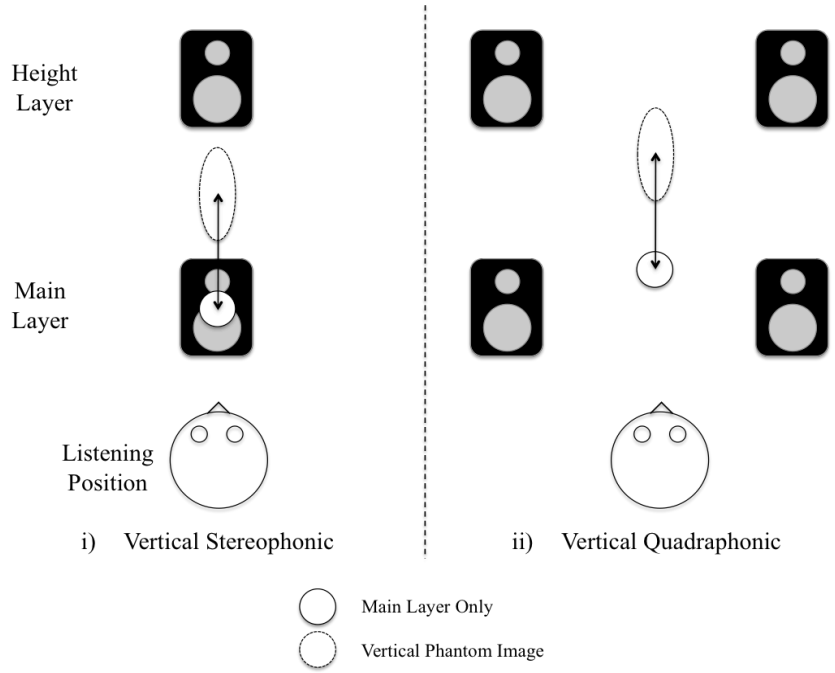

Fig. 4.10: Illustration to show how presentation method would not affect localisation thresholds despite the presence of the phantom image elevation effect.

4.1.4.3 The Localisation Dominance Effect

The results of the present experiment suggest that the only variable that had a significant effect on the localisation threshold was ITCD, with delays of both 1 and 10 ms requiring significantly less level reduction than did 0 ms. However, there was no condition whereby ICTD alone was sufficient for the localisation threshold; ICLD was always necessary. This result indicates that the precedence effect, in which an ICTD greater than 1.1 ms between coherent loudspeakers located on the horizontal plane will cause the resultant sound source to be localised at the exact position of the earlier loudspeaker [Wallach et al. 1949], is not a feature of median plane localisation. This therefore means that the null hypothesis that the precedence effect does not operate for vertically oriented phantom images can be accepted. This agrees with the conclusions reported in the Lee [2011] and Stenzl et al. [2014] studies, as well as those reported in Experiments One and

Two of the present thesis. What is suggested, however, is somewhat of a localisation dominance effect, whereby the presence of an ICTD biases localisation towards the earlier loudspeaker. This can be considered as being similar to summing localisation for horizontal stereophony [Blauert 1997] and has been shown to operate in numerous median plane localisation studies including those of Somerville et al. [1965], Blauert [1971], Litovsky et al. [1997] and Tregonning and Martin [2015]. If it is the case that the earlier loudspeaker becomes dominant in determining perceived source location in the median plane, then this might explain why significantly less ICLD was necessary to meet the localisation threshold for the 1 and 10 ms ICTDs compared to for the 0 ms condition. Simply, the presence of a delay resulted in less perceived difference in elevation between the main layer only and phantom image conditions, which meant that less ICLD was needed to meet the localisation threshold. Incidentally, this would also mean that there is some correlation between the perceived initial location of the source and the resultant localisation threshold, which is similar to what was reported for the broadband source in Chapter Three.

Despite the above hypothesis, it should be noted that higher localisation thresholds as a result of a localisation dominance effect has not been reported in previous localisation threshold experiments. Both Lee [2011] and Stenzl et al. [2014] reported that there was no significant difference between the localisation thresholds in the range of 0-10 ms. A potential reason for this difference could relate to the time and level alignment employed in the respective studies. In the present study, time and level alignment was applied to the main layer in order to accommodate for the difference in distance between the loudspeakers in each layer and the listening position. However, Lee [2011] chose not to conduct time and level alignment in order that his setup could more accurately emulate pre-existing reproduction formats. Based on the dimensions of his experimental setup, this therefore meant that for the 0 ms condition the height loudspeaker signal arrived at the listening position delayed with respect to that from the main loudspeaker by 0.7 ms, which is similar to the 1 ms condition of the present experiment. However, it should be noted that Stenzl et al. [2014] did choose to time and level align their loudspeaker layers and produced results similar to those reported by Lee [2011]. It is therefore unclear why in the present experiment significantly less level reduction was necessary

in the presence of delay when such a result was not recorded by either Lee [2011] or Stenzl et al. [2014]. This requires further study.

4.1.4.4 Practical Implications

The localisation thresholds obtained in the present experiment can be used to influence both the placement of microphones and the rendering of 3D images in the context of 3D audio. In either case, the results are indicative as to the necessary attenuation of direct sounds in the height layer in order that the perceived location of the main channel signal is unaffected. With respect to 3D microphone configurations, a series of techniques have already been proposed by Lee [2011]. In that study, it was suggested that, since the localization threshold needs to be applied to the height layer microphone signals in order that the source image is located at the position of the main layer, directional microphones should be used for the height layer rather than omni-directional ones. The results of the present experiment support this suggestion. In addition, it was proposed that, in case of using microphones with an 'ideal' cardioid polar pattern (i.e. -6 dB attenuation at 90°), the necessary ICLD could be achieved for both vertically coincident and spaced configurations by angling the height layer of microphones at least 90° away from the direct sound. However, the data reported in the present study indicates that a minimum angle of 105° would be necessary in the case that the main and height layers are spaced apart, whilst for a coincident configuration the angle should be 115°. This would provide attenuation of direct sounds in the height layer at the localization thresholds for 0 ms and 1 ms found in the present study: around -7 and -9.5 dB, respectively.

The results of the present experiment are also considered to be useful for vertical image rendering in 3D sound mixing and upmixing applications. They indicate that direct sounds can be present in the height layer provided they are attenuated with respect to those in the main layer by either 9.5 dB (in the case of 0 ms ICTD) or 7 dB (in the case of 1-10 ms ICTD) without the perceived location of the main channel signal being affected. Such a technique could have potentially pleasing effects, such as an increase in perceived

VIS. However, it is currently not clear how the timbre of the main channel signal would be affected by such a technique and, further, if the end result would be pleasing. It can be seen from Fig. 4.9, for example, that the resultant spectrum of the signal is different at the localisation threshold compared to main layer only presentation, with a notable peak in the 7-9 kHz range. Alongside this, Halmrast [2000] suggested that secondary vertical sources would result in orchestral music sounding 'boxy', whilst Barron and Marshall [1981] indicated that timbral colouration as a result of vertical reflections are more audible than for lateral reflections. It would be necessary then to evaluate first of all what the perceptual differences are between the main layer only and vertical phantom image conditions with the localisation thresholds applied and further if the threshold conditions are considered as being preferable. Such a study would make it possible to determine whether the localisation threshold should be applied or, conversely, if the direct sound in the height layer should be either masked or absent entirely. This would provide further insights on both image rendering and microphone techniques in the context of 3D audio production. This is considered in Chapter Five of the present thesis.

It should also be noted that there are some, limited, applications with respect to the vertical panning of sound sources. It is indicated by the results that, depending on the ICTD, the threshold value for a source to be fully panned to the main loudspeaker layer is in the range of 7-9 dB, which agrees with the vertical localisation studies of both Barbour [2003] and Wendt et al. [2014]. However, further study would be needed to determine if this value is applicable to source localisation at the position of the height layer and, further, how changes in both ICLD and ICTD affect the perceived localisation of the resultant phantom image in between these extremes.

4.1.5 Conclusion

The present experiment was an analysis of localisation thresholds for natural sound sources using the blanket reduction method. In the experiment, the effects of sound source, ICTD and presentation method on the

localisation threshold were examined. Anechoically recorded conga, quartet, speech, guitar and oboe sources were presented to subjects in a natural listening environment using two conditions: vertical stereophonic and vertical quadraphonic. For each condition, the loudspeakers were divided into two layers, being 'height' (30° elevation) and 'main' (0° elevation). Delays ranging from 0-10 ms were applied to the height layer with respect to the main. Subjects sat a listening test in which the minimum amount of attenuation necessary in the height layer for the resultant phantom image to match the position of the same source presented from the main layer alone was considered.

The results of the experiment showed that the localisation thresholds were affected only by ICTD. For delays of 0 ms the threshold was -9.5 dB, which was significantly lower than the -7 dB found for 1 and 10 ms. That less ICLD was necessary in the presence of a delay was interpreted based on the existence of a localisation dominance effect. In addition, attempts to explain the non-significant effect of sound source were made based on the hypothesis that the primary mechanism to determine whether or not the localisation threshold had been met was the balance of spectral energy provided by the main and height layer, particularly in the 7-9 kHz range, which is not related to the spectrum of the source itself. This hypothesis also explained the non-significant effect of presentation method, with it being demonstrated that the reduction in the difference in energy between the main layer only and phantom image conditions in the 7-9 kHz region was similar for both methods for a given ICLD.

The practical implications of the results obtained in the study were also discussed. In particular, differences between suggestions made in previous studies and those indicated by the present results were considered. It was also stated that further study would need to be conducted into the spatial and timbral effects when the localisation thresholds are applied in order to determine whether or not it would be more appropriate for the direct sound in the height layer to be masked.

4.2 EXPERIMENT FOUR: LOCALISATION THRESHOLDS FOR OCTAVE BANDS (BAND REDUCTION)

One of the primary aims of the present thesis is to develop a method of applying the localisation threshold whereby the direct sound in the height layer undergoes frequency-dependent attenuation (band reduction). The results of Experiment One suggested that such a method would be possible in theory, with a frequency dependency of localisation thresholds being identified. However, despite this finding, it is arguably not possible to use the results of that experiment directly in order to develop a band reduction method. The primary reason for this is that the experiment was conducted in anechoic conditions, as opposed to in a natural listening environment. This is a salient point given the discussions regarding the potential effects of perceived VIS on the frequency dependency of localisation thresholds. It could be the case that this perception is in some way affected by the presence of reflections, with no study, of which the author is aware, considering this issue. Therefore, seeing as the primary application for a band reduction method would arguably be in non-anechoic conditions, it is necessary to determine how the frequency dependency of localisation thresholds is affected by the presence of reflections in order that a band reduction method can be developed.

A further limitation of Experiment One was that localisation thresholds were only considered for stereophonic centre and height centre loudspeakers. It should be noted that conventional reproduction systems for 3D sound reproduction, such as Auro 3D [Auro Technologies 2016] do not always incorporate a height centre loudspeaker. Instead, elevated left and right loudspeakers are often used. It would be of interest then to determine whether the frequency dependency of localisation thresholds would differ significantly for vertical quadraphonic presentation compared to vertical stereophonic. Although the results of Experiment Three suggested that the localisation thresholds were consistent for both presentation methods, the ongoing discussions suggesting that different mechanisms are used to determine the localisation threshold for band-

limited stimuli compared to complex sources necessitates that such an analysis is undertaken in the context of a band-reduction method.

Additionally, it is of further interest to determine how the duration of the test stimuli affects the localisation thresholds obtained. Natural sources contain a blend of continuous and transient information that is characteristic to the source itself. If it is the case that the localisation thresholds for burst (i.e. transient) noise are different compared to those for continuous noise then it can be argued that more transient sources, such as percussion, should be treated differently under the band reduction method than should be continuous sources, such as strings.

From the above background the following research questions were derived:

- Can the frequency dependency of localisation thresholds be demonstrated in a room in which reflections are present?
- How are localisation thresholds under such conditions affected by the method of stimulus presentation, the duration of the source and the ICTD?

This section is organised as follows. Firstly, the experimental design is discussed. Following this, the results are presented and analysed. The section concludes with discussions pertaining to the results and their implications for the research aims.

4.2.1 EXPERIMENTAL HYPOTHESIS

The first null hypothesis for this experiment is that the frequency dependency of localisation thresholds will not be maintained in a natural listening environment. It is anticipated that this null hypothesis will be rejected for the following reasons. In Section 3.1.4.1, it was hypothesised that the frequency dependency of

localisation thresholds in anechoic conditions was caused by frequency-dependent differences in perceived VIS between the main layer only condition (the reference) and the vertically arranged stereophonic phantom image conditions (the test stimuli). Assuming this hypothesis is an accurate explanation of the results, the following can be suggested. When a sound source is in the presence of reflections, one of the most perceptible differences, with respect to the same source presented in anechoic conditions, will be an increase in the spatial extent of the image (i.e. the image will appear larger). However, it is also the case that the main layer only and phantom image conditions in the present experiment will be analysed in the same space. Therefore, any increases in the size of the main layer only image, as a result of the presence of reflections, would arguably also be apparent for the phantom image conditions. Consequently, it is expected that the frequency-dependent differences in VIS would be maintained in a natural listening environment, with the resultant localisation thresholds being similar to those obtained in anechoic conditions.

An additional null hypothesis for the present experiment is that the frequency dependency of localisation thresholds is not dependent on the presentation method. It is anticipated that this null hypothesis will be rejected. It has already been shown in Experiment Two, for example, that the pitch-height effect governs the perceived elevation of octave band stimuli presented as vertically oriented phantom images from directly in front of the subject. Therefore, differences in perceived elevation as a result of the phantom image elevation effect are not thought to be relevant to band-limited stimuli. In addition, even if the phantom image elevation effect was a factor, the results of Experiment Two indicated that there was no correlation between the perceived elevation of the octave band stimuli and the localisation thresholds found in Experiment One. Based on these discussions, there seems to reason to believe that the localisation thresholds would not differ greatly for vertical quadraphonic and stereophonic presentation.

4.2.2 EXPERIMENTAL DESIGN

This experiment was generally similar to Experiment One, in that the effect of frequency on the localisation thresholds for octave bands of noise was analysed. The general methodology required subjects to compare the difference in perceived elevation between the main layer only and phantom image conditions, and adjust the amplitude of the height layer until the two sources were perceived to be co-located. The physical setup, subjects and test method were identical to those used in Experiment Three. As with that experiment, the vertical quadraphonic and stereophonic presentation methods were tested, with ICTDs of 0, 1 and 10 ms being applied to the height layer with respect to the main.

The test stimuli used for the experiment were created by generating pink noise in MATLAB and filtering it into nine octave bands, with centre frequencies ranging from 63 Hz to 16 kHz. A limitation with the FFT filter used for Experiments One and Two is that it gave little control over the frequencies outside the pass band. So that the frequency content of the octave band stimuli could be more tightly controlled for the present experiment, an 8^{th} order Butterworth filter (-48 dB/oct roll off) was used. The original pink noise source was also included, giving 10 sources in total. The stimuli were presented to subjects both continuously (30 s duration, 1 s onset/offset) and as bursts (200 ms duration, 10 ms onset/offset, 1 per second). In total there were 120 stimuli (ten sources, three delay times, two presentation methods, two durations).

The amplitude of the broadband pink noise source at the listening position was 70 dB LAeq when presented from the main layer only. The amplitude of each octave band was kept relative to the broadband pink noise, as opposed to being individually level matched. This was because ultimately the localisation thresholds obtained would be applied to natural sound sources. It was therefore decided to obtain thresholds for octave bands at equal energy as opposed to at equal amplitude. The amplitude for each stimulus when presented as a phantom image was dependent on the amplitude of the height layer relative to that of the main, which was to be varied by the subject using the AMOA method described in Section 4.1.2.4.

During the test, subjects were strictly instructed to face forwards, keeping their head still and using only their eyes to look at the test interface. The heads of subjects were not fixed, however head movements were monitored using a motion tracker device [Johnson et al. 2016]. The tracker instructed subjects if their head position had deviated from an acceptable range of natural motion, as described in Section 4.1.2.4. Additionally, a guide point for the ear height and distance was placed on the right hand side of the subject to help maintain the correct listening position throughout the test. Prior to the start of each test, all subjects sat a supervised practice, which utilised a speech source, in order to ensure that the instructions were understood. The test was completed in four sittings of 30 stimuli each, with each sitting taking around 15 minutes. The order of tests and stimuli was randomised for each subject.

4.2.3 DATA ANALYSIS AND RESULTS

Levene and Shapiro-Wilk tests were first conducted, using the SPSS software, in order to determine the suitability of the collected data for parametric statistical analysis. The results of the Levene test showed homogeneity of variance for all sound sources, whilst the Shapiro-Wilk test showed that not all scores in each condition featured normal distribution. This therefore meant that the assumptions of Analysis of Variance (ANOVA) were violated. For these reasons, non-parametric tests were chosen for the statistical analysis.

4.2.3.1 The Effect of Presentation Method

Fig 4.11 shows the effect of presentation method on the median localisation thresholds obtained in the study. The medians have been plotted with notch edges. The notch edges for each presentation method generally show considerable overlap, which would seemingly indicate that the effect of presentation method on the localisation thresholds obtained was not significant. A series of Wilcoxon tests were further conducted in order to determine whether the effect of presentation method was significant for any of the stimuli. The

Wilcoxon test results agreed with the notch edges in that no significant pairs were identified ($p > 0.05$ for all pairs). It can therefore be concluded that the effect of presentation method on the localisation thresholds was not significant.

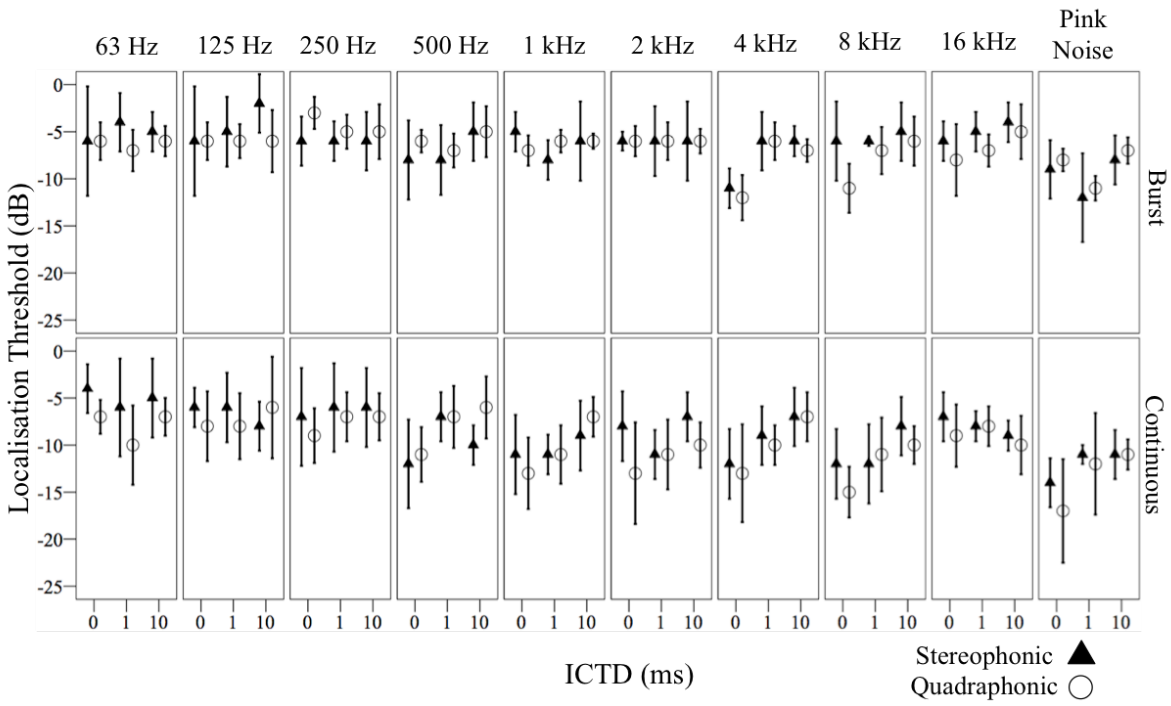

Fig 4.11: Medians and associated notch edges showing the results of Experiment Four.

4.2.3.2 The Effect of Signal Duration

As it was determined that the effect of presentation method on the localisation thresholds obtained was not significant, the results for the presentation methods for each stimulus were amalgamated. Fig. 4.12 shows the effect of signal duration on the localisation thresholds for the combined presentation methods. The notch edges suggest that there was a significant effect of signal duration for the 500 Hz (0 and 10 ms) and the 1 (0 ms), 2 (0 and 1 ms) and 8 kHz (0 and 1 ms) octave bands, as well as for the broadband source (0 and 10 ms). In each of these cases, the median localisation threshold was lower for the continuous stimuli than for the bursts. Wilcoxon tests were conducted to further analyse this. The results suggested the following. Firstly,

the effect of signal duration was significant for the broadband source for 0 and 10 ms (p = 0.001 and p = 0.009). In addition, for the 1 and 2 kHz sources the effect was significant for both 0 and 1 ms (p = 0.002 and p = 0.006 for 1 kHz, p =0.003 and p = 0.017 for 2 kHz). However, the Wilcoxon test results also suggested signal duration was significant for 1 kHz at 10 ms (p = 0.008). Despite this result, there is clear overlap of the notch edges. In addition, the Pearson's correlation coefficient (r), which was calculated based on Cohen [1988], was 0.441, which does not indicate a large effect [Cohen 1988]. Therefore, it can be concluded that signal duration was not significant for 1 kHz with 10 ms ICTD.

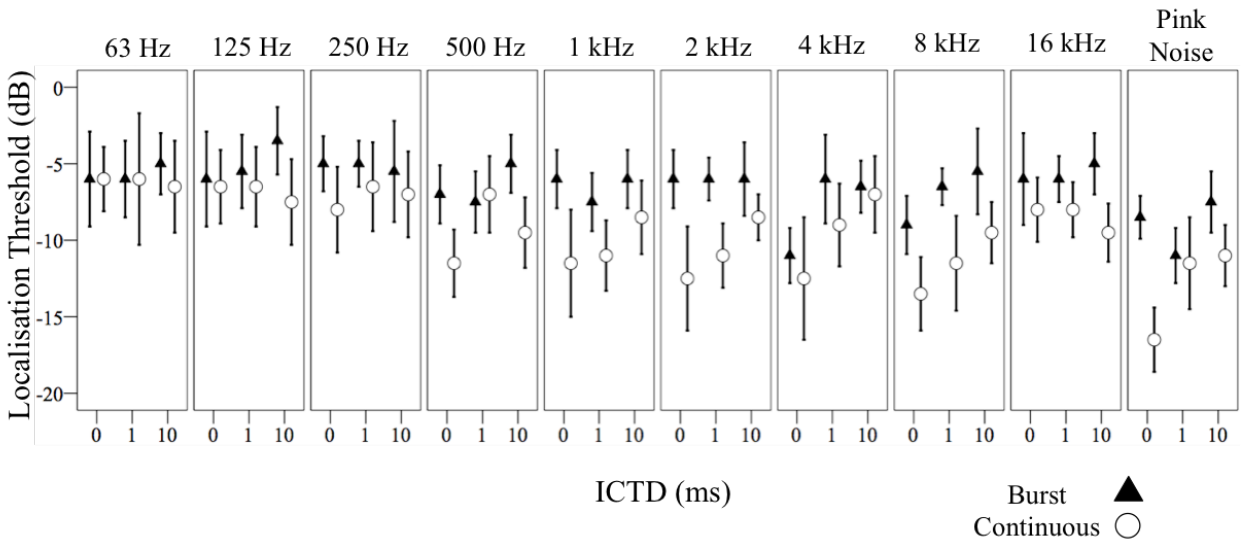

Fig 4.12: Median and associated notch edges showing the results of Experiment Four with the results for each presentation method combined.

With respect to the 500 Hz source, there was some agreement with the notch edges and the Wilcoxon test results in that the effect of signal duration was significant for 10 ms (p = 0.004). However, the results also showed no significant effect for 0 ms (p = 0.084). For this ICTD, r was 0.29, which indicates a small effect [Cohen 1988]. However, despite the fact that the Wilcoxon and Pearson's correlation coefficient results indicate that the effect of signal duration was not significant, there is a clear lack of overlap between the notch edges. Based on this, it can be reasonably be concluded that the effect of signal duration was significant for the 500 Hz band with 0 ms ICTD.

The Wilcoxon test results also suggested that the effect of signal duration was significant for the 8 kHz octave band for all ICTDs (p = 0.001, 0.001 and 0.009). This mostly agrees with the notch edges, although there is overlap for 10 ms, which suggests that the effect of signal duration was not significant. Additionally, the results suggested significant effects for the 63 and 125 Hz bands with 10 ms ICTD (p = 0.041 and p = 0.006) and further that at 16 kHz, signal duration was significant for all ICTDs (p = 0.012, 0.009 and 0.018). Each of these results disagrees with the notch edges. Despite these results, the Pearson's correlation coefficient suggested no large effect sizes for the 63 Hz, 125 Hz or 16 kHz bands ($r < 0.5$ in all cases). Therefore, it can be concluded that signal duration was not significant at these frequencies.

4.2.3.3 The Effect of ICTD

Fig. 4.13 shows the effect of ICTD on the median localisation threshold. As the effect of signal duration was found to have a significant effect, the data for the continuous and burst stimuli has not been amalgamated. For the burst stimuli at 4 kHz, the median localisation threshold appears to be significantly lower for the 0 ms ICTD than for both the 1 and 10 ms ICTDs. Additionally, for the pink noise bursts the 1 ms ICTD looks significantly lower than the 10 ms ICTD, whilst the overlap between 0 and 1 ms is notably small. These observations were confirmed with the results of a Friedman test (p = 0.017 for 4 kHz, p = 0.015 for pink noise). It should be noted that the Friedman test also identified a significant effect for the 500 Hz octave band (p = 0.035). This result is in positive disagreement with the notch edges, which overlap heavily. A series of Wilcoxon tests were conducted to analyse which pairs of ICTDs were significantly different from one another for the 500 Hz and 4 kHz octave band and broadband pink noise sources. As the analysis required multiple pairwise comparisons the Bonferroni correction was used to avoid any type-I errors [Simner 1986]. The Wilcoxon test results showed no significantly different pairs for the 500 Hz stimuli. In addition, the effect size (Kendall's W), which was calculated during the Friedman test, was low (0.185). It can therefore be concluded that the effect of ICTD was not significant for this octave band. For the pink noise source, the Wilcoxon test also showed no significantly different pairs. However, given the minimal

overlap between 1 and 10 ms and the results of the Friedman test, it can be concluded that ICTD was significant for this source. For the 4 kHz band, the Wilcoxon test identified a significant difference between 0 and 10 ms (p = 0.009), which agrees with the notch edges. There was however no significant difference between 0 and 1 ms according to the Wilcoxon test (p = 0.180). Despite this, there is clearly no overlap between the notch edges so it is reasonable to conclude that the difference here was significant.

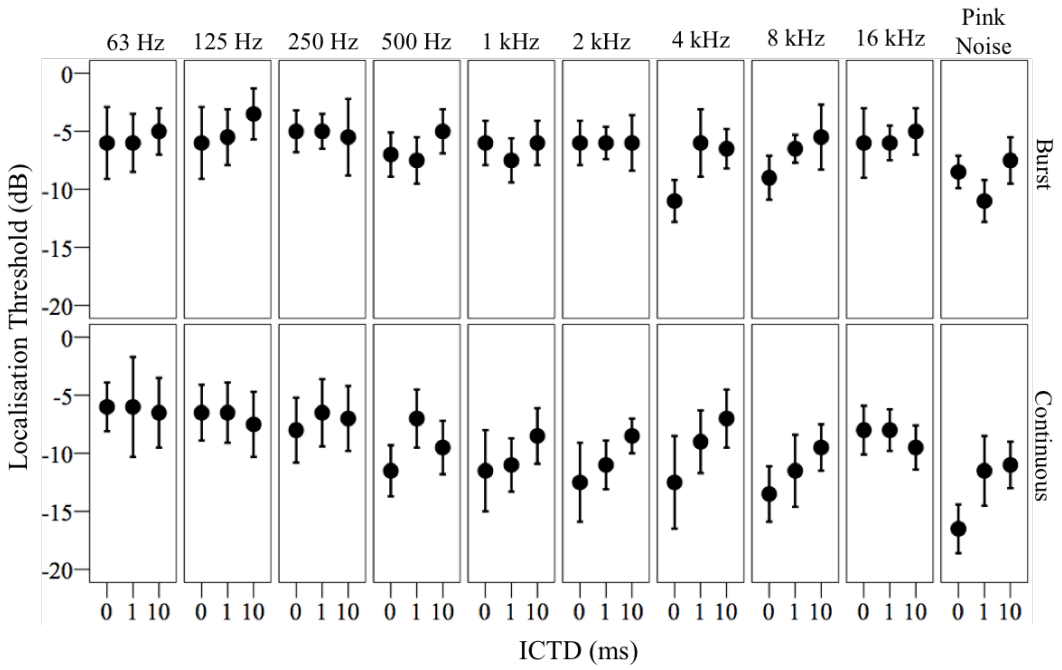

Fig. 4.13: Medians and associated notch edges showing the results of Experiment Four. The data has been arranged to show the effect of ICTD.

For the continuous noise, ICTD appeared to have little effect for the majority of the stimuli. However, there looks to have been some effect at 8 kHz, with the threshold for the 0 ms ICTD appearing to be significantly lower than that for the 10 ms ICTD. The same effect can be seen for the broadband pink noise. This significance was confirmed with a Freidman test. However, along with the 8 kHz (p = 0.001) and broadband pink noise (p = 0.025) showing statistical significance, the Friedman test also suggested significance for both the 1 (p = 0.049) and 4 kHz (p = 0.005) octave bands. This is at odds with the notch edges, which show considerable overlap between each ICTD for both stimuli. A series of Wilcoxon tests were conducted to

further analyse the effect of ICTD for the 1, 4 and 8 kHz octave bands, as well as for the broadband pink noise. The Bonferroni correction was applied. The results agree with the notch edges in that there was no significant effect for 1 kHz and that at 8 kHz the significant difference was between 0 and 10 ms ($p = 0.003$). However, the Wilcoxon test results also suggested that the interactions between 0 ms and both the 1 and 10 ms ICTDs was significant for 4 kHz ($p = 0.036$ and 0.003 respectively), whilst for the broadband pink noise no significant pairs were identified. However, for the 4 kHz octave band the results show a clear overlap of notch edges. In addition, the effect size (Kendall's W) was low (0.293), whilst the Pearson's correlation coefficients suggested that the effect was not large ($r < 0.5$ for all comparisons). Therefore, it can be concluded that the effect of ICTD was not significant for 4 kHz. With respect to the broadband pink noise, the Friedman test suggested statistical significance, whilst there was also no overlap between the notch edges between 0 and 10 ms. As a result, it can be concluded that the effect of ICTD was significant for the continuous broadband pink noise.

4.2.3.4 The Effect of Frequency

Fig. 4.14 shows the effect of frequency for each stimulus duration at each ICTD. For the burst stimuli, the notch edges overlap for all frequencies for the 10 ms ICTD, suggesting that the effect of frequency was not significant. The median thresholds are also similar, each being around -6 dB. However, for the 0 and 1 ms ICTDs there are pairs of stimuli that show no overlap. At 0 ms, the most noticeable difference is for the 4 kHz octave band, which has a median threshold that looks to be significantly lower than all frequencies below it (-12 dB compared to around -6 dB). Additionally, for 1 ms the threshold for the broadband pink noise appears to be significantly lower than all other stimuli (around -11 dB compared to around 6-8 dB). The significant effect of frequency at 0 and 1 ms was confirmed with a Friedman test ($p = 0.000$ for both). The effect was not significant for 10 ms ($p = 0.100$).

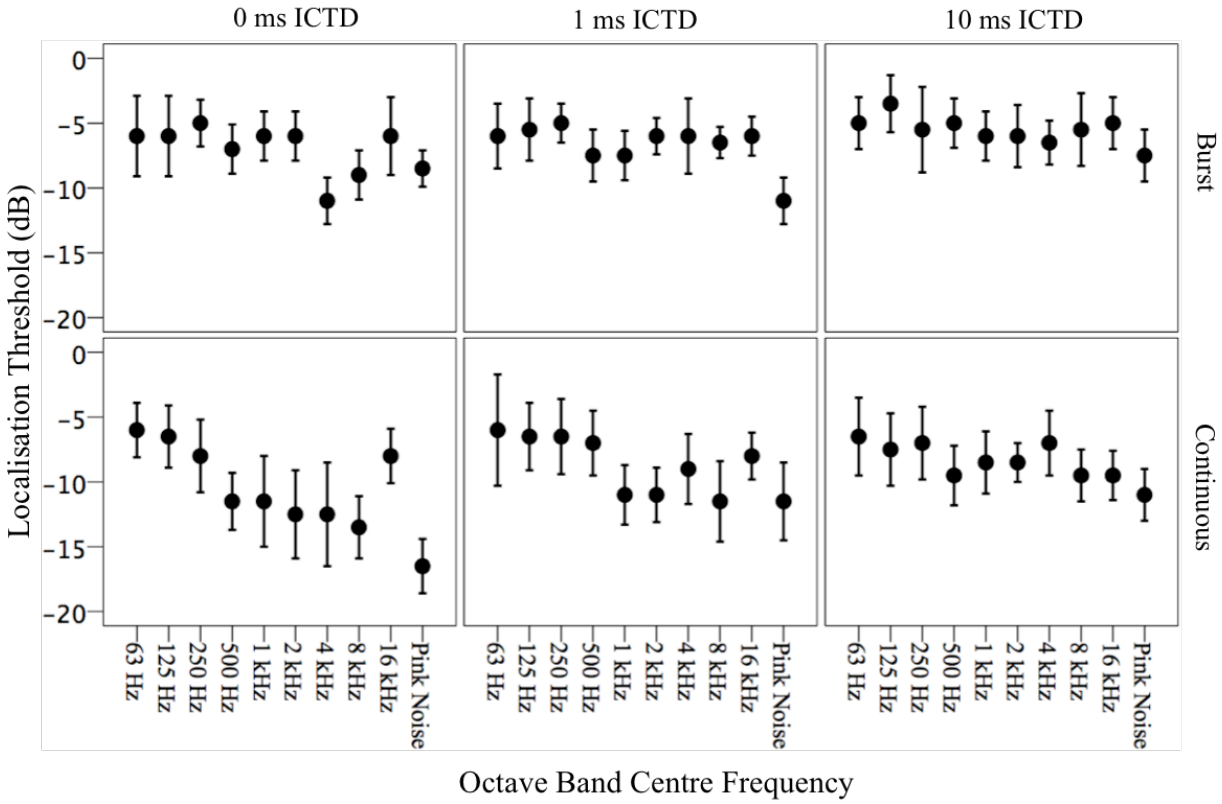

Fig. 4.14: Medians and associated notch edges showing the results of Experiment Four. The data has been arranged to show the effect of frequency.

Wilcoxon tests were conducted for the 0 and 1 ms ICTDs, with the Bonferroni correction being applied. For 0 ms, the Wilcoxon test results suggest that there were no significantly different pairs. This disagrees with the notch edges, which show the threshold for 4 kHz in particular to be notably lower than that those for the octave bands lower in frequency. This result might be a type-II error, being a false negative [Lieberman and Cunningham 2009], as a result of the use of the Bonferroni correction, as it is clear from the notch edges and Friedman tests that the effect of frequency was significant for the 0 ms burst stimuli. For the 1 ms ICTD, the threshold for broadband pink noise was significantly lower than the 63 Hz (p = 0.000), 125 Hz (p = 0.045), 250 Hz (p = 0.045) and 16 kHz (p = 0.045) octave bands. This shows some agreement with the notch edges although the notch edges do in fact suggest that the threshold for the broadband pink noise was significantly lower than all of the octave bands. Additional differences were found between 8 kHz and 250 Hz (p = 0.045), which is not in agreement with Fig 4.14.

For the continuous stimuli, the notch edges suggest that the effect of frequency was only significant for the 0 ms ICTD. In particular, the 16 kHz octave band appears to be significantly higher than both the 8 kHz band and the broadband pink noise. The notch edges for the other ICTDs all overlap. A Friedman test conducted on the data showed that the effect of frequency was significant for the 0 ms ICTD (p = 0.000), agreeing with the notch edges. However, the test also suggested that the effect was significant for the 1 ms (p = 0.000) and 10 ms (p = 0.001) ICTDs. Wilcoxon tests were subsequently conducted for each ICTD (Bonferroni correction applied). For 0 ms, the results suggested that the broadband pink noise was significantly lower than the octave bands with centre frequencies of 500 Hz and below, which agrees with the notch edges. In addition, the 8 kHz band was significantly lower than the 250 Hz bands and below, whilst 4 kHz was significantly lower than 63 Hz (p = 0.045). This suggests overall that at 0 ms the high frequency octave bands (4 and 8 kHz) have significantly lower localisation thresholds than do the low frequency stimuli (63-250 Hz). The results for 1 ms suggest that 8 kHz was significantly lower than 250 Hz (p = 0.045) and that the pink noise was significantly lower than the 16 kHz band (p = 0.045) and the frequencies at and below 250 Hz. This result disagrees with the notch edges. Moreover, for 10 ms the Wilcoxon test revealed significant differences between the broadband pink noise and 63 Hz (p = 0.000). However, it is clear from the notch edges that there is some overlap between the two stimuli and, in addition, the effect size (Kendall's W) was low (0.179). Overall this analysis suggests that the effect of frequency was only significant for the 0 ms ICTD.

4.2.4 DISCUSSION

The results of the present study have shown that the frequency dependency of localisation thresholds is maintained in the presence of reflections. However, it is also the case that this was somewhat dependent both on the ICTD and the duration of the signal. When the stimuli were presented as bursts, the effect of frequency was notably weak, with the median thresholds generally being consistent across the spectrum. The only exception to this was for the 0 ms ICTD, where the thresholds for 4 and 8 kHz were notably lower than those for the other bands (although only 4 kHz was significantly lower). For the continuous stimuli, the

effect of frequency was more apparent. For all three ICTDs, a general decrease in the median threshold with increases in frequency was observed, with there being a significant effect for both 0 and 1 ms. This result then was somewhat similar to the results of Experiment One, although in that study there were more significantly different pairs with respect to the effect of frequency. Consequently, the null hypothesis that the frequency dependency of localisation thresholds is not maintained in a natural listening environment can be rejected.

Additional results of interest obtained in the study were as follows. Firstly, the effect of presentation method on the localisation thresholds obtained was not significant. This indicates that the same frequency-dependent thresholds can be used for both vertical stereophonic and quadraphonic presentation in situations whereby the resultant image is formed on the median plane. This also means that the null hypothesis that the frequency dependency of localisation thresholds is not affected by the presentation method can be accepted. It should also be noted that the duration of the signal had a significant effect on the thresholds obtained for both the mid (500 Hz – 2 kHz) and high (8 and 16 kHz) frequencies, as well as for the broadband pink noise (although there were notable differences in the median threshold for other octave bands).

4.2.4.1 Comparison with the Results of Experiment One

Between the results of the present experiment and those reported in Experiment One, it can be seen that localisation thresholds for continuous octave band stimuli are frequency dependent both for natural and anechoic listening environments. In Fig. 4.15, the results of the respective experiments have been directly compared. In the case of the present study, the median thresholds for 125 Hz – 8 kHz and broadband pink noise, when presented using the vertical stereophonic condition, have been shown. The medians have been plotted with notch edges.

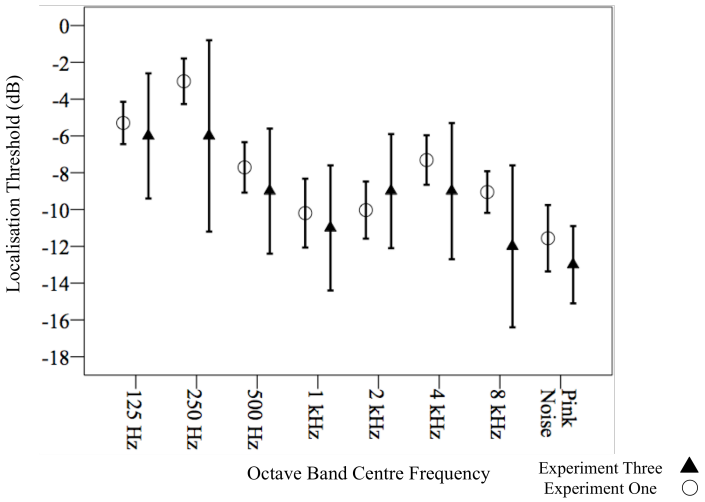

Fig 4.15: Comparison between band reduction localisation thresholds for continuous stimuli presented using the vertical stereophonic condition.

In Section 4.2.1, it was hypothesised that the presence of reflections would result in localisation thresholds similar to those obtained in anechoic conditions. It was supposed that the test stimuli would appear larger, as a result of the reflections, but that any changes would apply to both the main layer only and phantom image conditions. This would therefore mean that the frequency-dependent differences in perceived VIS, which were postulated as being the reason for the frequency dependency of localisation thresholds, would be maintained. By extension, there would be no reason for the localisation thresholds to differ compared to those obtained in Experiment One. However, if Fig. 4.15 is considered, it can be seen that this hypothesis was not accurate. Instead, the thresholds obtained in the presence of reflections were generally in the range of 1-3 dB lower than those for anechoic conditions (although that for 2 kHz was around 1 dB higher), whilst the notch edges are also much wider (7-9 dB compared to 2-4 dB). This latter result indicates that, although there was as significant effect of frequency for both conditions, the effect was much stronger for Experiment One (anechoic) compared to in the present study (listening room). The reasons for these results can arguably be explained in two ways, i) the effect of reflections and, ii) differences in the test method between the two experiments.

With respect to reflections, consideration was given to the effect that the first arriving reflection would have on the spectral content of the test stimuli. Given the dimensions of the listening room in which the experiment was conducted, it can be argued that the first arriving reflection would have been from the floor, as demonstrated in Fig. 4.16. Further, based on the nature of the test setup, as described in Section 4.1.2.1 it can be calculated that this reflection travelled around 2.32 m to reach the listening position. This would have meant that the time delay between the arrival of the direct sound and the first arriving reflection was around 1.4 ms. Fig. 4.17 shows the resultant spectrum when a sine sweep is summed with a version of itself delayed by this amount. Here it can be seen that there is a strong comb filtering effect, with heavy and frequent attenuation across the frequency spectrum. Based on this analysis, the following can be suggested. The presence of reflections, most noticeably from the floor, would have resulted in comb filtering of the test stimuli. It can be argued that the destructive nature of this effect might have had an influence on the frequency dependent differences in perceived VIS, which have been postulated as being important for a frequency dependency of localisation thresholds. It should be noted, however, that this would have to be studied further. The primary reason for this is that the hypothesis relies on a simulation only, with the effect of reflections on the resultant spectrum of the test stimuli not having been directly measured. In addition, the simulation did not account for the absorption coefficients of the floor, which might mean that the effects of comb filtering as shown in Fig. 4.17 are exaggerated somewhat. Nevertheless, the results do at least show that reflections may have affected the spectral content of the test stimuli and this might explain why the effect of frequency was less strong in the presence of reflections compared to in anechoic conditions.

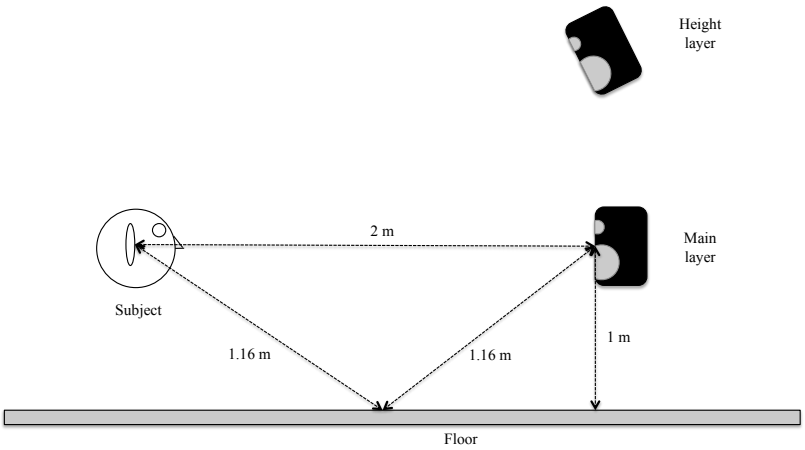

Fig. 4.16: Difference in distance travelled between the direct sound and the first arriving reflection (floor).

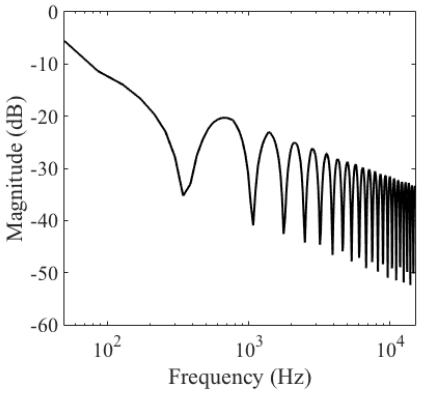

Fig. 4.17: The effect of a floor reflection delayed by 1.4 ms on the resultant spectrum of a sine sweep.

In terms of the test method, the following can be suggested as having caused the differences in results between Experiment One and the present experiment. As was described in Section 4.2.2, the octave bands used in the present experiment were created using an 8th order Butterworth filter. Conversely, for Experiment One an FFT filter was used. The effect that each filter had on the resultant spectrum of the 1 kHz octave band is shown in Fig. 4.18. It can be seen that the predominant difference was a much more gradual roll-off for the Butterworth filtered stimuli, which would have meant that more frequencies outside the pass band would have been audible in the signal. It might be the case then that the presence of frequencies outside the pass band might have had an effect on the frequency dependent differences in VIS between the main layer

only and phantom image conditions. This in turn would have lessened the effect of frequency on the localisation thresholds obtained, which matches what was observed in the experiment.

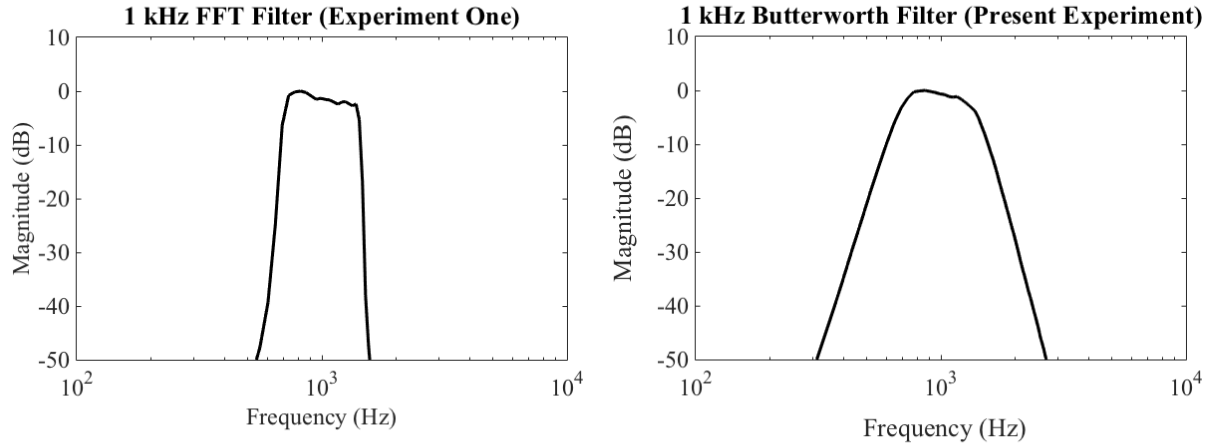

Fig. 4.18: FFTs for 1 kHz octave band filtered using Butterworth filter (right) and FFT filter (left).

In addition, the threshold detection methods used in the respective experiments might be the reason that the thresholds obtained for the present experiment were lower than those for Experiment One. The effect that the methodology has on the subsequent thresholds obtained has been examined in the literature. Hesse [1986], for example, reported a 'significant' decrease in thresholds of around 2 dB for the two alternative forced choice method compared to tracking, adjustment and yes/no methods when investigating the threshold of audibility of a sine wave in noise. Furthermore, Bech [1998] analysed the threshold of detection of reflections, reporting that the use of MOA resulted in thresholds around 3-5 dB lower than two-alternative forced choice. The 1-3 dB difference between the results of the present experiment and those in Experiment One falls within the range of variation reported in the aforementioned studies. Therefore, it can be suggested that the differing threshold detection methods used for each experiment (MOA for Experiment One, AMOA for the present experiment) contributed to the variation in localisation thresholds. This would require further study, as the AMOA method is novel and the thresholds obtained using it have not been directly compared to

other, pre-existing, methods. Additionally, the effect of threshold detection method has not been explored in relation to localisation thresholds.

4.2.4.2 The Effect of Signal Duration

From the data, it can be seen that the frequency dependency of localisation thresholds had a notable dependency on the duration of the signal. Indeed, for the burst stimuli localisation thresholds were generally consistent across the spectrum, whilst the pattern for the continuous stimuli was somewhat similar to that reported in Experiment One. In addition, the continuous stimuli generally yielded lower localisation thresholds than did the burst stimuli. In a bid to ascertain the reasons for this result, the author conducted a series of informal listening exercises. Given the aforementioned hypothesis that the frequency dependency of localisation thresholds might be related to frequency-dependent differences in perceived VIS between the main layer only and phantom image conditions, the listening exercises primarily focused on how the perception of VIS varied for each signal duration. The primary observation made is as follows. When a sound source is presented as a vertically oriented phantom image, the full extent of the source's VIS is not immediately apparent. Instead, the vertical spread of the source increases to a maximum over the course of a few seconds.

Based on the above observation, the following can be hypothesised. For the continuous stimuli, the signal duration was such there was time for the full extent of VIS to build up. This enabled to subjects to identify the frequency-dependent differences in perceived VIS between the main layer only and phantom image conditions more clearly, which resulted in frequency-dependent localisation thresholds. On the other hand, for the burst conditions the signal duration (200 ms) was such that the full extent of VIS did not have sufficient time to build up. Consequently, it was more difficult to perceive any frequency-dependent differences in VIS and, as a result, the thresholds were more consistent across the spectrum. That burst stimuli do not enable a full build up of VIS might also explain why the thresholds were lower for the

continuous stimuli than they were for the bursts. In this regard, it could be that the differences in perceived VIS for the main layer only and phantom image conditions would be smaller for the burst stimuli than they would be for the continuous stimuli, which would mean less ICLD would be necessary in order to meet the localisation threshold.

It should be noted, however, that the build up of VIS over time has not been demonstrated experimentally. The above hypothesis is therefore based on the informal listening experiences of the author alone. Subsequently, this phenomenon would require further study to determine both if it is prevalent for all subjects and further, if it does operate, what the minimum necessary time is for the full extent of VIS to be built up. In addition to this, the hypothesis does not explain why the effect of frequency was still significant for the burst stimuli when the ICTD was 0 ms. Further, the reason why signal duration was not significant for all octave bands tested and, even when it was, why there was a dependency on ICTD, is not known. It is clear then that further study is necessary with respect to this hypothesis.

4.2.4.3 The Effect of Presentation Method

A further aim of the present experiment was to analyse how the localisation thresholds would be affected by different presentation methods. The experimental data obtained shows that although there were some differences in the median localisation threshold when presentation changed from vertical stereophonic to vertical quadraphonic, these differences were not significant. With respect to the octave band stimuli, this result can arguably be explained based on the aforementioned VIS hypothesis. In this case, it would appear that the frequency-dependent differences in perceived VIS when presentation changed from main layer only to vertical phantom image did not differ significantly between stereophonic and quadraphonic presentation. This would have to be studied further, however, as frequency-dependent increases in VIS between main layer only and phantom image presentation have not yet been demonstrated experimentally.

However, the above hypothesis is arguably insufficient at explaining why presentation method did not significantly affect the localisation threshold for broadband pink noise. Firstly, it is known from Experiment Two that there are increases in perceived elevation for this source when presentation changes from main layer only to vertical phantom image, which is not the case for the octave band stimuli. In addition, presentation using the quadraphonic condition is likely to be affected by the phantom image elevation effect, which would not be a factor for vertical stereophonic presentation. In order to explain this result, consideration was given to the results of Experiment Three, which showed that the effect of presentation method was not significant for natural sound sources. In Section 4.1.4.2, it was hypothesised that this result was obtained as a result of the following mechanism. Firstly, as was discussed in Section 3.1.4.1, the localisation threshold for complex sources might be determined by the difference in energy in the 7-9 kHz region between the main layer only and phantom image conditions; when this energy difference is sufficiently small the resultant stimuli are perceived as being in the same location. Following Experiment Three then, it was demonstrated that this difference in energy undergoes a similar reduction for each presentation method when the localisation threshold is applied; this was used to explain why the effect of presentation method was not significant. Therefore, given that the balance of spectral energy in the 7-9 kHz region has previously been used to explain the localisation thresholds for broadband pink noise, it seems reasonable to suggest that this is the reason that the effect of presentation method was not significant for this source.

4.2.4.4 The Effect of ICTD

The effect of ICTD was found to be significant for a limited number of stimuli tested in the present experiment. For the continuous stimuli, significant differences were identified between 0 and 10 ms for both the 8 kHz octave band and for the broadband pink noise. In addition, for the burst stimuli the threshold for the 4 kHz octave band with 0 ms ICTD was significantly lower than those for both 1 and 10 ms, whilst for the pink noise source the threshold for 1 ms was significantly lower than for 10 ms. That less level reduction

is necessary in the presence of a delay in the height layer could be indicative of a localisation dominance effect, which is discussed in detail in Chapter Two. However, it should be noted that a number of the results obtained in the present study disagree with this suggestion. For the pink noise bursts, for example, there was no significant difference between the results for 0 and 10 ms, with the median thresholds being almost identical. In addition to this, the effect of ICTD was only significant for a limited number of stimuli; had the effect operated than arguably a more consistent effect of ICTD would have been observed. It could be the case then that the significant effect of ICTD as observed in the present study was due to the random effects of comb filtering. It has already been discussed, following Experiment Two, that this might affect the perceived location of the broadband pink noise source and perhaps there may have been some effect with respect to the localisation thresholds obtained. It should be noted however that this would have to be studied further. It also remains unclear how this would affect the thresholds obtained for the band-limited stimuli.

4.2.4.5 Practical Implications

The results of the present experiment are valuable in providing a number of details with respect to the nature of a band reduction method. Firstly, and perhaps most importantly, that the frequency dependency of localisation thresholds is maintained in the presence of reflections shows that there is a suitable experimental basis from which a band reduction method can be developed. Further to this point, the results also give exact values as to how much each octave band in the height layer should be attenuated in order to reach the localisation threshold. It is also interesting to consider the discussions regarding the mechanism for determining the localisation threshold for the broadband pink noise. If it is the case that the elevation effects of vertical interchannel crosstalk are related to specific frequency regions then manipulating said regions alone could result in the localisation threshold being met without the rest of the direct sound in the height layer being attenuated. This is considered further in Experiment Five.

A further point of note relates to the effect of signal duration. Primarily, this result indicates that band reduction techniques should consider each instrument individually based on its blend of continuous and transient characteristics. For example, a more continuous source, such as a violin, would require a greater amount of level reduction in the high frequency range than would something more transient, such as percussion. It should be noted that this would be difficult to apply in a practical recording situation, however it is certainly applicable to 3D image rendering techniques. Additionally, the results suggest that band reduction for transient sources might not even be entirely necessary, as the thresholds for each band were generally consistent across the spectrum when burst sources were used as the test stimuli. Furthermore, although generally minimal, there would also have to be some consideration as to the effect of ICTD, with the present experimental data showing that its effect was significant at 4 kHz for the burst stimuli and at 8 kHz for the continuous stimuli. This therefore indicates that consideration should be given as to how delayed the direct sounds are in the height layer with respect to those in the main and the effect that this will have on how the aforementioned bands will be treated.

4.2.5 CONCLUSION

The present experiment built upon Experiment One by analysing the frequency dependency of localisation thresholds in the presence of reflections and therefore providing the experimental basis for the development of a band reduction method. Broadband pink noise and nine octave bands, with centre frequencies ranging from 63 Hz - 16 kHz, were presented to subjects as vertically arranged phantom images using both stereophonic and quadraphonic conditions. The stimuli were presented both in bursts and continuously. The height layer of loudspeakers was delayed with respect to the main by 0, 1 and 10 ms. Subjects were required to reduce the amplitude of the height layer until the resultant phantom image position matched that of the same stimulus presented from the main layer alone using an AMOA method.

The experimental data obtained showed that the frequency dependency of localisation thresholds was maintained for the continuous stimuli with 0 and 1 ms ICTD. However, it should be noted that the effect of frequency was generally not as strong as was reported in Experiment One. This result was first interpreted based on comb filtering effects caused by the first arriving reflection, which might have affected the frequency dependent differences in VIS between the main layer only and phantom image conditions that are thought to be important with respect to the frequency dependency of localisation thresholds. Additionally, differences in the test method between the two experiments, in terms of the use of different filters to create the test stimuli, as well as different threshold detection methods used for the listening tests, might have been important.

Stimulus duration was found to have a significant effect for at least one ICTD for each of the frequency bands in the range of 500 Hz - 16 kHz, with the exception of 4 kHz. The effect was also significant for the broadband pink noise source. In each case, the threshold was lower for the continuous stimuli. Also, the effect of frequency was less strong for the bursts, with the median thresholds generally being consistent across the spectrum. Informal listening conducted by the author identified that the build up of VIS is not instantaneous. Instead, a few seconds are necessary for the full extent of VIS to be realised. It was reasoned then that the duration of the burst stimuli was too short to enable such a build up, which might have affected the frequency-dependent differences in perceived VIS necessary for frequency dependent thresholds to be identified. This might also explain why the thresholds for the burst stimuli were higher than those for the continuous stimuli.

The localisation thresholds obtained in the experiment were not significantly affected by the presentation method. For the band-limited stimuli, it was thought that this was because both methods resulted in similar differences in perceived VIS between the main layer only and phantom image conditions. With respect to the broadband pink noise source, this result was explained using the hypothesis that the reduction in the difference in energy between the main layer only and phantom image conditions in the 7-9 kHz range was similar for both presentation methods for a given ICLD. This was the same hypothesis as was used to explain

the non-significant effect of presentation method for the natural sound sources, as observed in Experiment Three.

ICTD was found to have had a limited effect on the localisation thresholds. Indeed, the only significant differences were identified for the 8 kHz continuous stimuli, the 4 kHz bursts and for both pink noise sources. The pattern followed by these results was somewhat indicative of a localisation dominance effect, whereby less level reduction is necessary in the presence of an ICTD. Despite this, inconsistencies with the thresholds for the pink noise source, as well as an overall inconsistent effect of ICTD, indicated that this result might be related to comb-filtering effects.

Overall these results provide the impetus for the development of a band reduction method, which is discussed in detail in Experiment Five.

4.3 EXPERIMENT FIVE: DEVELOPMENT OF A BAND REDUCTION METHOD AND THE VERIFICATION OF LOCALISATION THRESHOLDS

In Experiments Three and Four, localisation thresholds were obtained for both the band and blanket reduction methods. The purposes of the present experiment are twofold. The first is to derive a series of different band reduction methods based on the experimental data provided from Experiment Four. To reiterate, this would involve the frequency-dependent manipulation of the amplitude of the direct sound in the height layer. The second is to verify that such methods, alongside the blanket reduction methods derived in Experiment Three, are effective at preventing the height layer from affecting the perceived location of the main channel signal.

Verifying the effectiveness of the localisation thresholds is considered as being important in light of the overall aims of the research. One of the key aims is to analyse how the method of applying the localisation threshold influences the perceived timbre and spaciousness of the main channel signal. Further, it is also desired to determine which method is the most preferred by subjects. In order that these aims can be realised, it is essential to ensure that the stimuli are being presented to subjects in such a way that the perceived location of the main channel signal is not being affected by the signal in the height layer. Hence, verification tests are necessary.

From the above background the following research questions were derived:

- Can a series of band reduction methods be developed that are successful in the application of localisation thresholds?
- Can the blanket reduction thresholds obtained in Experiment Four be verified?

4.3.1 EXPERIMENTAL DESIGN

4.3.1.1 Physical Setup

The physical setup for the experiment is shown in Fig 4.19. The experiment was conducted in the same room as was used in Experiments Three and Four and used an almost identical setup. However, as both experiments demonstrated that the localisation thresholds were not affected by presentation method, only the L, R, HL and HR loudspeakers were used (i.e. the vertical quadraphonic condition); the C and HC loudspeakers were removed. The vertical quadraphonic condition was favoured to the vertical stereophonic condition as existing 3D audio systems, such as Auro 3D [Auro Technologies 2016], tend to make use of elevated L and R loudspeakers, however do not always use an elevated centre loudspeaker. It was therefore considered that the vertical quadraphonic condition would be more relevant to practical situations. A light-

emitting diode (LED) strip was positioned directly in front of the listening position. This was located behind the acoustically transparent curtain and was to be used by subjects when making their localisation judgments.

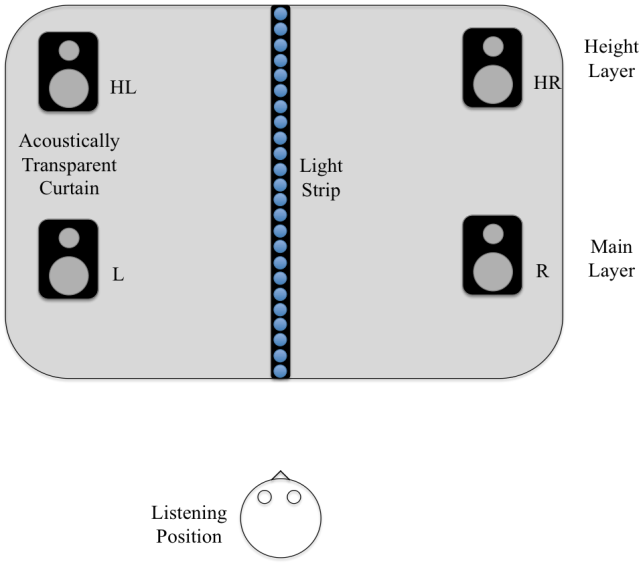

Fig. 4.19: Physical setup for the localisation threshold verification test.

4.3.1.2 Test Stimuli

For the present experiment, the goal was not simply to analyse band reduction by conducting frequency-dependent attenuation of the direct sound in the height layer across the full frequency spectrum. Indeed, the data obtained in the experiments conducted thus far indicates that there may be other ways to apply localisation thresholds using the band reduction technique. For example, it has been discussed throughout the thesis that there may be some perceptual dominance of the 7-9 kHz region in determining the localisation threshold for complex sources. In addition, the results presented in Experiments One and Four generally indicated that the thresholds for continuous octave bands are higher for the low frequencies (125-500 Hz) than they are for the mid-high frequencies (1 kHz and above). Furthermore, in the literature it has been reported that the primary cues for elevation perception lie in the 4-10 kHz range [Shaw and Teranishi 1968,

Hebrank and Wright 1974a, Asano et al. 1990]. More specifically, 8 kHz is related to 'above' perception [Blauert 1969], whilst 4 kHz is associated with the elevated front [Appendix A].

If it is the case that certain frequency regions are dominant with respect to the elevation effects of vertical interchannel crosstalk, it might be possible to apply the localisation threshold by manipulating either single bands or groups of bands within the direct sound signal in the height layer. This seems particularly salient given the results of a study conducted by Chun et al. [2011], who boosted the amplitude of musical sources, which were presented from horizontally arranged loudspeakers, in the 4-9.5 kHz range. As a consequence of this 'directional band boosting', the resultant sources were perceived as being elevated by up to 20° with respect to the horizontal plane. It could be suggested then that the opposite effect (i.e. directional band attenuation) could prevent vertical interchannel crosstalk from affecting the perceived location of the main channel signal.

From the above discussions, the following band reduction methods were considered. In each case the attenuation was only applied to the direct sound in the height layer:

1. Full band (FB): Every octave band in the signal underwent frequency-dependent attenuation.
2. 1 kHz and above (1+B): Frequency-dependent attenuation was applied to the 1 kHz band and above. The frequencies below 1 kHz were not attenuated.
3. 4 kHz reduction only (4B).
4. 8 kHz reduction only (8B).

The stimuli used for the experiment were the same natural sound sources as were used in Experiment Three. The test stimuli were presented to subjects in the following conditions: (i) main layer only; (ii) height layer only; (iii) vertically oriented phantom image with 0 dB ICLD and; (iv-viii) vertically oriented phantom image with a localisation threshold applied to the height layer (blanket, FB, 1+B, 4B and 8B). Although not necessarily integral to the verification test, the 0 dB ICLD and height layer only conditions were included in

the experiment in order to reduce any expectation biases. During preliminary tests, in which only the main layer only and localisation threshold conditions were considered, subjects reported that hearing all stimuli originate from similar positions in a localisation test was confusing. Furthermore, some were led to believe that the stimuli couldn't all be coming from the same location and this forced them to provide different answers to what was actually being perceived. The height layer only and 0 dB conditions were therefore included in order to introduce stimuli that were in a position away from the main layer only condition. This was found to prevent the issue.

The ICTDs used for the experiment were 0 and 1 ms. In general, when a significant effect of ICTD was identified in either Experiment Three or Four, the significant difference tended to be between 0 ms and either 1 or 10 ms. For example, for Experiment Three the blanket reduction thresholds were -9.5 dB for 0 ms, which was significantly lower than the -7 dB for both 1 and 10 ms. Equally, in Experiment Four significant differences were identified for the 4 kHz bursts between 0 ms and both 1 and 10 ms, whilst the 8 kHz continuous stimuli showed a significant effect between 0 and 10 ms. This therefore suggests that different thresholds should be applied in the case that an ICTD is present compared to when one is not. It was however decided that both 1 and 10 ms did not need to be tested, as there was consistently no significant difference between the thresholds for these ICTDs for either Experiment Three or Four. As was discussed earlier, in the context of microphone techniques for recording for 3D audio formats, a 10 ms ICTD corresponds to a 3.4 m path difference between the direct sound arriving at each layer, which is fairly large in practice. It was therefore decided that the 1 ms condition would be more representative of a practical configuration, with the path difference being only around 0.34 m.

The threshold values used for the blanket and band reduction methods were as follows. For band reduction, the thresholds applied to the speech, oboe, guitar and quartet sources were the median thresholds obtained for the continuous stimuli in Experiment Four. This choice was made due to the mixture of transient and continuous characteristics in continuous noise [Hartmann 1983], arguably making them more applicable for use on the aforementioned natural sound sources. However, due to its transient nature, band reduction for the

conga used the thresholds obtained for the burst stimuli. A summary of the thresholds used for each band is shown in Table 4.1. It should be noted that, due to the significant effect of ICTD identified in Experiment Four, the thresholds for both the continuous 8 kHz band and the 4 kHz bursts were dependent on the delay applied to the height layer. For blanket reduction, the median thresholds derived in Experiment Three were used were used (i.e. -9.5 dB for 0 ms and -7 dB for 1 ms).

Table 4.1: Band reduction values for continuous and burst octave bands.

Centre Frequency	**Attenuation (dB)**	
	Continuous	**Bursts**
63 Hz	-6	-6
125 Hz	-7	-6
250 Hz	-7	-5
500 Hz	-8.5	-6
1 kHz	-11	-6
2 kHz	-10	-6
4 kHz (0 ms)	-9.5	-11
4 kHz (1 ms)	-9.5	-6
8 kHz (0 ms)	-13.5	-6
8 kHz (1 ms)	-10	-6
16 kHz	-8	-6

The test stimuli were created as follows. For the band reduction method, each source was first broken down into octave bands using the same Butterworth filters that were used for Experiment Four. Each octave band then underwent the relevant amplitude reduction. Following this, the bands were combined in Logic Pro X and were bounced as the right channel of a stereo wav file. The difference in the waveform of the guitar and speech sources as a result of filtering the signals into octave bands and re-combining them can be seen in Fig. 4.20. Here it is noticeable that the re-combined signals are generally similar to the original signals albeit there are notches in the signal of around 2 dB at the boundary frequencies for each octave band that is identical for both sources. Informal listening conducted by the author identified that the audibility of the difference between the waveforms was low for each source, with no other audible artifacts being present as a result of filtering and re-combining each signal. The left channel of the stereo file was the unaltered

(original) source. Delays of 0 and 1 ms were applied to the right channel with respect to the left. During the test, the left channel was routed to the main layer (L and R), whilst the right was routed to the height (HL and HR). For blanket reduction the process was identical, except that the sources were not broken down into octave bands; the amplitude of the right channel was reduced as a whole.

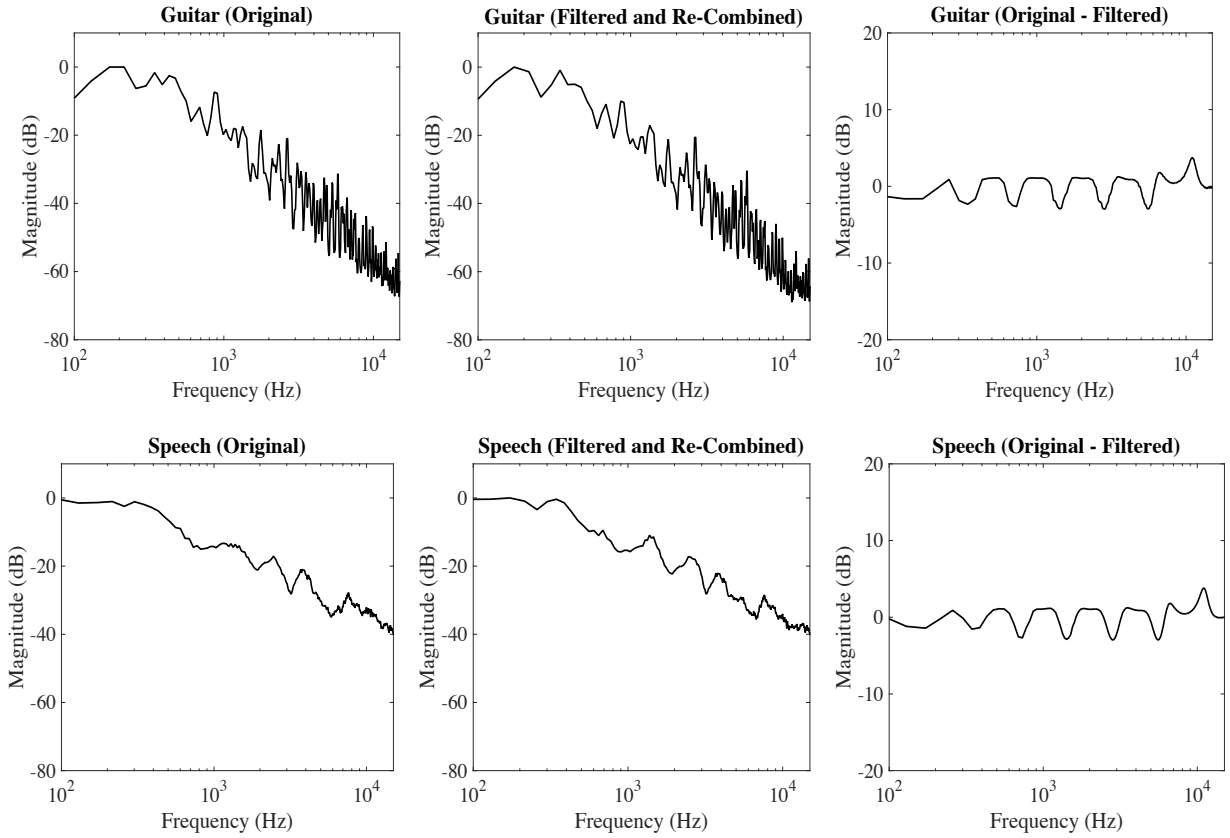

Fig. 4.20: waveforms for guitar (top) and speech (bottom) sources showing i) original frequency content (left), ii) frequency content after source has been broken down into octave bands using an 8th order Butterworth filter and re-combined (centre) and iii) difference between original and recombined signals (right).

All stimuli were presented at 70 dB LAeq at the listening position when presented from the main loudspeaker layer only. The increase in amplitude when the stimuli were presented as vertically arranged quadraphonic phantom images was dependent on the localisation threshold applied to the height layer, as shown in Table 4.2. It should be noted that it was decided against level matching the stimuli, as amplitude

increases as a result of vertical interchannel crosstalk are inevitable. As a result, not level matching was considered as being more representative of a practical situation in which the effect is present. In total, there were 70 stimuli, being the main and height layer only conditions (10), the localisation threshold conditions (50 – five sources, two ICTDs, five threshold methods) and the 0 dB ICLD conditions (10 – five sources, two ICTDs).

Table 4.2: Average amplitudes for stimuli presented using each localisation threshold method.

Method	Amplitude (dB LAeq)
FB	70.0
1+B	71.2
4B	72.2
8B	72.4
Blanket	70.4
0 dB	73.1

4.3.1.3 Test Method

The test was completed by the same 10 subjects who participated in Experiments Three and Four. Localisation judgments were made using the LED strip located directly in front of the listening position. For each test, subjects were provided with a handheld knob, which controlled which LED on the strip was turned on. They were required to adjust the knob until the position of the active LED matched the perceived location of the focal point of each stimulus. This method was chosen following research conducted by Lee et al. [2016], who found that it was faster and produced results with greater accuracy and consistency compared to the numbered scale method, which had been used previously for Experiment Two. The position of the LED selected for each stimulus was converted into an elevation angle within a Max/MSP patch. The LEDs on the strip were spaced apart by 3 cm, which gave a resolution of 0.779° at a distance of 2 m from the listening position. The heads of subjects were not fixed, however they were instructed to sit up and face forwards at all times, using only their eyes to look at the light strip. To help maintain the correct seating

position, a small headrest was positioned behind the head of each subject. The test was completed four times by each subject, with each sitting containing all 70 stimuli and taking around 20 minutes to complete. The presentation order of stimuli was randomised for each test.

4.3.2 DATA ANALYSIS AND RESULTS

Levene and Shapiro-Wilk tests were first conducted, using the SPSS software, in order to determine the suitability of the collected data for parametric statistical analysis. The Shapiro-Wilk test showed that not all scores in each condition featured normal distribution, although the results of the Levene test showed homogeneity of variance for all sound sources. For these reasons, non-parametric tests were chosen for the statistical analysis.

Fig. 4.21 shows the median perceived elevation of each of the test stimuli, plotted with notch edges. With respect to the localisation thresholds, it can be observed that each of the proposed band reduction methods, as well as the blanket thresholds obtained from Experiment Three, resulted in the perceived location of each stimulus being similar to that for the main layer only condition. This was the case for all sources, with the median difference in perceived elevation between the main layer only and localisation threshold conditions ranging between 2.4° and -4.0° for the 0 ms ICTD and between 3.9° and -3.2° for the 1 ms ICTD. Alongside this, the notch edges between the main layer only and localisation threshold conditions all overlap. It is also interesting to note that for the 0 ms ICTD, the median perceived elevation for the stimuli with the localisation threshold applied was slightly lower than that for the main layer only condition for a large number of sources.

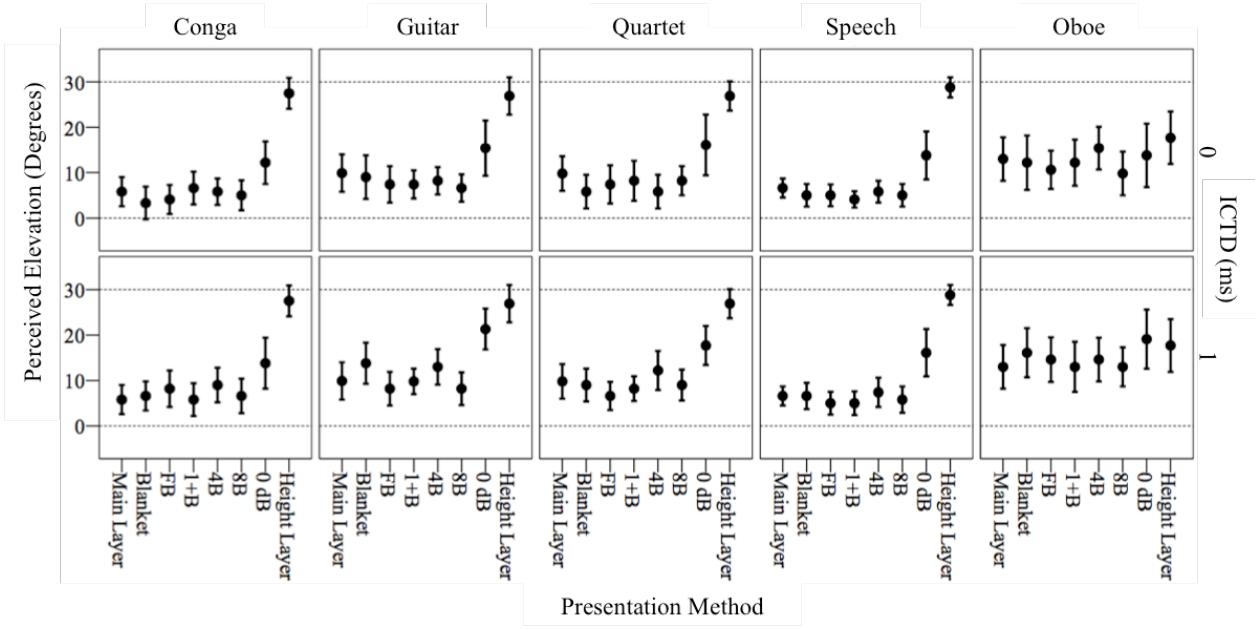

Fig 4.21: Median perceived elevation for each stimulus in the verification experiment. The dotted lines at 0° and 30° represent the positions of the main and height layers respectively.

In order to further determine whether or not each of the localisation thresholds were successful in preventing vertical interchannel crosstalk from affecting the perceived location of the main channel signal, Wilcoxon tests were conducted (p = 0.05). The results suggested the following. Firstly, for both the conga and oboe sources, there were no significant differences between the judgments for each of the localisation threshold conditions and those for the main layer only condition. However, significant differences were identified for the other sources as follows; guitar – 8B (0 ms, p = 0.030), blanket (1 ms, p = 0.002); quartet – blanket (0 ms, p = 0.041), 4B (1 ms, p = 0.06); speech – 8B (0 ms, p = 0.033), blanket (1 ms, p = 0 .025). However, in each of these cases it can be seen that there is overlap between the main layer condition and each of the notch edges. In addition, the effect size r did not indicate a large effect in any case (< 0.5 for all). As a result, it can be concluded that there were no significant differences between the localisation threshold and main layer only conditions for any of the sources tested, with the significance identified in the Wilcoxon tests likely being type-I errors. Therefore, each of the proposed methods was successful in preventing vertical interchannel crosstalk from affecting the perceived location of the main channel signal.

Furthermore, it is interesting to note that median judgments for the vertically oriented phantom image conditions with 0 dB ICLD were generally higher than those for the main layer only condition. The only source that did not follow this trend was the oboe, whose perceived elevation was similar for all conditions. This indicates that this source was less affected by the migration of the main channel signal from the main layer as a result of vertical interchannel crosstalk. This result might suggest that the application of localisation thresholds would not always be necessary and would have somewhat of a source dependency. Excluding the oboe, the smallest increase in median perceived elevation for the 0 dB condition compared to the main layer only condition was 5.5° (0 ms guitar), whilst the largest increase was 11.4° (1 ms guitar). Additionally, as a whole the 0 ms condition yielded a smaller difference in median perceived elevation between the 0 dB and main layer conditions compared to the 1 ms condition. Wilcoxon tests showed that the difference between the 0 dB and main layer only conditions was significant for all sources, with the exception of the oboe at 0 ms (p = 0.823). This result generally agrees with the notch edges in Fig. 4.21, although it is clear that in some cases, such as for the quartet at 0 ms, there is some overlap. Nevertheless, it is apparent that there was a notable, and generally significant, increase in median perceived elevation when the stimuli were presented as vertically oriented phantom images, with 0 dB ICLD. Further, this difference was markedly reduced when each of the localisation threshold methods were applied to the height layer.

A further result of note can be seen with respect to the main and height layer only conditions. Firstly, for the latter condition it would appear that perceived elevation judgments were generally accurate for all sources, excluding the oboe, with respect to the physical position of the height layer. Conversely, for the main layer only condition the judgments were less accurate, with perceived source elevation being in the range of 5.8°-13.0° with respect to the main layer's physical position. This elevation of the sound source with respect to the main layer was also maintained for the conditions whereby a localisation threshold was applied to the height layer. The results of a Wilcoxon signed rank test, which compared the results for the main layer only condition to the physical position of the main layer (0°), showed that each source was perceived to be significantly higher than the physical height from which the source was presented (p = 0.000 for all sources).

4.3.3 DISCUSSION

4.3.3.1 The Effectiveness of the Localisation Threshold Methods

The data provided in the present experiment has shown that a shift in the perceived elevation of natural sound sources is apparent when a sufficient amount of direct sound is present in the height layer. In addition, using any of the band or blanket reduction methods that were tested can prevent this from happening. That both the blanket reduction and FB methods were successful was somewhat expected. The blanket reduction thresholds, for example, had already been derived in Experiment Three and so were reasonably expected to work. Additionally, it can be deduced from Table 4.1 that the influence of the height channel on the resultant amplitude of each source for the FB method was low, with the average peak amplitude being equal to that for the main layer only condition. Therefore, as the audibility of the direct sound in the height layer was low with respect to that in the main layer, it seems reasonable that the perceived location of the main channel signal would not be affected. It was considered as being more interesting that localisation thresholds could be applied through the selective manipulation of frequency bands within the height layer, with the 1+B, 4B and 8B conditions all being effective.

In order to explain the effectiveness of the 1+B, 4B and 8B methods, consideration was first given to the discussions regarding the mechanisms that might determine the localisation threshold for complex sources. As was discussed in Chapter Three, a vertically oriented phantom image will have more energy in the 7-9 kHz region compared to main layer only presentation. Given that 8 kHz has an association with above localisation [Blauert 1969], it is reasonable to conclude that sufficient energy difference in the 7-9 kHz region will result in the phantom image condition being elevated with respect to main layer only presentation. By extension, this would also mean that providing sufficient attenuation of the phantom image in this range would result in the localisation threshold being met. Based on this hypothesis, it is relatively simple to explain why the 8B and 1+B methods were effective. In either case, the methods necessitated the attenuation

of the 8 kHz octave band by a minimum of 6 dB (depending on the transient nature of the source and the ICTD). It can therefore be suggested that this attenuation was sufficient for the phantom image condition to not sound elevated with respect to the main layer only condition and hence the 1+B and 8B methods were effective at preventing vertical interchannel crosstalk from affecting the perceived location of the main channel signal.

However, it should be noted that the above discussions do not explain why the 4B method was effective. For this method, only the 4 kHz octave band was attenuated, which would have meant that the energy differences in the 7-9 kHz range between the main layer only and phantom image conditions would have been maintained. This result would seemingly indicate that energy differences in the 7-9 kHz region are not as dominant in determining the localisation threshold for natural sound sources as had been previously suggested. In order to gain further objective insights into this result, the difference in spectral energy between the height and main layers for the vertical quadraphonic condition were considered using HRIRs obtained from the KEMAR dummy head database [Gardner and Martin 2000]. The resultant measurements are shown in Fig. 4.22, in the range of frequencies covered by the 4 and 8 kHz octave bands (2840 – 11360 Hz). Any point where the spectrum falls below 0 dB represents dominance of the main layer over the height and vice versa. From Fig. 4.22, it can be seen that the difference in spectral energy between the main and height layers is much smaller for the 4 kHz octave band (2840-5680 Hz), with differences being in the range of ± 5 dB, compared to the 8 kHz octave band (5680-11360 Hz, ± 20 dB). Although it is known from the literature that the spectral cues for elevation exist as low as 4 kHz [Hebrank and Wright 1974a, Asano et al. 1990], it can be argued that the spectral energy differences in the region governed by the 4 kHz octave band are not sufficient to contribute to differences in perceived elevation, with the peak between 7 and 9 kHz likely being more important. Based on this analysis, it is perhaps the case that the balance of spectral energy in the 7-9 kHz region is not the primary mechanism for the determination of the localisation threshold for natural sound sources. This, however, would require further study.

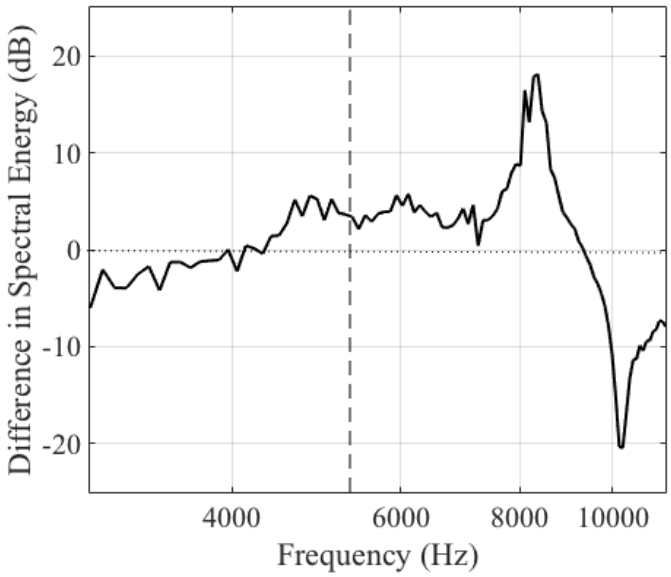

Fig. 4.22: Difference in spectral energy between the main and height layers of a vertical quadraphonic configuration. The vertical dashed line separates the frequency range for the 4 kHz octave band (left) from that for the 8 kHz octave band (right).

4.3.3.2 The Relationship Between Perceived Source Elevation and the Localisation Threshold

The results obtained in Experiment Three indicated that the blanket reduction localisation thresholds were not source dependent. This is interesting when the results of the present experiment are considered, which showed that the difference in perceived elevation between the main layer only and phantom image conditions differed for each source. For example, the speech source presented as a vertically oriented phantom image with 0 dB ICLD was perceived as being significantly higher than that for main layer only presentation for both of the ICTDs tested. On the other hand, judgments for the oboe source were generally consistent irrelevant of how the source was presented to subjects. The remaining sources were affected more inconsistently. The increases in median perceived elevation for the conga, guitar and quartet sources were significant when the ICTD was 1 ms. Despite this, for the 0 ms ICTD the difference was not significant, even though the median perceived elevation notably increased in each case. These results would seemingly

indicate that vertical interchannel crosstalk has a source dependent effect with respect to the perceived migration of the main channel signal from the position of the main layer.

With respect to the oboe source, an explanation of the results obtained is offered thus. According to the literature, narrowband stimuli incident from the median plane are localised on the basis of frequency, with increases in frequency corresponding to increases in perceived elevation [Roffler and Butler 1968a, Cabrera and Tiley 2003]. This phenomenon is known as the 'pitch-height effect' [Cabrera and Tiley 2003]. As can be seen from Fig. 4.2, the spectrum for the oboe source was notably narrow, with a bandwidth ranging from around 500 Hz to 4 kHz and with its predominant energy focused around 1 to 2 kHz. According both to the literature and to the results presented in the present thesis, band-limited stimuli in this frequency range are localised at a similar vertical position, regardless of which loudspeaker layer presented the source, for both vertical stereophonic [Experiment Two, Cabrera and Tiley 2003] and vertical quadraphonic [Lee 2016] loudspeaker arrangements. Therefore, it might be that localisation judgments for the oboe were determined by the pitch-height effect, with no relation to the difference in energy in the 7-9 kHz region between the phantom image and main layer only conditions. Should this be the case, then it would indicate that that hypothesis is predominantly applicable to broadband sources, with less relevance to sources that are both narrowband and absent in high frequency energy.

From the above discussions, there are two important points to consider with respect to the results of Experiment Three. Firstly, for the oboe source ICLD was always required to reach the localisation threshold, despite there being little difference in perceived elevation between different conditions. Secondly, there was not a significant effect of sound source, even though the results of Experiment Five suggested that vertical interchannel crosstalk had a source dependent effect on the perceived elevation of the main channel signal. Combined, these results indicate that there is perhaps not necessarily a connection between the difference in perceived elevation between the main layer only and 0 dB phantom image conditions and the subsequent localisation threshold obtained, which is similar for what was observed for octave band stimuli in Chapter Three. In that chapter, the results were explained in the following way. When sound source presentation

shifts from main layer only to vertical phantom image, a key difference is an increase in perceived VIS. Therefore, given that the test conditions required a direct comparison between the positions of stimuli presented using the main layer only and vertical phantom image conditions, it is possible that differences in perceived VIS were perceived as elevation differences. Further, these differences would have decreased as the amplitude of the height layer was reduced. At the localisation threshold then, the difference in VIS is sufficiently small for stimuli presented using the two conditions to be perceived as being in the same location. Arguably, the results of the present study show that this mechanism can also explain the localisation thresholds obtained for natural sound sources in Experiment Three, with the differences in perceived VIS between the main layer only and phantom image conditions not being source dependent. This would have to be studied further, however, as reductions in the amplitude of the height layer would simultaneously decrease the differences in perceived VIS and the energy difference in the 7-9 kHz region between the phantom image and main layer only conditions. Further study would therefore be required to ascertain which mechanism is the more dominant for determining the localisation threshold for natural sound sources.

4.3.3.3 The Localisation Dominance Effect

In Experiment Three, it was identified that the blanket reduction threshold was significantly higher in the case that the height layer was delayed with respect to the main. It was subsequently hypothesised that this was due to the presence of a localisation dominance effect, in which perceived source location is biased towards the position of the earlier loudspeaker layer. The data provided in the present experiment enables further analysis as to whether or not such an effect exists. Fig. 4.23 shows the experimental data for the main and height layer only conditions alongside those for the 0 dB ICLD conditions (both 0 and 1 ms ICTD). The median perceived elevation for each has been plotted with notch edges. From the results, it is clear that there is no evidence to support the existence of a localisation dominance effect, with the median perceived elevation for all stimuli increasing in the presence of an ICTD. This result is somewhat similar to those reported in Experiment Two, in which the perceived elevation of broadband pink noise presented from

vertically arranged stereophonic loudspeakers in anechoic conditions increased as the ICTD increased from 0-1 ms. With respect to the results of Experiment Three, the hypothesis that the localisation thresholds are higher in the presence of an ICTD due to the operation of a localisation dominance effect can be rejected. As a consequence of this, further study would be required in order to adequately explain this result.

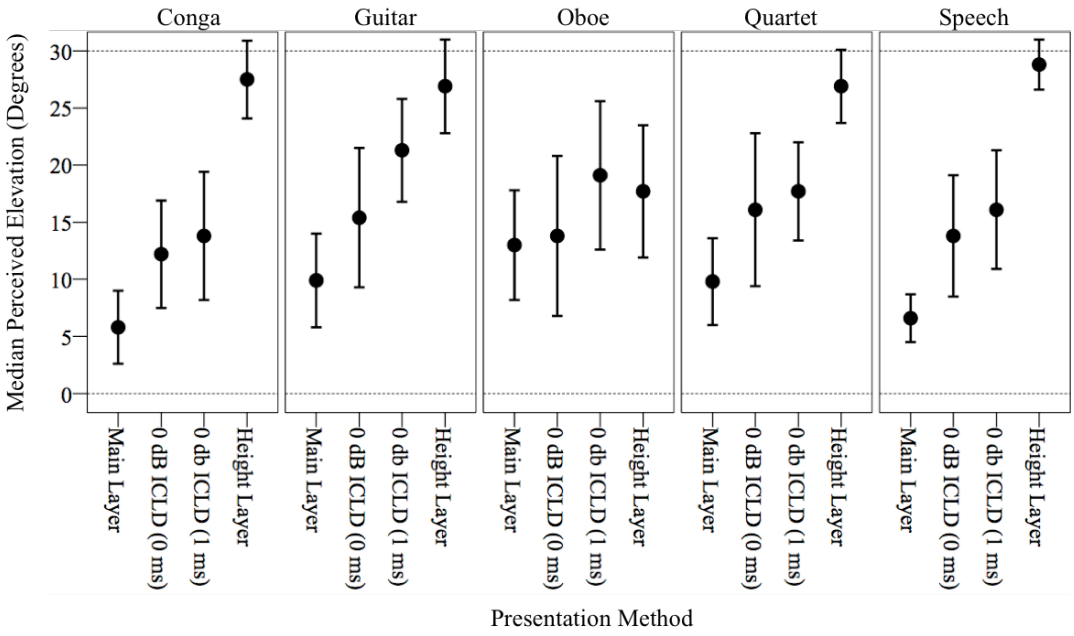

Fig. 4.23: Analysis of the localisation dominance effect.

4.3.3.4 Relationships with Vertical Amplitude Panning and the Phantom Image Elevation Effect

It is possible to compare the results of the present experiment to those reported by Barbour [2003], who conducted median plane stereophonic localisation studies using pink noise and speech sources. The test stimuli were presented to subjects as phantom images from vertically arranged stereophonic loudspeakers. The lower (main) loudspeaker was not elevated with respect to the listening position, whilst that of the upper

(height) loudspeaker varied between 45°, 60° and 90°. Subjects were required to identify the perceived elevation of each stimulus when the ICLD varied between -15 dB (height layer attenuated) and 15 dB (main layer attenuated) in 3 dB steps. Before the results of the respective studies are compared, it is first necessary to identify a series of differences in the experimental setup. Barbour [2003], for example, used stereophonic loudspeakers for his study as opposed to the quadraphonic configuration used in the present. In addition, the elevation of the height layer with respect to the main was different, with the present experiment only considering elevation angles of 30°. Further differences in Barbour's [2003] study include that the loudspeaker setup was not obscured from the view of subjects, the effect of ICTD was not considered and neither were main and height layer only conditions tested.

Although it was mentioned that Barbour [2003] did not consider perceived source location for the main and height layer only conditions, it can be inferred, based on Lee's [2011] -9 to -10 dB masking thresholds for cello and bongo sources presented from vertically arranged stereophonic loudspeakers, that the signal in the height layer was inaudible for the -15 dB ICLD condition. For 45° height layer elevation, Barbour [2003] reported that the perceived location of sources for the -9 dB ICLD was similar to that for -15 dB. This result can be interpreted in the following way. When the ICLD was -9 dB, the resultant phantom image was localised at same position as when the same source was presented from the main layer only. Further, as the ICLD increased to -6 dB, differences in the perceived elevation of the main layer only and phantom image conditions became more apparent. This result closely matches what was reported in the present experiment. When there was 0 ms ICTD, the localisation threshold using the blanket reduction method was -9.5 dB, with the results of the present experiment showing that localisation judgments were not significantly different compared to those for the main layer only condition. Both results therefore suggest that vertical interchannel crosstalk will affect the perceived location of the main channel signal if the ICLD is greater than -9 dB (when there is no ICTD present).

With respect to the perceived elevation of stimuli, an interesting difference can be seen between the two sets of data in the case that the ICLD was 0 dB. Barbour's [2003] results suggested the following. When a sound

source is presented as a vertically oriented phantom image with 0 dB ICLD, the perceived location of the resultant image is not directly in between those of the same sound source presented from the main and height layer alone (i.e. the ±15 dB ICLD conditions). Instead, localisation judgments are at a position biased towards those for the main layer only condition. Although a similar result was reported in the present study, a key difference should be identified. In the Barbour [2003] study, localisation judgments for the 0 dB condition were biased towards the physical position of the main layer, being around 15° for height layer elevation of 45°, 25° for 60° elevation and 22° for 90° elevation. Conversely, in the present experiment the stimuli presented with 0 dB ICLD and 0 ms ICTD were localised roughly in the middle of the physical position of each layer (12°-16°). The difference then lies in the perceived elevation of the main layer only conditions. Where Barbour [2003] reported the perceived location of the stimuli presented with -15 dB ICLD to be at the exact position of the main layer, the main layer only condition in the present study yielded localisation judgments that were significantly higher than the physical position of the loudspeaker layer, with median perceived elevation ranging from between 5.8° for the conga and 13.0° for the oboe.

The difference in the perceived location of the main layer only conditions in the respective studies can be explained by the phantom image elevation effect, which is described in detail in Chapter One. According to this effect, the perceived elevation of the phantom centre image emitted from coherent loudspeakers arranged on the horizontal plane is determined by the base angle between the loudspeakers; a greater angle corresponds to an increased sense of elevation [De Boer 1947, Lee 2017]. Part of the motivation of analysing the effect of presentation method in Experiments Three and Four was that it is known that the effect would operate for the quadraphonic condition and not for the stereophonic, which might have affected the subsequent thresholds. That there was no significant effect was explained as follows. Previous research has indicated that the phantom image elevation effect is also maintained for elevated loudspeakers [Lee 2016]. Therefore, it can be argued that the effect would also operate for vertically oriented phantom images. This would mean that the 0 dB phantom image for the quadraphonic condition would be perceived as being elevated with respect to that for the vertical stereophonic condition, with the difference in elevation resulting from the two methods being similar to that for main layer only presentation. By extension, this would mean

similar perceptual elevation differences between the 0 dB and main layer only conditions for each method, which was one of the hypotheses used to explain the non-significant effect of presentation method in Experiment Three.

The results of the present experiment can be used to further scrutinize the above hypothesis. However, although evidence of the phantom image elevation effect can be seen for main layer only presentation, the data in Fig. 4.21 does not support the hypothesis that the effect would be maintained for elevated loudspeakers. Instead, with the exception of the oboe, median localisation judgments for all sources presented using the height layer only condition were slightly lower than the physical position of the height layer. If it is the case that the effect does not operate for elevated loudspeakers, then it seems reasonable to conclude that there would equally be no effect for vertically oriented phantom images. This would therefore suggest that the perceived difference in elevation between the main layer only and phantom image conditions would be different for each presentation method. This result would further indicate that the localisation thresholds for natural sound sources are not related to the perceived difference in elevation between the main layer only and 0 dB phantom image conditions. It should be noted, however, that this would have to be studied further. This is partly because the phantom image elevation effect has not yet been explored fully for elevated loudspeakers or for vertically oriented phantom images.

4.3.3.5 Practical Implications

The results of the present experiment indicate that there are numerous ways to incorporate direct sounds in the height layer without the perceived location of the main channel signal being affected by vertical interchannel crosstalk. This has implications both for microphone techniques for recording for 3D audio formats and for the rendering of 3D images. It was earlier discussed how the localisation threshold did not represent a complete masking of the direct sound in the height layer and further that attributes such as the timbre and spaciousness of the main channel signal would be affected when the localisation threshold is

applied. It is somewhat apparent that the perception of each of these would vary depending on the localisation threshold method being used. For example, given that the 1+B, 8B and 4B methods do not require any attenuation of the low frequencies in the height layer signal, it could be the case that the resultant main channel signal would sound fuller compared to the blanket reduction method. In addition, given that Furuya et al. [1995] reported that the perception of VIS was related to the amplitude of a vertical reflection relative to the direct sound, it could be that the degree of VIS afforded by each method would differ. This is based on the data in Table 4.2, which shows that the amplitude of the height layer when the localisation threshold was applied was dependent on the method being used. It is apparent then that further study is needed to ascertain what the most salient effects of vertical interchannel crosstalk are, how these vary when the different localisation threshold methods are applied and which method is the most preferred by subjects. Each of these points is considered in Experiment Six.

4.3.4 CONCLUSION

In the present experiment, band and blanket reduction thresholds, which were obtained from Experiments Three and Four respectively, were verified. The test stimuli were the same natural sound sources as were used in Experiment Three and were presented to subjects using the vertical quadraphonic method. Stimuli were presented to subjects from the height and main layers only and as vertically oriented phantom images. For the phantom image conditions, the height layer was presented either with no attenuation (0 dB ICLD) or with one of five localisation threshold methods applied (blanket, FB, 1+B, 4B, 8B). The aim of the experiment was to determine which of the localisation threshold methods resulted in perceived source location similar to that of the same source presented from the main layer only. Delays of 0 and 1 ms were applied to the height layer with respect to the main.

The experimental data showed that the both the blanket reduction and each of the four tested band reduction methods were successful in preventing the effects of vertical interchannel crosstalk on the perceived location

of the main channel signal. Of particular interest was the result that the 1+B, 4B and 8B methods were all effective. For the 8B and 1+B methods, this result was interpreted based on the supposed importance of the 7-9 kHz region in determining the localisation threshold for complex signals. However, this hypothesis was insufficient for explaining the results for 4B, which did not necessitate attenuation in this region. This result suggested that the 7-9 kHz region might not be as important in determining the localisation threshold as had previously been suggested. Additionally, that the difference in spectral energy between the main and height layers in the frequency region governed by the 4 kHz octave band was small suggested that a mechanism other than the relative balance of spectral energy might be important in determining the localisation threshold for natural sound sources.

The results of the study also suggested that vertical interchannel crosstalk had a source dependent effect on the perceived elevation of the sound source. This was interesting for two reasons. Firstly, in Experiment Three the localisation thresholds using the blanket reduction method were not source dependent. Secondly, judgments for the oboe source were consistent no matter how the source was presented to subjects, with localisation likely being related to the pitch-height effect, and yet ICLD was always necessary for this source for the phantom image position to match that of the main layer only. These results suggest that the localisation threshold for natural sound sources is not related to the perceived difference in elevation between the main layer only and 0 dB phantom image conditions. Instead, differences in VIS between the two conditions might explain the results obtained, although this would have to be studied further.

In addition, the experimental data showed no evidence to support the operation of the precedence effect or localisation dominance. With respect to the latter effect, the presence of a delay caused the phantom image to move closer to the later loudspeaker, in a manner similar to that observed for broadband pink noise in Experiment Two. Therefore, the hypothesis raised following Experiment Three, that the significant effect of ICTD on the localisation thresholds using the blanket reduction method was due to the operation of a localisation dominance effect, was rejected. Further study would therefore be needed to explain this result.

A further result of note was that localisation judgments for the main layer only condition were found to be significantly higher than the physical position of the main layer. This was interpreted based on the phantom image elevation effect. Following Experiment Three, it was hypothesised that this would also operate for height layer only presentation. As a consequence of this, it was thought that the perceived difference in elevation between the main layer only and phantom image conditions for each presentation method would be similar, which was one reason postulated as being the cause of the non-significant effect of presentation method observed in Experiments Three and Four. However, the results of the present experiment showed no evidence of the effect for elevated loudspeakers. This further suggests that the localisation threshold is not related to the perceived difference in elevation between the main layer only and 0 dB phantom image conditions.

The results obtained in the study show that there are numerous ways in which direct sounds can feature in the height layer without the perceived location of the main channel signal being affected. However, it remains unclear how the timbre and spaciousness of the main channel signal would vary as a result of each method and this will be considered in Experiment Six.

4.4 SUMMARY

This chapter has explored localisation thresholds in detail for both the blanket and band reduction methods. Experiment Three was conducted as expansion of the work conducted by Lee [2011] and Stenzl et al. [2014], considering localisation thresholds for natural sound sources using the blanket reduction method. The study also considered the effects of presentation method, signal duration and ICTD. Experiment Four built upon the data obtained in Experiment One, analysing localisation thresholds both for octave bands of noise and broadband pink noise in the presence of reflections. As with Experiment Three, the effects of presentation method and ICTD were considered. In Experiment Five, a series of band reduction methods were derived based on the data provided from Experiment Four. These were then tested in localisation experiment along

with the blanket reduction thresholds from Experiment Three. The perceived location of sources presented from the height layer only was also considered, along with vertically oriented phantom images with 0 dB ICLD (i.e. maximum crosstalk). The key findings from each experiment are as follows:

Experiment Three:

- The localisation thresholds for the natural sound sources tested were only significantly affected by changes in ICTD. The threshold was -9.5 dB for 0 ms ICTD and -7 dB for 1 and 10 ms.
- The non-significant effect of sound source might be explained by the suggestion that the mechanism that determines whether or not the localisation threshold has been met for complex sources is the balance of spectral cues provided by the main and height layers, particularly in the 7-9 kHz region. This is determined primarily by the ICLD and is not affected by the spectral content of the source itself.
- The non-significant effect of presentation method was explained on the basis that increases in ICLD resulted in similar differences in spectral energy in the 7-9 kHz region between the main layer only and phantom image conditions for both presentation methods.
- The results were indicative of a localisation dominance effect of the earlier loudspeaker, although the precedence effect itself was not observed.

Experiment Four:

- The frequency dependency of localisation thresholds was maintained in the presence of reflections, albeit the effect was not as strong as was observed in anechoic conditions (Experiment One).
- The effect of frequency was less strong compared to Experiment One. It was suggested that this might be related to comb-filtering effects as a result of floor reflections.

- Differences in the filtering and threshold detection methods might also have caused differences in the results between the two experiments.
- The localisation thresholds obtained were not affected by presentation method. For the octave band stimuli, this might be because the frequency-dependent differences in perceived VIS between the main layer only and phantom image conditions differ little between vertical stereophonic and quadraphonic presentation.
- For the broadband source, the result might have been caused by the same mechanism as was suggested for natural sound sources following Experiment Three.
- The effect of frequency was dependent on signal duration, being stronger for continuous presentation compared to bursts. This might be because the full-extent of perceived VIS requires time to build up. As such, the duration of the burst stimuli might have been too short to enable frequency-dependent differences in VIS between the main layer only and phantom image conditions to be perceptible. This would also explain why the threshold was higher for the bursts compared to the continuous stimuli.
- ICTD had a random and inconsistent effect on a limited number of stimuli. Significant effects were observed for 4 kHz bursts, 8 kHz continuous and broadband pink noise for both durations. This might be related to the random effects of comb filtering, with the effect being too inconsistent to be indicative of a localisation dominance effect.
- No evidence was found to support the operation of the precedence effect in the median plane.

Experiment Five:

- The blanket reduction thresholds obtained in Experiment Three are sufficient at preventing vertical interchannel crosstalk from affecting the perceived location of the main channel signal. Each of the band reduction methods tested (FB, 1+B, 4B and 8B) were also effective.

- That the 4B method was effective indicates that the balance of spectral energy in the 7-9 kHz region between each layer might not be as important in determining the localisation threshold for natural sound sources as had been suggested previously.
- Vertical interchannel crosstalk has a source dependent effect on the perceived location of the main channel signal. That the localisation thresholds obtained in Experiment Three were not source dependent indicates that the threshold does not necessarily relate to the perceived elevation differences between the main layer only and 0 dB phantom image conditions.
- The oboe source might have been localised based on the pitch height effect. If this is the case, then the balance of spectral energy hypothesis would not explain the localisation thresholds obtained for this source. Instead differences in VIS might be the key mechanism. This might also apply to other natural sound sources.
- No localisation dominance effect was observed; a delay in the height layer instead caused perceived source elevation to increase. Equally no evidence was found for the precedence effect. It is therefore unclear why less level reduction was necessary in the presence of an ICTD in Experiment Three.
- The phantom image elevation effect might not operate for vertically oriented phantom images.

5 EXPERIMENT SIX: THE PERCEPTUAL EFFECTS OF VERTICAL INTERCHANNEL CROSSTALK

The present chapter consists of one experiment, which is divided into four parts. In Part One, the perceptual effects of vertical interchannel crosstalk were elicited. Following this the audibility of each of the elicited attributes was graded in order to establish which were the most salient (Part Two). Then, it was determined how the perception of those salient attributes was affected when natural sound sources were presented using the various localisation threshold methods developed in Chapter Four (Part Three). The final part of the experiment (Part Four) considered the subjective preference of the localisation threshold methods.

Up to this point, the experiments reported in the present thesis have predominantly been focused on preventing vertical interchannel crosstalk from affecting the perceived location of the main channel signal. However, it is somewhat apparent that this represents just one effect that would be perceived as a result of the presence of excessive direct sounds in the height layer. As was discussed in Chapter Two, it is likely that the perceived timbre and spaciousness of the main channel signal would be affected, although the precise nature of this has not yet been reported in the literature. The primary aim of the present experiment then is to conduct such an analysis and further to determine those effects of vertical interchannel crosstalk that are the most salient.

Alongside the elicitation of the most salient effects of vertical interchannel crosstalk, a further aim relates to the localisation threshold methods derived from Chapter Four. As was discussed by Lee [2011], the localisation threshold does not represent a complete masking of the direct sound in the height layer. As a result, it can be argued that some of the more salient effects would be somewhat audible, even if the perceived location of the main channel signal were unaffected. Given that the aforementioned band and blanket reduction methods involve attenuating the direct sound in the height layer in different ways, it can be

argued that the perception of such effects would depend on the localisation threshold being used. It is of interest then to determine how the most salient effects of vertical interchannel crosstalk vary when the different localisation thresholds, as derived in Chapter Four, are applied.

An additional aim is to analyse the subjective preference of localisation thresholds. Firstly, it is of interest to determine if conditions whereby direct sound is present in the height layer at the localisation threshold are more preferred with respect to the direct sound being either masked or absent entirely. Should subjects prefer the latter condition then this would suggest that the masking threshold, whereby the direct sound in the height layer is inaudible [Lee 2011] should be analysed in further detail. Conversely, should the localisation threshold conditions be preferred then it would be of interest to determine how blanket reduction compares to each of the different band reduction methods. It is also considered as being important to determine the reasons behind the preference gradings for each, as this might influence which method is chosen for a given practical situation.

From the above background the following research questions were derived:

- What are the most salient perceptual effects of vertical interchannel crosstalk?
- How does the audibility of the most salient effects change when the amplitude of the vertical interchannel crosstalk signal is reduced to the localisation threshold using the different methods developed in Chapter Four?
- Which localisation threshold methods are the most preferred by subjects?

5.1 EXPERIMENTAL HYPOTHESIS

The first null hypothesis for this experiment is that vertical interchannel crosstalk has no audible effect on the main channel signal. It is thought that this null hypothesis will be rejected based on the plethora of

literature describing the perceptual effects of secondary vertical sources as reviewed in Chapter Two of the present thesis, as well as the informal listening conducted following Experiment One-Five. From these sources, it is hypothesised that two of the more apparent effects of vertical interchannel crosstalk will be increases in perceived loudness and source elevation. The former stems from the law of addition of energies, in that the inclusion of additional direct sound in the height layer will naturally increase the perceived amplitude of the main channel signal. The latter effect has been demonstrated in the present thesis, most notably in Experiment Five. With respect to the effect on perceived timbre, Lee [2006] found that horizontal interchannel crosstalk affects the brightness, fullness and hardness of the main channel signal, among other attributes. However, the audibility of such perceptual changes was found to be relatively low. It is thought in the present study that similar perceptual effects will be reported although they will be much more audible than was reported in the Lee [2006] study. This is based on the data reported by Barron [1971] and Barron and Marshall [1981], that timbral colouration is more audible for vertical reflections than for horizontal. However, it should also be noted that the amplitude of the crosstalk signal in the Lee [2006] study was reduced with respect to the main signal by between -4 and -20 dB, which may have limited the audibility of the perceived timbral effects. In terms of spatial impression, the key effect is expected to be an increase in perceived VIS in line with the informal observations reported following Experiment One.

The second null hypothesis for the present experiment is that the perceptual effects of vertical interchannel crosstalk will not differ between different localisation threshold methods. It is anticipated that this null hypothesis will be rejected for the following reasons. As is shown in Table 4.2, each of the derived localisation threshold methods results in a different amount of attenuation of the direct sound in the height layer, with the amplitude of the 8B and 4B conditions (72.4 dB LAeq, 72.2 dB LAeq) being notably louder overall than the blanket and FB conditions (70.4 dB LAeq, 70.0 dB LAeq). It is therefore thought that that the audibility of the most salient perceptual effects of vertical interchannel crosstalk will be dependent on the relative amplitudes of the direct sound in each layer. This seems logical given that changes in both timbre and spatial impression have a notable dependency on the amplitude of the secondary source. As a result, it is thought that the most salient perceptual effects of vertical interchannel crosstalk will be more audible for 8B

and 4B than they will be for either FB or blanket. Moreover, as its amplitude falls in between that of the aforementioned stimuli (71.2 dB LAeq), it is though that perception for 1+B will be somewhere in-between those for the other methods.

5.2 GENERAL METHODOLOGY

The attribute elicitation method chosen for the present experiment was based on the Quantitative Descriptive Analysis (QDA) method. This method involves a panel of assessors developing a set of common attributes to describe their perceptions of the stimuli under investigation [Lorho 2005]. Bech and Zacharov [2006] considered the QDA method to have six 'phases':

1. Subjects are presented with a representative group of stimuli and are required to provide descriptive terms for the attributes they perceive.
2. Duplicate attributes are either removed or grouped by subjects to minimize the number available.
3. The attributes are further discussed and reduced. Additionally, stimuli that represent good examples of the attributes are identified.
4. Simple test stimuli are introduced that activate all attributes at a wide range of intensities. These stimuli are used to introduce the concept of scaling. Subjects will discuss the use of the scale and define end-point terms for each scale.
5. Stimuli with smaller perceptual differences are introduced and the scaling exercise is repeated. The inclusion of repetitions is so that the consistency of subjects can be monitored. If there is a large inter-subject variance for a given attribute then this may indicate the presence of a multidimensional attribute; this should be divided up into other attributes.
6. Actual tests are conducted in which the elicited attributes are graded for each stimulus.

One of the key benefits of the QDA method is that it provides a complete description for the sensory properties of a given stimulus [Stone and Sidel 2004]. This was considered as being important as the perceptual effects of vertical interchannel crosstalk have not previously been elicited and, as such, it was desired to utilise a method that would elicit as many attributes as possible. In addition, the use of a common scale for all subjects enables statistical analysis that can infer the performance of a wider population [Berg and Rumsey 1999]. However, despite these benefits there are a number of drawbacks. One of the most notable of these is the potential duration of the process. Stone and Sidel [2004] suggested that each training session could take around 90 minutes, with potentially 4-5 sessions being necessary before the process is completed. Francombe [2014], in elicitation experiments conducted to determine the perceptual attributes of audio-on-audio interference, suggested that the attribute grouping process alone required three sessions that lasted around 5 hours in total. The organization of this can become especially difficult given that ideally between 10 and 12 subjects are necessary for the technique to be effective and all must be present for all stages Stone and Sidel [2004].

The potential duration of the QDA method was seen as being somewhat of an issue. As such, alternative methods of attribute elicitation were considered including the Repertory Grid Technique (RGT). This method involves subjects being presented with triads of stimuli and describing how two of them are similar and how the third one is different [Berg and Rumsey 1999, Bech and Zacharov 2006]. An individualized set of attribute scales is subsequently derived, with subjects using these scales to grade their perceptions. Berg and Rumsey [1999] argued that subjects are more reliable in tests in which they have developed their own attributes rather than being provided with them, as would be the case with QDA. Additionally, the method is well suited to naïve subjects [Lorho 2005].

Despite the benefits of RGT there are notable drawbacks that made it insufficient for use in the present Experiment. Bech and Zacharov [2006] suggested that the presentation of elements as triads does not specifically ask subjects to produce bipolar attributes. As a result, key attributes may be missed by subjects. Given the previous discussions on the desire to elicit as many attributes as possible this was viewed as being

problematic. In addition, Berg and Rumsey [1999] suggested that RGT is not suited for simple statistical analysis, as individual subjects may develop differing constructs. Therefore, more advanced methods such as principal component analysis (PCA) and multidimensional scaling (MDS) may become necessary for the analysis of results. Such complex analysis is not necessary as a result of QDA, which uses the same scale for all subjects and is therefore considered as being much simpler to implement.

Based on the above discussion it was decided that a modified version of the QDA method, one that limits the duration of the process, should be considered for testing. Such an approach was taken previously by Lee [2006]. For that elicitation experiment, subjects were presented with a list of potential attributes that had been elicited from previous studies. They were then asked to grade the audibility of changes within each attribute, with the most audible (salient) being used for the main grading experiment. Additionally, the grouping of elicited attributes was performed by the experimenter, following informal discussions with subjects, rather than through group discussions by the subjects themselves. The benefit of such a method is that it removes the lengthy discussion processes required by QDA, which makes the experiment both quicker and simpler to organise. However, it is arguable that this process biased the experiment somewhat, as the attributes graded were based on one individual's interpretation of each of the descriptions provided by subjects. Based on this it is apparent that, when using a modified QDA method, some consideration must be given to the effect of trying to limit the timescale for the experiment on the results obtained.

In order that the QDA method might be simplified for the present study, whilst limiting any potential bias, the experiment was conducted in three separate parts:

- Part One: Elicitation of the perceptual effects of vertical interchannel crosstalk.
- Part Two: Grading of the audibility of the perceptual effects.
- Part Three: Grading of the most salient effects when the localisation thresholds are applied.

It should be noted that a fourth part (Part Four) was added in order to determine the subjective preference of the localisation threshold methods. For each of the aforementioned parts the physical setup (room, loudspeaker arrangement, equipment, etc.) was identical to that used for Experiment Five. The quartet, speech, conga and guitar sources were used as the test stimuli and were presented using the vertical quadraphonic condition. The oboe source was removed based on the results of Experiment Five, which showed that the source was little affected by the elevation effects of vertical interchannel crosstalk. ICTDs of 0 and 1 ms were applied to the height layer with respect to the main layer. The range of ICLDs tested for each part varied as described over the next few sections.

5.3 PART ONE: ELICITATION OF THE PERCEPTUAL EFFECTS OF VERTICAL INTERCHANNEL CROSSTALK

The purpose of this part of the experiment was to elicit the perceptual effects of vertical interchannel crosstalk and to subsequently group them into a set of common attributes.

5.3.1 Test Method

This part of the experiment was divided into two sections. The first section required subjects to complete an individual free elicitation test using a Max/MSP interface (Fig. 5.1). For each trial, subjects were presented with a given stimulus presented using two different conditions; the stimulus presented from the main layer only (no crosstalk) and; the same stimulus presented from both the height and main layers together with 0 dB ICLD (maximum crosstalk). The purpose of the latter condition was to emulate the maximum possible interference of the height channel signal on the main in order that each of the perceptual effects of vertical interchannel crosstalk would be as audible as possible. For each trial, each of the two conditions was

allocated randomly to buttons ‘A’ and ‘B’ on the test interface. The task for subjects was to write down any way in which sound ‘A’ was perceptually different from sound ‘B’ for each of the eight trials (four sources, two ICTDs). They were provided with a test sheet that contained a table made up of 8 rows (numbered from 1-8) and were to record all responses in the table row corresponding to the current trial. At the end of the test, the order of the stimuli was noted so that each subject’s responses could be allocated to the relevant stimuli in preparation for the second part of the test. To avoid bias, subjects were given no guidance on what sort of differences they should listen for and instead were encouraged to list every audible variation. Each subject required 20-30 minutes to complete this exercise.

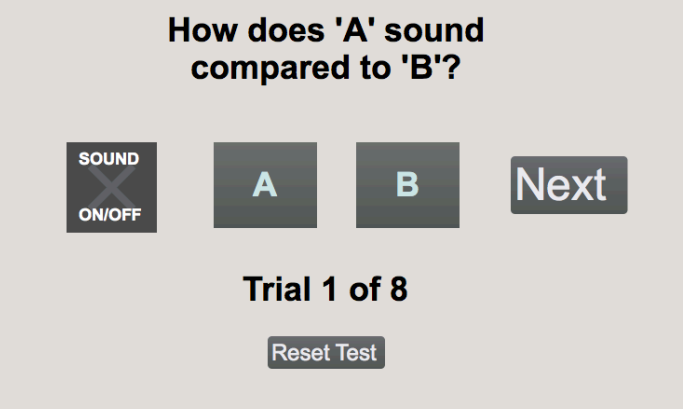

Fig. 5.1: Max/MSP interface used for the individual elicitation test.

The second exercise for this part of the experiment was a group discussion. For this, each of the subjects who had participated in the free elicitation test grouped all of the elicited terms into a set of common attributes and developed descriptions for them. In order that the timescale for this process might be reduced, a series of attributes and descriptions were provided at the start of the discussion. The majority of the terms and descriptions were those elicited by Lee [2006] for horizontal interchannel crosstalk. However, additional terms, such as ‘loudness’ and ‘vertical image spread’ were added due to their relevance to the present study. In all 11 attributes and descriptions were provided (Table 5.1). The group required around 2 hours to complete this exercise.

Table 5.1: Attributes provided for the group discussion.

Attribute	Definition
Horizontal Source Width	The perceived horizontal width of a sound source
Vertical Image Spread	The perceived vertical width of a sound source.
Locatedness	The easiness of localisation of a sound source.
Source Elevation	The perceived elevation of a sound source.
Fullness	The timbral characteristics of a sound depending on the level of low frequencies.
Source Distance	The perceived distance from the listener to a sound source.
Hardness	The timbral characteristics of a sound depending on the level of mid/high frequencies (especially 2-4 kHz).
Brightness	The timbral characteristics of a sound depending on the level of high frequencies.
Naturalness	The perceived degree of 'realism' of a sound.
Envelopment	The sense of being surrounded by a sound or that the sound is all around.
Loudness	A perceived change in the amplitude of a sound.

A number of steps were taken in order to limit bias in this part of the experiment. Firstly, the list of attributes provided for the group discussion was not available to subjects for their individual free elicitation task; this was to keep the elicitation as free as possible and therefore dependent solely on the subject. Additionally, subjects were encouraged to use the provided attributes in the group discussion only for those elicited terms that fit them. If an elicited term did not fit any of the attributes then the group was required to name a new attribute and develop a new description. Moreover, whilst the author sat in on the group discussion his role was solely to act as the chair and to move the discussion along. At no point did the author suggest a specific term or attempt to influence the direction of the discussion in any way.

In total six subjects participated in this stage of the experiment. Each of these were postgraduate students and members of the University of Huddersfield's Applied Psychoacoustics Lab. All were highly experienced

with critical listening. As mentioned earlier, it is recommended that between 10 and 12 subjects participate in QDA experiments in order that the results are not biased towards the perception of a few individuals [Stone and Sidel 2004]. However, given the nature of the experiment it was found that assembling such numbers proved difficult. Instead, expert listeners were specifically chosen in order that as many perceptual attributes could be elicited as possible with the limited numbers available. In addition, all subjects were encouraged to contribute equally in the group discussion in order that the results did not reflect the opinions of any particular individual too strongly.

It should be noted that it was chosen to not loudness match the no crosstalk and maximum crosstalk conditions. Instead, the amplitude of the stimuli when presented from each individual layer was 70 dB LAeq. As a result, the maximum crosstalk condition (main and height layers together) was around 3 dB louder than the no crosstalk condition (main layer only). This naturally leads to the question of whether any attribute changes identified by subjects would have arisen purely as a result of this amplitude difference. This was seen as a somewhat unavoidable issue as, by its very nature, vertical interchannel crosstalk relates to the presence of an increased amount of direct sound. Therefore, amplitude increases are somewhat inevitable. As a result of this, it was regarded that level matching the maximum and no crosstalk conditions would not be representative of a practical situation in which vertical interchannel crosstalk was present.

5.3.2 Results and Discussions

Table 5.2 shows the elicited terms and the attributes in which they were grouped. All 11 of the attributes provided for the group discussion were used. In addition, the group proposed two further attributes, ‘richness’ and ‘resonance’. ‘Richness’ was described as ‘an even balance and extension of frequencies across the spectrum’, whilst ‘resonance’ was considered to be ‘a boosting of certain frequency bands within a source’. A high number of the elicited terms were found to relate to source fullness (43); vertical and horizontal image spread (32/30); loudness (34) and elevation (30). However, as the audibility of each attribute has not

yet been determined, it should not be concluded that the aforementioned attributes are the most salient. Equally, this means that the attributes that featured few elicited terms do not necessarily have low audibility. For example, even though only two elicited terms were considered as relating to 'hardness' it may be that when subjects are specifically listening for changes in that attribute that variations might become more apparent.

Table 5.3 shows the frequency that changes in each attribute were perceived for each sound source. Here it can be seen that the majority of the attributes were perceived consistently for each sound source, with the range for 10 of the 13 sources being 3 or less. However, there does appear to be some minor source effect for vertical image spread, horizontal source width and fullness. For example, for fullness it can be seen that the conga, speech and quartet sources each had either 4 or 5 occurrences, whilst for the guitar source there were 9 occurrences for 0 ms ICTD and 7 for 1 ms. Equally, for vertical and horizontal spread some sources featured as many as 6 occurrences, whilst others featured as few as 2. The specific reasons for certain sources and ICTDs causing fluctuations in the frequency that certain attributes were perceived is unknown and requires further study.

Table 5.2: Results of the elicitation and grouping exercise.

Attribute	Number of Elicited Terms	Elicited Terms
Horizontal Source Width	30	Spacious (2), narrow (5), larger horizontally, larger image (2), narrower spatially (5), wider (4), horizontal spread (3), squashed, spread out (2), frontal envelopment (2), open, wider horizontally, wider spatially.
Vertical Image Spread	32	Spacious (2), vertically spread (3), wider height spread (3), larger vertically, larger image, narrow (4), wider vertical spread (3), sonically spread vertically, squashed, spread out, more sound from above, frontal envelopment (2), more separation between instruments, open, wider vertically, spread out, larger image, more vertical image spread (3), wider.
Locatedness	12	Focused (4), localizability, precision (3), more separation between instruments, more clarity, blurred localisation, less focused, hard to localise.
Source Elevation	30	Same height (2), vertical image shift (6), elevated (6), lower height (6), more sound from above (2), image shift (2), image is elevated, localised in the sky, voice of god, lower in space, higher source position, localised lower.
Fullness	43	Thin (15), full (7), less low frequency content (7), less bottom end, low/mid frequency boost (2), warmth (6), body (2), full bodied, more low energy, delicate
Source Distance	4	Distance, further away (2), closer.
Hardness	2	Soft, heavier.
Brightness	19	Clearer (7), duller (4), more high frequencies (3), muffled (2), brighter, woolly, less clarity.
Naturalness	6	Coloured (3), natural, detailed (2).
Envelopment	2	Spacious, reverberant.
Loudness	34	Louder (23), quieter volume (8), lower volume, quieter, delicate
Richness	2	Richer (2).
Resonance	9	Resonant (3), boomy (4), nasal, sibilant.

Table 5.3: The frequency that changes in each attribute were perceived for each sound source.

Source	**Guitar**		**Quartet**		**Speech**		**Conga**		**Range**
ICTD (ms)	**0**	**1**	**0**	**1**	**0**	**1**	**1**	**0**	
Horizontal Source Width	4	6	4	5	4	2	2	3	4
Vertical Image Spread	6	4	6	2	4	3	3	4	4
Locatedness	1	2	1	2	1	2	1	2	1
Source Elevation	4	2	4	4	4	5	4	3	3
Fullness	9	7	5	5	4	4	5	4	5
Source Distance	1	0	1	1	1	0	0	0	1
Hardness	1	0	0	0	0	0	1	0	1
Brightness	1	3	2	3	2	3	3	2	2
Naturalness	1	0	1	0	1	1	1	1	1
Envelopment	0	0	0	0	0	2	0	0	2
Loudness	6	4	5	4	4	4	4	3	3
Richness	1	0	0	0	1	0	0	0	1
Resonance	1	0	1	1	1	0	2	3	3

In Table 5.4 the elicited perceptual attributes have been categorised according to the perceptual effects of secondary vertical sources (location, spatial impression and timbre) as discussed in Chapter Two. It can be seen that six out of the thirteen elicited attributes related to changes in perceived timbre, with a further three each for location and spatial impression. Loudness was considered to be a separate perceptual attribute. At this stage it is not possible to determine whether the elicited attributes are increased or decreased as a result of vertical interchannel crosstalk as the maximum and no crosstalk conditions were randomly allocated to buttons 'A' and 'B' on the test interface. The only exceptions to this are 'source elevation', which was shown in Experiment Five to increase as a result of vertical interchannel crosstalk, and 'loudness', which would naturally have increased due to the aforementioned presence of increased levels of direct sound.

Table 5.4: Elicited attributes grouped by type.

Spatial Impression	**Timbre**	**Location**	**Other**
Horizontal Source Width	Fullness	Locatedness	Loudness
Vertical Image Spread	Hardness	Source Elevation	
Envelopment	Brightness	Source Distance	
	Naturalness		
	Richness		
	Resonance		

As was mentioned in Section 5.3.1, it could be argued that subjects might have perceived a number of the elicited terms as a result of the differences in loudness between the maximum and no crosstalk conditions. At this stage any discussions are somewhat speculative, mainly because it is unclear whether the majority of attributes increased or decreased with changes in condition. However, suggestions as to the potential effects of loudness can still be made. For example, with respect to spatial impression, Cabrera and Tiley [2003] demonstrated that the perceived vertical and horizontal image spread of broadband and octave band pink noise increased with loudness. In this case the stimuli were presented from single loudspeakers, with the amplitude of stimulus presentation varying between 64 and 84 phon. In addition, perceived source distance may have decreased with increases in loudness, as the louder source would naturally have seemed closer to the listener. Furthermore, the timbral attributes may have been affected by loudness. Attributes such as 'fullness' for example, might have appeared to change simply because the low frequencies became more audible as the amplitude increased. This of course would rely on assumptions regarding the direction in which changes in the attribute were perceived as a result of vertical interchannel crosstalk, which are not possible to make with certainty based on the present experimental data. However, this at least highlights the influence that loudness may have had on the results obtained.

Despite the above discussion, it should not be considered that the attributes elicited were done so purely because the maximum and no crosstalk conditions were not loudness matched. For example, subjects would likely have perceived increases in VIS as a result of the introduction of the direct sound in the height layer. In this case then it can be argued that the difference in amplitude between the two conditions might only have made the variations in perceived VIS even more audible. In addition to this, variations in source elevation are primarily related to the ICLD between the upper and lower layers, as demonstrated by Somerville et al. [1965] and Barbour [2003]. As a result, differences in amplitude between the maximum and no crosstalk conditions are not considered as being particularly important with regards to that particular attribute. Also, in terms of timbre, vertical reflections have been shown to alter the perception of the direct sound. Barron and Marshall [1981] noted that lateral reflections caused musical sources to gain body and fullness, further noting that the effects are more noticeable for vertical reflections. Additionally, Halmrast [2000] found that vertical reflections result in orchestral music sounding 'boxy'. Therefore, timbral variations would also have been expected, even if the amplitude of the two conditions were equal.

Based on the above discussion, it can be argued that the attributes elicited in the present experiment were not done so purely because the max and no crosstalk conditions were different in amplitude. However, it should be noted that this would have to be studied further. Instead, it is more likely that the primary effect of loudness was to influence the overall audibility of the elicited attributes. At present, a series of perceptual effects of vertical interchannel crosstalk have been elicited. Part Two will analyse how audible each of these are in order to determine those that are the most salient.

5.4 PART TWO: GRADING OF THE AUDIBILITY OF THE PERCEPTUAL EFFECTS

In Part One of the present experiment, 13 perceptual effects of vertical interchannel crosstalk were obtained through a process of individual elicitation tests and group discussions. However, it should be noted that the audibility of changes within each attribute as a result of the effect are not yet known. By extension, this

means that the most salient effects of vertical interchannel crosstalk have not yet been determined. The aim of this part of the experiment was to determine the most salient effects of vertical interchannel crosstalk by considering the audibility of each of the 13 attributes elicited in Part One. Once this has been completed it will then become possible to grade the audibility of each of the most salient attributes when the different localisation thresholds are applied (Part Three).

5.4.1 Test Method

For this part of the experiment, subjects were required to sit 13 audibility tests, each of which focused on one of the attributes elicited during Part One. The general methodology was similar to that used for Part One in that subjects were asked to compare 'no crosstalk' and 'maximum crosstalk' conditions, which were randomly allocated to buttons 'A' and 'B' on a Max/MSP interface (Fig. 5.2). In this case, however, subjects were required to identify how audible changes in a given attribute were between 'A' and 'B' using a 10-point scale. The scale included anchor points corresponding to the perceived degree of audibility, with 1 being labeled as 'Just Audible', 4 as 'Slightly Audible', 7 as 'Audible' and 10 as 'Very Audible'. Bech and Zacharov [2006] recommend the use of anchors in order to minimise 'end-point effects' (i.e. subjects being reluctant to use the end points of scales in anticipation of extreme stimuli), which can result in responses being grouped towards the middle of the scale. They suggest that, ideally, audible anchors 'that illustrate the range of impressions' should be used during a training phase to limit this issue. However, the use of audible anchors was problematic for the present experiment because the test conditions required subjects to directly compare no crosstalk and maximum crosstalk conditions. It was therefore not possible to provide subjects with stimuli that represented the maximum audibility of the effect because arguably these would have been identical to the test stimuli. It was also difficult to gague how changes in ICLD would have affected the audibility of a given attribute and therefore intermediate anchors were also difficult to provide with certainty. As a result of these limitations, it was decided to use semantic labels on the scale as anchors instead of providing audible anchors. It should be noted that Bech and Zacharov [2006] urge caution when using this

method, as semantic anchors can affect the resolution of the scale. It was deemed, however, that this would be unavoidable given the nature of the experiment. The same eight stimuli that had been used in Part One (four sources, two ICTDs) were used for Part Two. It should be noted that subjects were specifically instructed to grade the perceived *difference* in the magnitude of each attribute between 'A' and 'B'. For example, they were informed that if both stimuli featured high amounts of the attribute being tested then the grading should be low; a high grading should only be given if there was a notable difference between the two stimuli with respect to the attribute in question.

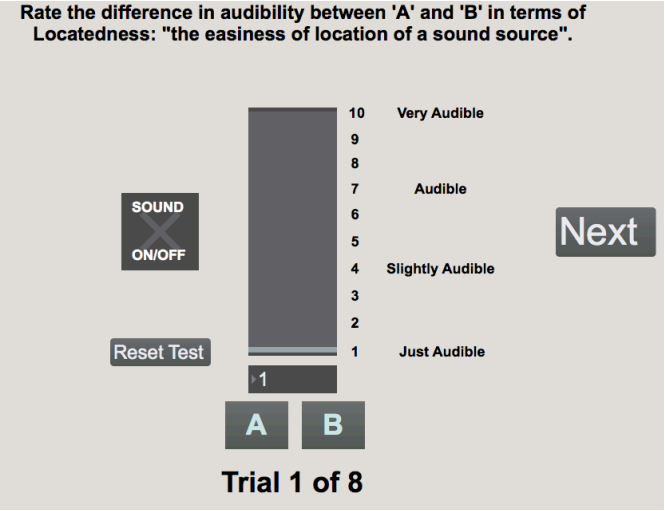

Fig. 5.2: Interface used for the audibility tests (the above focuses on the 'locatedness' attribute).

The attribute that subjects were listening for during each test was displayed at the top of the interface. In each case, subjects were provided with both the name of the attribute (e.g. hardness, brightness, etc.) and a description for it. The descriptions provided for each attribute were the same as those used for the group discussion (Table 5.1). For each test, subjects were required to first determine the attribute that they were listening for and were then encouraged to listen to the first few trials in order to familiarize themselves with that attribute. When this process was complete they were to reset the test and sit it fully. On average subjects required around 5 minutes to complete each test. Subjects completed all 13 tests in two sittings, each of

which lasted no more than 45 minutes. Due to the length of each sitting they were encouraged to take breaks if they so required in order to minimize the effects of fatigue. 12 experienced subjects completed the test, with the test order being randomised for each of them.

In order to determine the overall audibility of each attribute the audibility index was derived. This represents the average audibility for each attribute and was calculated by dividing the sum of the grading values obtained for each attribute by the number of subjects [Lee 2006]. A limitation with considering the average audibility for a given attribute is that inter-stimuli differences are not taken into account. It might be the case, for example, that changes for a particular attribute are highly audible for one stimulus but much lower for another, as was alluded to in Table 5.3. This would therefore mean that the resultant audibility index would be somewhere in the middle of the scale. Because of this, it could be argued that key salient attributes might be missed in the test. Despite this, the audibility index was utilised for the test for the following reason. As was discussed earlier, to the knowledge of the author the most salient perceptual effects of vertical interchannel crosstalk are not yet known. The predominant aims of this experiment are therefore to determine what these effects are and, further, how their audibility varies when the different localisation threshold methods are applied. Although inter-stimulus differences are considered as being an interesting topic of research, it was thought that a more gerneral study into the most salient effects would be more appropriate in light of the aims of the present thesis. Had these effects already been known then a study that analysed how they varied for different test stimuli would have been considered but, as it stands, it was thought that this analysis lay beyond the scope of the research. Therefore, the present experiment aimed to determine the most salient effects of vertical interchannel crosstalk as a whole and this allows for future study on how these effects differ for different test stimuli.

It should also be noted that the audibility index method serves only to determine those attributes of vertical interchannel crosstalk that are the most salient. It will not give information as to the direction in which the attribute is perceived. For example, it may be that changes in source ‘fullness’ are highly audible however it would remain unclear whether the degree of fullness increases or decreases; this issue is resolved in Part

Three. The present methodology simply aimed to filter out those attributes that are barely audible, leaving the most salient for subsequent testing.

5.4.2 Results and Discussions

The results of the audibility-grading test are shown in Fig. 5.3. From the results it is clear that the most audible effects of vertical interchannel crosstalk are variations in source elevation, which had an audibility index of 6.4, VIS (6.4) and loudness (6.3). Further to this, it can be deduced that the source elevation and loudness attributes are increased as a result of vertical interchannel crosstalk, as was discussed in section 5.3.2. However, at present it remains ambiguous as to how the perception of VIS is affected. Hypothetically perception of this attribute should increase as a result of vertical interchannel crosstalk although this has not yet been shown experimentally.

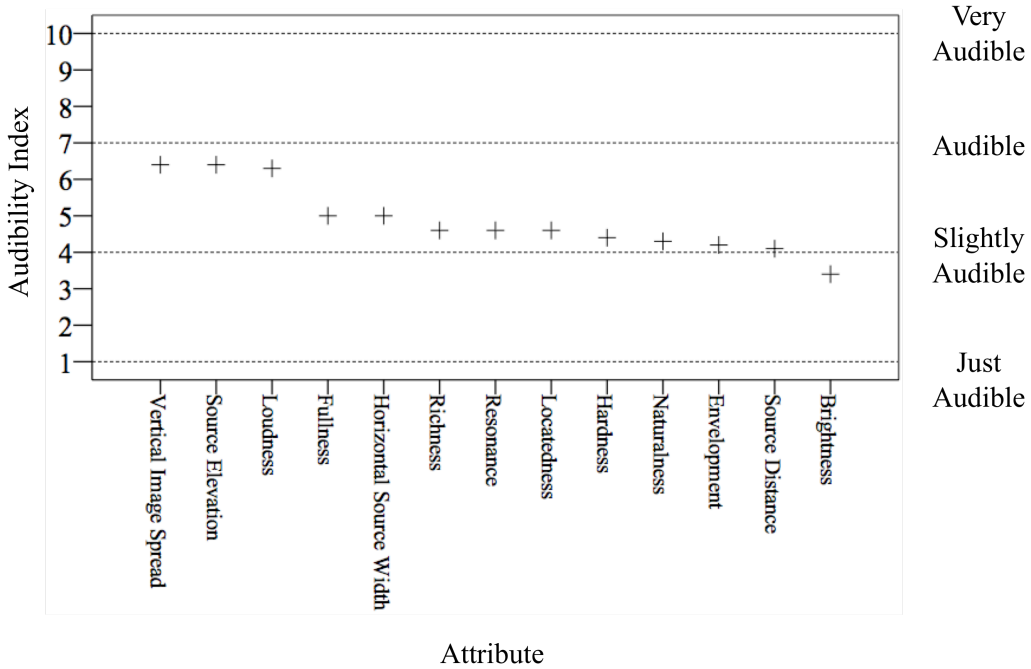

Fig. 5.3: Results of the audibility-grading test.

That loudness, VIS and source elevation changes were found to be the attributes most perceptually affected by vertical interchannel crosstalk agrees with what was hypothesised. The loudness increases, for example, would have been apparent to subjects, as the maximum crosstalk condition was effectively twice the amplitude of the no crosstalk condition. Furthermore, the results of Experiment Five demonstrated that the perceived elevation of the test stimuli was between 5.5° and 11.4° higher for the maximum crosstalk condition, compared to when no crosstalk was present. Again, this would have been fairly easy for the experienced listeners tested in the experiment to identify. Also, as was discussed in Section 5.3.2, audible variations in VIS would have been expected as the maximum crosstalk condition effectively introduced a secondary vertical source. That the amplitude of this secondary source was equal to that of the main channel signal indicates that the change in perceived VIS would have been quite large. Alongside this would have been the difference in amplitude between the two conditions, which has also been shown to affect the perception of VIS [Cabrera and Tiley 2003]. It therefore seems obvious that subjects would have considered variations in perceived VIS as being amongst the most audible of changes.

It is interesting to note that, based on the categorization of the elicited attributes in Part One, the most salient perceptual effects of vertical interchannel crosstalk related to changes in spatial impression, loudness and location. With respect to the timbral attributes it can be seen from Fig. 5.3 that the audibility indexes were much lower. For example, ‘fullness’, being the most audible timbral attribute affected by vertical interchannel crosstalk, was given an audibility index of 5.0. This index was between 1.3 and 1.4 lower than those for the most salient attributes in the other three categories. This result was initially interpreted based on the perceptual effects of comb filtering on the audibility of timbral changes. When a delay is present (i.e. for the 1 ms condition) timbral variations as a result of comb filtering would be fairly audible. Conversely, for the 0 ms condition comb filtering is absent and therefore the audibility of timbral variations would theoretically be much lower. Consequently, as the audibility index considers all scores for all stimuli, it seems reasonable that timbral variations would be rated as being somewhere in the middle of the scale, which was generally the case. The effect of ICTD would have had less of an influence on perceived VIS and source elevation as each of these is predominantly related to ICLD, which did not vary throughout the

experiment. Therefore, the audibility of variations within these attributes would be consistently high throughout the test.

In order to verify the above hypothesis, the audibility indexes for each of the timbral attributes were compared for each ICTD. This can be seen in Table 5.5. From the comparison it is noticeable that, rather than being greater for the 1 ms condition, the audibility indexes were generally consistent. The only exception to this was the 'hardness' attribute, which varied by 0.7 between the two conditions, although in this case the greater index was given to the 0 ms condition. As a result, variations in the audibility of timbral attributes due to changes in ICTD do not explain why the most salient timbral effects were considered as being less audible than the most salient effects within other categories. Further study would be required before the reason for this could be established. However, at this point it is sufficient to conclude that vertical interchannel crosstalk seemingly has less of an influence on the perceived timbre of the main channel signal compared to its effect on perceived loudness, location and spatial impression.

Table 5.5: Comparison of audibility indexes between the 0 and 1 ms conditions.

Attribute	Audibility Index		
	0 ms	1 ms	Variance
Fullness	5.1	4.8	-0.3
Richness	4.6	4.6	0.0
Resonance	4.6	4.7	0.1
Hardness	4.8	4.1	-0.7
Naturalness	4.3	4.2	-0.1
Brightness	3.4	3.5	0.1

In addition, it was initially hypothesised that timbral variations would be more audible for vertical interchannel crosstalk than they were for horizontal interchannel crosstalk. This hypothesis was based on the results of Barron [1971] and Barron and Marshall [1981], which found that timbral coloration as a result of vertical reflections was more audible than for horizontal reflections. Table 5.6 compares the audibility indexes for the timbral attributes in the present study to those obtained by Lee [2006] for horizontal interchannel crosstalk. Initial consideration of the difference in gradings seemingly supports the conclusions

of Barron [1971] and Barron and Marshall [1981]. For all attributes that were perceived for both horizontal interchannel crosstalk and vertical interchannel crosstalk, the higher audibility index was given to the latter effect. However, it should be noted that differences in the test method render such a comparison difficult. In the present study the crosstalk signal was equal in amplitude to the main channel signal. Conversely, in the Lee study [2006] it was between 4 and 20 dB lower, depending on the microphone technique being emulated. As was demonstrated in Chapter Two, the magnitude of timbral variations depends in part on the amplitude of the reflection relative to the direct sound. It therefore stands to reason that the audibility indexes for the timbral stimuli would be greater in the present study compared to in Lee's [2006]. Further study would therefore be necessary to determine how the audibility of timbral changes varies between vertical interchannel crosstalk and horizontal interchannel crosstalk.

Table 5.6: Comparison between the audibility indexes for timbral attributes between the present study and Lee [2006].

Attribute	Audibility Index	
	Present Study	Lee [2006]
Fullness	5.0	3.5
Richness	4.6	X
Resonance	4.6	X
Hardness	4.4	2.3
Naturalness	4.3	1.3
Brightness	3.4	1.4
Phasiness	X	0.5

The primary aim of this part of the experiment was to determine the most salient perceptual effects of vertical interchannel crosstalk. Part of the motivation behind this was to reduce the number of attributes to be tested for the different localisation thresholds (Part Three). However, it can be argued that two of the most salient effects, being source elevation and loudness, do not require further testing. This is partly because it is already known that each of these would increase as a result of vertical interchannel crosstalk. In addition to this, with respect to the source elevation attribute it is clear that there would be no variation when the localisation threshold is applied. This is because the nature of the localisation threshold means that the perceived elevation of the main channel signal is unaffected by the influence from the vertical interchannel

crosstalk signal. Moreover, it is apparent that the difference in perceived loudness would depend on the amplitude of the height layer when the localisation threshold is applied, as was discussed in Section 5.1. Therefore, neither the source location nor loudness attributes were considered for testing in Part Three.

In Part Three then, the attributes considered for testing were VIS and fullness. The former attribute was chosen, as its audibility index was the joint highest out of all the attributes tested in the experiment. Further, although it is hypothesised that this attribute would increase as a result of vertical interchannel crosstalk this is not yet known for certain and needs to be verified in line with the aims of the research. The fullness attribute was chosen for two reasons. Firstly, this was the attribute with the highest audibility index after the loudness, VIS and source location attributes. In addition, as loudness, spatial impression and location effects had already been considered, analysing the timbral effects of vertical interchannel crosstalk varied when the different localisation thresholds were applied was considered as being important.

5.5 PART THREE: GRADING OF THE MOST SALIENT EFFECTS WHEN THE LOCALISATION THRESHOLDS ARE APPLIED

In Part Two, it was determined that the most salient effects of vertical interchannel crosstalk include increases in source elevation and loudness. In addition, audible variations in both VIS and fullness were identified, although at present it is not certain whether such attributes are increased or decreased as a result of the effect. The purposes this part of the present experiment then are twofold. The first is to determine the direction in which changes in both VIS and fullness are perceived. The second is to determine how the perception of both fullness and VIS are affected by the method of applying the localisation threshold to the direct sound in the height layer. This will provide useful information on the audible effects of vertical interchannel crosstalk on the main channel signal, even after the effect on source elevation has been removed.

5.5.1 Test Method

The test setup was identical to that used for Parts One and Two. The experiment was divided into two tests, one focusing on the 'vertical image spread' attribute and the other on 'fullness'. A key consideration for the experiment was the scaling method that was to be used. According to Bech and Zacharov [2006] a pre-existing scale should be used where possible. Furthermore, they asserted that one of the most common scale types used for audio evaluation is category scaling. In category scaling subjects are provided with a fixed number of categories, the intervals between which represent equal sensory differences [Lawless 2013]. For a given stimulus, subjects are required to select the category that best describes the perceived magnitude of the attribute being tested [Bridges and Lisagor 1975]. ITU-T P.800 [ITU 1996] presents a series of different category scaling methods with respect to the grading of audio quality, being absolute category rating (ACR), degradation category rating (DCR) and comparison category rating (CCR), each of which are summarised in Fig. 5.4. As it was necessary that the chosen scale was able to give information as to the direction in which the attribute was perceived (i.e. greater or less than the reference), it would appear that the CCR scale is the only one appropriate for testing, as this is the only one that meets this criterion.

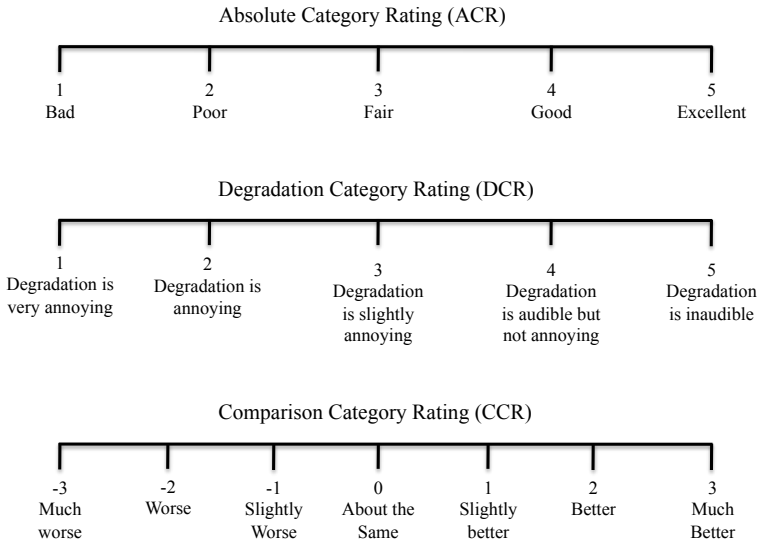

Fig. 5.4: Examples of category rating scales [ITU-R 1996].

A number of drawbacks of the use of category rating scales were discussed by Lodge and Tursky [1979]. They considered primarily that information is often lost due to the limited resolution of each category. For example, if there are minor differences between two stimuli then the scale might force them to be placed in the same category. As such, this differential information becomes lost. This was considered as being problematic with respect to the present study for the following reason. When the localisation thresholds are applied to the height layer, it is hypothesised that the differences in perceived VIS and fullness between the no crosstalk condition and the localisation thresholds would be fairly small. This is due to the amplitude reductions of the crosstalk signal, as a result of the thresholds being applied, which would somewhat limit its perceived interaction with the main channel signal. Therefore, the use of category rating scales was not considered as being appropriate for the present experiment.

An alternative to category rating scales, and one that solves the above issue, is the use of magnitude estimation scales. For such methods, subjects are presented with a series of test stimuli in a random order and are instructed to match numbers proportional to the perceived magnitude of each stimulus relative to a reference [Lodge and Tursky 1979]. Bech and Zacharov [2006] noted that the use of such scales allow subjects to determine numerically whether a given stimulus is perceived as being larger or smaller than the reference with respect to the attribute in question. This was considered as being ideal in light of the aforementioned need to test each attribute in both directions. In addition, the improved resolution of scale allowed for more information to be obtained as to the differences between each localisation threshold method and the no crosstalk condition.

The test interface was created using Max/MSP (Fig. 5.5). The individual fullness and VIS tests were each divided into eight trials, each of which focused on a single sound source (guitar, quartet, speech, conga) with one of the test ICTDs applied (0 and 1 ms). For each trial, subjects were presented with a reference, being the given source presented from the main layer only, and 6 test sounds, being the same source presented as quadraphonic phantom images with both a time delay and one of the six localisation threshold methods

applied to the height layer (blanket, FB, 1+B, 4B, 8B and 0 dB). Subjects were required to compare each of the test sounds to the reference with respect to the attribute being tested.

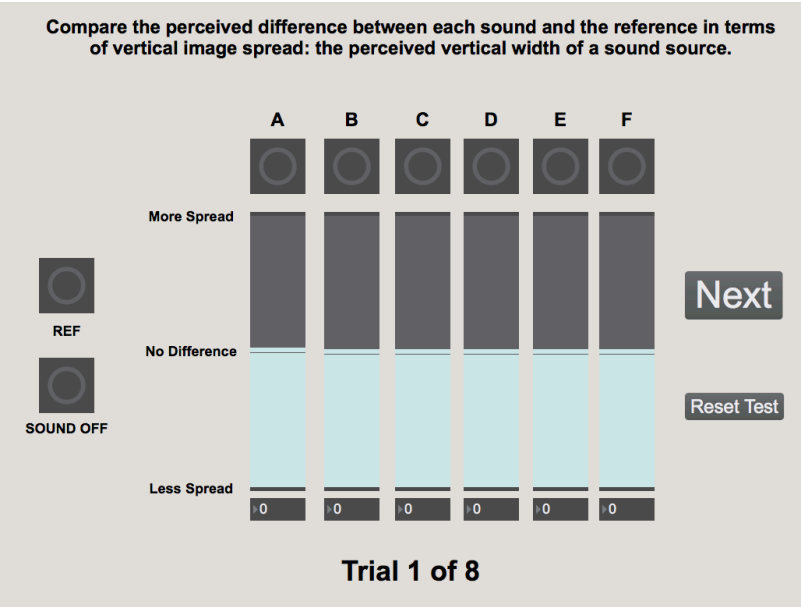

Fig. 5.5: Max/MSP interface used for the grading tests. Interface for VIS test shown.

Judgments were made using a series of sliders that were located underneath each of the test sounds on the interface. According to Waltz et al. [2010] there is generally no restriction as to the range of numbers used on magnitude estimation scales, with the only necessary feature being that an increase in the perceived magnitude of a particular attribute corresponds to an increase in the number assigned to it. As such a 100-point scale was chosen, with values arbitrarily ranging from -50 (lowest possible rating) to 50 (highest possible rating) in increments of 1. A score of 0 therefore related to no perceptual difference between the stimulus and reference. It was considered that the resolution of such a scale would be appropriate in determining what was anticipated to be subtle differences between each of the localisation threshold methods being tested. With respect to the labeling of the scale, it was suggested in ITU-R BS.1284-1 [ITU 2003] that introducing pre-defined points might introduce bias. In order that this could be avoided the labeling on the scale was kept to a minimum. The only labels provided for subjects were ‘No Difference’ at the midpoint of

the scale and 'More…' and 'Less…' at the maximum and minimum respectively. It should be noted however that the use of such a method necessitates that the data be normalized using the formula below [ITU 2003]:

$$Z_i = \frac{(x_i - x_{si})}{s_{si}} \cdot s_s + x_s \qquad (5.1)$$

Where:

Z_i: normalized result.

x_i: score of subject i.

x_{si}: mean score for subject i in session s.

x_s: mean score for all subjects in session s.

s_s: the standard deviation for all subjects in session s.

s_{si}: the standard deviation for subject i in session s.

A further point of note is that caution should be exercised when using the above formula in the present context. The reason for this is that the modification of scale values could result in stimuli that were perceived as being no different from the reference (i.e. those that scored '0' on the scale) having normalized scores that are either positive or negative, thus indicating a perception that was not actually there. In order to prevent this, for each individual subject it was first determined what a score of '0' on the scale would be converted to when normalized. This number was then subtracted from their normalized data to give their final subjective gradings. It was this data that was subsequently used for analysis.

Twelve subjects, being staff and both postgraduate and undergraduate students from the University of Huddersfield's Music Technology department, sat each experiment. The test order was randomised for all subjects. In addition, within each test the order of trials was random, as was the allocation of each localisation threshold method to each slider. Each reference stimulus was presented at 70 dB LAeq, with the

amplitude increase as a result of vertical interchannel crosstalk dependent on the localisation threshold method used.

5.5.2 Data Analysis and Results

Levene and Shapiro-Wilk tests were conducted in order to determine the suitability of the collected data for parametric statistical analysis. The results of the Levene's tests showed that the data was not homogenous. In addition, the results of the Shapiro-Wilk test showed that not all results featured normal distribution. Therefore, as the assumptions of ANOVA were violated, non-parametric tests were used for the statistical analysis.

5.5.2.1 VIS

The Effect of ICTD

Fig. 5.6 shows the effect of ICTD on the perception of VIS for each sound source, with each localisation threshold method applied. The data has been plotted with notch edges. A Wilcoxon signed-rank test was first conducted in order to determine those stimuli whose perceived VIS was significantly greater than that for the main layer only condition. When the ICTD was 0 ms, significant increases were observed for the following stimuli; conga: all methods; guitar: 8B (p = 0.022) and 0 dB (p = 0.003); quartet: blanket (p = 0.05), 4B (p = 0.008) and 0 dB (p = 0.003) and; speech: blanket (p = 0.009), 4B (p = 0.021), 8B (p = 0.021) and 0 dB (p = 0.003). For the 1 ms ICTD condition, the significant increases were for; conga: 0 dB (p = 0.003); guitar: 8B (p = 0.021) and 0 dB (p = 0.006); quartet: 8B (p = 0.05) and 0 dB (p = 0.004) and; speech: 0 dB (p = 0.003). These results generally agree with the notch edges. It can be seen then that the perception of VIS increases as a result of vertical interchannel crosstalk. Further, significant increases in VIS are still apparent for some of the localisation threshold methods, most notably for the 4B and 8B conditions, although this is dependent both on the ICTD and the source.

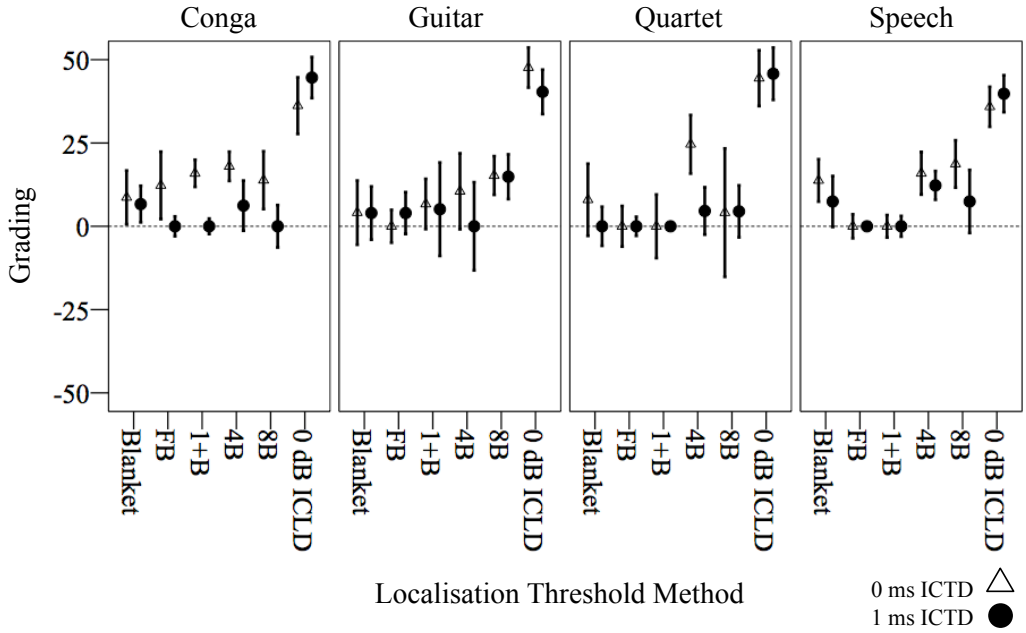

Fig. 5.6: Medians and associated notch edges showing the effect of ICTD on the perception of VIS.

With respect to the effect of ICTD, the following can be inferred from the notch edges. Firstly, there appears to have been no significant effect for either the speech or guitar sources, as there is notable overlap of the notch edges for all cases. Conversely, there appears to have been a significant difference for the quartet source when the 4B threshold method was used. In addition, the effect appears to be significant for the conga for both the FB, 1+B and 4B conditions. A series of Wilcoxon tests were conducted in order to verify this. The results showed no significantly different pairs for the speech source, whilst for guitar the only significant difference was for 0 dB ($p = 0.022$). However, given that there is a notable overlap between notch edges for the 0 dB guitar and also that the effect size was not large ($r = 0.49$), it can be concluded that the effect of ICTD was not significant in this case. For the conga, significant differences were identified for 4B ($p = 0.028$), FB ($p = 0.008$) and 1+B ($p = 0.015$). This generally agrees with the notch edges, although there is some small overlap for FB. However, the effect size, r, indicated a large effect (0.52). Therefore, it can be concluded that the effect of ICTD was significant for this threshold method. The results for the quartet agreed with the notch edges in that the only significant effect was for the 4B condition ($p = 0.028$).

The Effect of Sound Source

Fig. 5.7 shows the effect of sound source on the perception of VIS for each localisation threshold method. The median scores have been plotted with notch edges. For the 0 ms condition, the effect of sound source appears to have been significant for 1+B (conga and both quartet and speech), whilst the overlap between the conga and all other sources for the FB condition is notably small. A Friedman test conducted on the data showed that the effect of sound source was significant for these two localisation threshold methods (p = 0.005 for FB, p = 0.03 for 1+B). For all other methods the effect of sound source was not significant. Wilcoxon tests were conducted for each of these threshold methods in order to find the significantly different pairs. The Bonferroni correction was used to reduce Type-I errors. For the FB condition, a significant difference was identified between the conga and speech (p = 0.048), which agrees with the notch edges. In addition, there were no significant differences between the conga and either the quartet (p = 0.834) or guitar (p = 0.102). This analysis agrees with the notch edges, although it should be noted that the overlap between the conga and the quartet and guitar is minimal. For the 1+B condition the Wilcoxon test results identified a significant difference between the conga and both the quartet (p = 0.03) and guitar (p = 0.048). Based on the notch edges however there is clearly significant difference between the conga and speech. This appears to be a Type-II error, with the unadjusted data showing a significant difference (p = 0.013). Also, although there was significant difference between the guitar and the conga according to the Wilcoxon test, the overlap between notch edges and medium effect size (r = 0.38) suggests that the difference was not significant. For the 1 ms ICTD the notch edges all overlap, with the exception of the guitar and conga for the 8B condition, indicating that, generally, the effect of sound source was not significant. This was confirmed with the results of a Friedman test. Overall, for 0 ms the effect of sound source was only significant for the 1+B and FB methods. For 1 ms there was no significant effect.

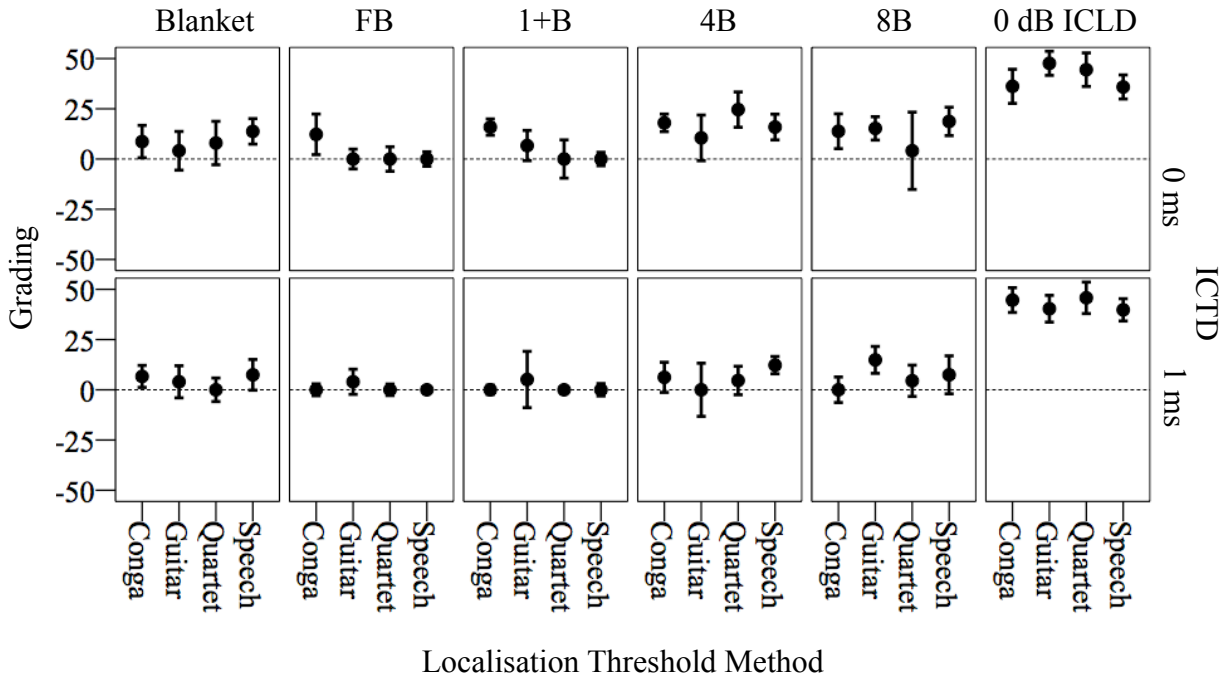

Fig. 5.7: Medians and associated notch edges showing the effect of sound source on the perception of VIS.

The Effect of Localisation Threshold Method

Fig. 5.8 shows the effect of localisation threshold method on the perception of VIS. The medians have been plotted with notch edges. Prior to any analysis, it would appear that the 0 dB condition resulted in a significantly greater perception of VIS for all sources at both ICTDs. Friedman tests conducted on each source at each ICTD suggested that the effect of localisation threshold method was significant ($p < 0.01$ for all). A series of Wilcoxon tests were conducted to analyse the significantly different pairs, with the Bonferroni correction being applied. This analysis agreed that the 0 dB condition resulted in a significantly higher perception of VIS compared to all the other methods for all the stimuli tested.

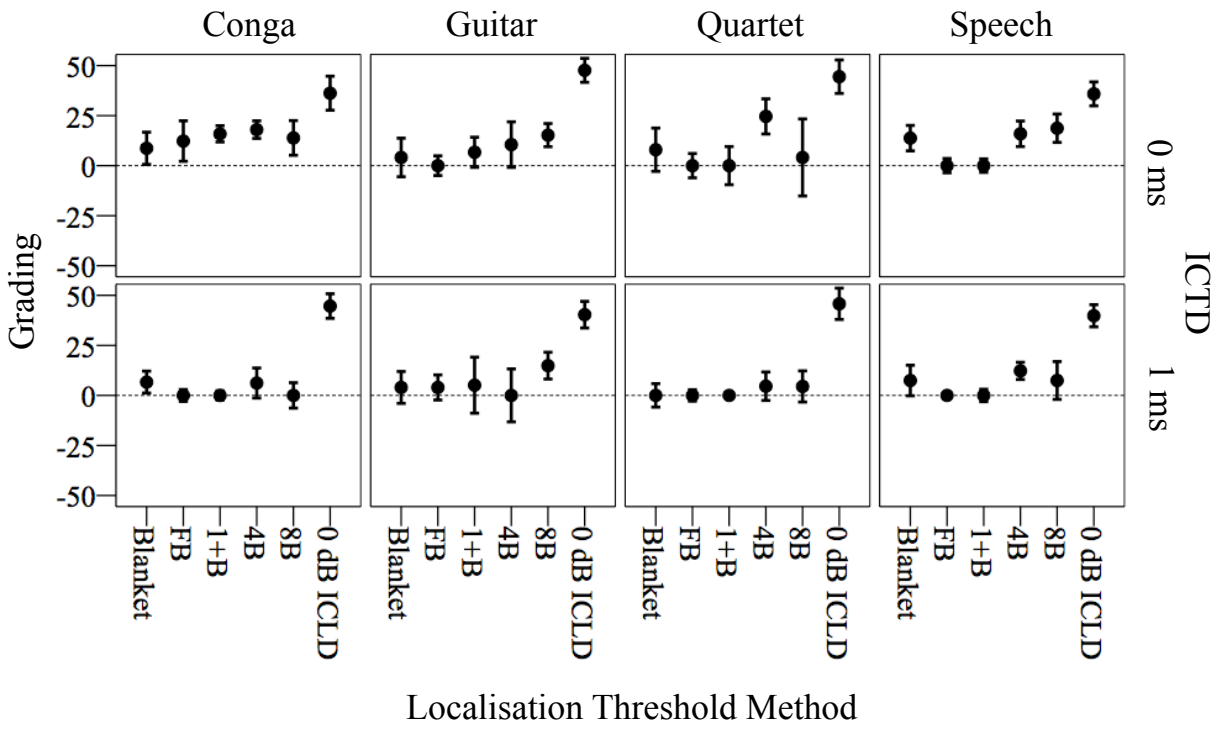

Fig. 5.8: Medians and associated notch edges showing the effect of localisation threshold method on the perception of VIS.

With respect to differences between the localisation threshold methods, the Wilcoxon test results indicated the following for the 0 ms ICTD condition. For the conga, there were no significantly different pairs, which agrees with the notch edges. For both the speech and quartet sources no significantly different pairs were identified, however this may be due to Type-II errors from the use of the Bonferroni correction. For the quartet, for example, the grading for 4B looks to be significantly higher than that for the 1+B and FB methods. When the Bonferroni correction was removed, and the unadjusted results considered, the Wilcoxon test results matched what is shown by the notch edges. Equally, for the speech source the unadjusted data showed that the gradings for FB and 1+B were significantly lower than those for blanket, 4B and 8B, which agrees with the notch edges. Interestingly, the results for the guitar show no significantly different pairs, even when the Bonferroni correction was removed. However, there is clearly no overlap between 8B and FB.

This may therefore be an example of a Type-I error, caused by the number of comparisons for the Wilcoxon test.

For 1 ms, the results for the conga were the same as for 0 ms in that there were no significant differences between the threshold methods. This result was also found for the other three sources. For the guitar and quartet sources this analysis matches what can be seen with the notch edges, although it is clear from Fig 5.8 that the overlap between FB and 8B is small for the guitar. For the speech source the interaction between FB and 4B was found to be non-signifcant (p = 1.000), as was that between 1+B and 4B (p = 1.000) despite the clear lack of overlap between notches. This result therefore appears to be a Type-I error, especially as the p value for the unadjusted data for the interaction between FB and 4B was 0.075. It can therefore be concluded that the effect of localisation threshold method was significant for the speech source with 1 ms ICTD between 4B and both 1+B and FB. From this analysis it can therefore be concluded that the VIS gradings were significantly affected by the localisation threshold method for both ICTDs for all sources, with the exception of the conga when the ICTD was 0 ms, whilst significant differences were also found for the speech at 1 ms. In addition, when there were significant differences the general trend was the 4B and 8B were significantly higher than FB and 1+B.

5.5.2.2 Fullness

The Effect of ICTD

Fig. 5.9 shows the effect of ICTD on the perception of fullness for each source, with each localisation threshold method applied. A Wilcoxon signed-rank test was first conducted to analyse the difference in perceived fullness between each stimulus and the main layer only condition. For both ICTDs, the 0 dB condition resulted in significant increases in perceived fullness. Further, for 0 ms, neither the blanket reduction nor the FB technique resulted in a significant increase in perceived fullness for any source. On the

other hand, significant increases in fullness were identified for the 4B condition for all sources, whilst 8B was significant for all sources except the speech (p = 0.093). Also, significant increases for the 1+B condition were observed for the guitar (p = 0.012) and speech (p = 0.018) sources. For 1 ms, there were notably fewer significant increases. For the conga only the 4B condition was significant (p = 0.012), whilst for the speech only 1+B (p = 0.012) and 8B (p = 0.028) were significant. There were no significant increases for the quartet or guitar. Overall this analysis shows good agreement with the notch edges.

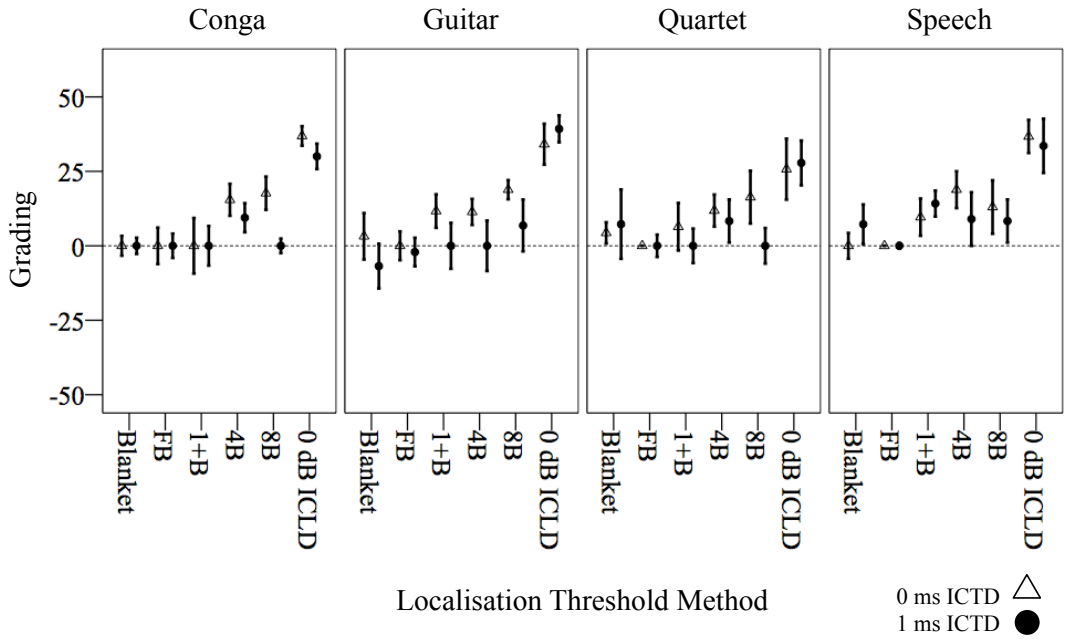

Fig. 5.9: Medians and associated notch edges showing the effect of ICTD on perceived Fullness.

With respect to the effect of ICTD, the notch edges indicate that the fullness gradings were significantly affected for the conga and quartet (both 8B). A series of Wilcoxon tests were conducted in order to verify this. The results for the quartet matched the notch edges, with there being a significant difference for the 8B condition (p = 0.047). Interestingly, the Wilcoxon test also suggested that the effect of ICTD was significant for the 4B condition for the guitar (p = 0.033) however in this case there is clear overlap between the notch edges. Addiitonally, ICTD did not have a significant effect at 8B (p = 0.091) for this source despite the

minimal overlap between notch edges. For the conga, significant differences were identified both for 8B (p = 0.007), which agrees with the notch edges, and also for 0 dB (p = 0.013), which does not. The result for 0 dB is arguably a Type-I error, with there being a notable overlap of notch edges, indicating no significant difference. Additionally, significant differences were identified for the speech source for the 4B condition (p = 0.021). It should be noted that this does not agree with the notch edges. In addition, the effect size, r, did not indicate a large effect (0.49). Therefore, the effect of ICTD was not significant for the speech source. Overall then, ICTD had a significant effect on the perception of fullness for the conga and quartet sources (both 8B).

The Effect of Sound Source

The effect of sound source on the perception of fullness for each localisation threshold method at each ICTD is shown in Fig. 5.10. The medians have been plotted with notch edges. It can be seen for 0 ms that there is notable overlap of all notch edges, indicating no significant effect. This was confirmed with the results of a Friedman test ($p > 0.05$ for all). Further, for the 1 ms condition the only significant difference appears to be between the speech and all other sources for the 1+B condition. However, the results of a Friedman test suggested that there was no significant effect of sound source ($p > 0.1$ for all). Despite this result, a Wilcoxon test was conducted for 1+B to further analyse the difference between the speech source and the conga, quartet and guitar. The result showed that the differences in each case were not significant (p = 0.131 for conga, p = 0.059 for guitar and p = 0.066 for quartet). In addition to this, the effect size, r, indicated a medium effect for all sources ($p > 0.4$ for all), whilst Kendall's W was also low (0.01). However, despite these results it is clear that there is no overlap between the notch edges, which, according to McGill et al. [1978] and Kirchner [2001] indicates that pairs of stimuli are significantly different from each other with 95% confidence. It can therefore be concluded that the effect of sound source was significant for the 1+B condition with 1 ms ICTD, with the difference coming between the guitar and all other sources. For all other conditions the effect was not significant.

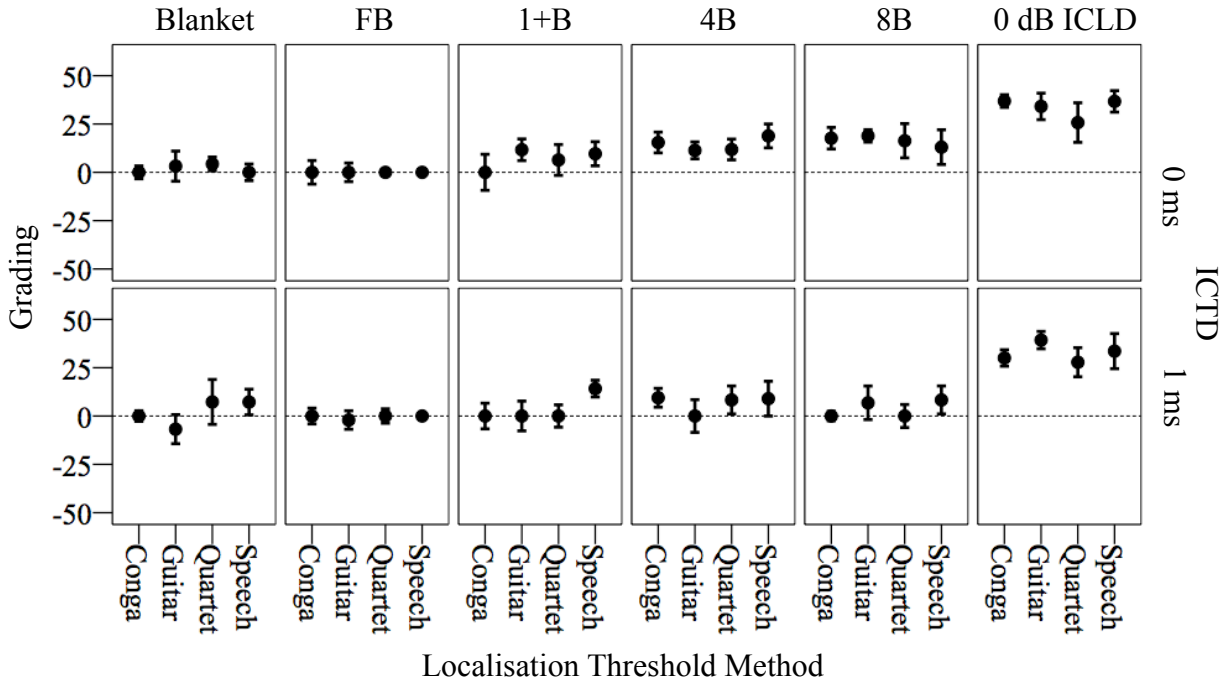

Fig. 5.10: Medians and associated notch edges showing the effect of sound source on the perception of fullness for each localisation threshold method.

The Effect of Localisation Threshold Method

Fig. 5.11 shows the effect of localisation threshold method on the perception of fullness for each source at each ICTD. The medians have been plotted with notch edges. From the notch edges it would appear that the 0 dB condition resulted in a significant increase in perceived fullness with respect to the localisation threshold methods. The only result that does not appear to conform to this is the 0 ms quartet, where there is notable overlap between the 0 dB and 8B conditions. A Friedman test conducted separately for each ICTD showed that the effect of localisation threshold method was significant for all sources, with Wilcoxon tests showing that in all cases the 0 dB condition was graded significantly higher than each of the other localisation threshold methods. The only exception to this was the quartet at 0 ms, with the Wilcoxon test results showing that the interaction between 0 dB and 8B was not significant (p = 0.062). This agrees with the notch edges.

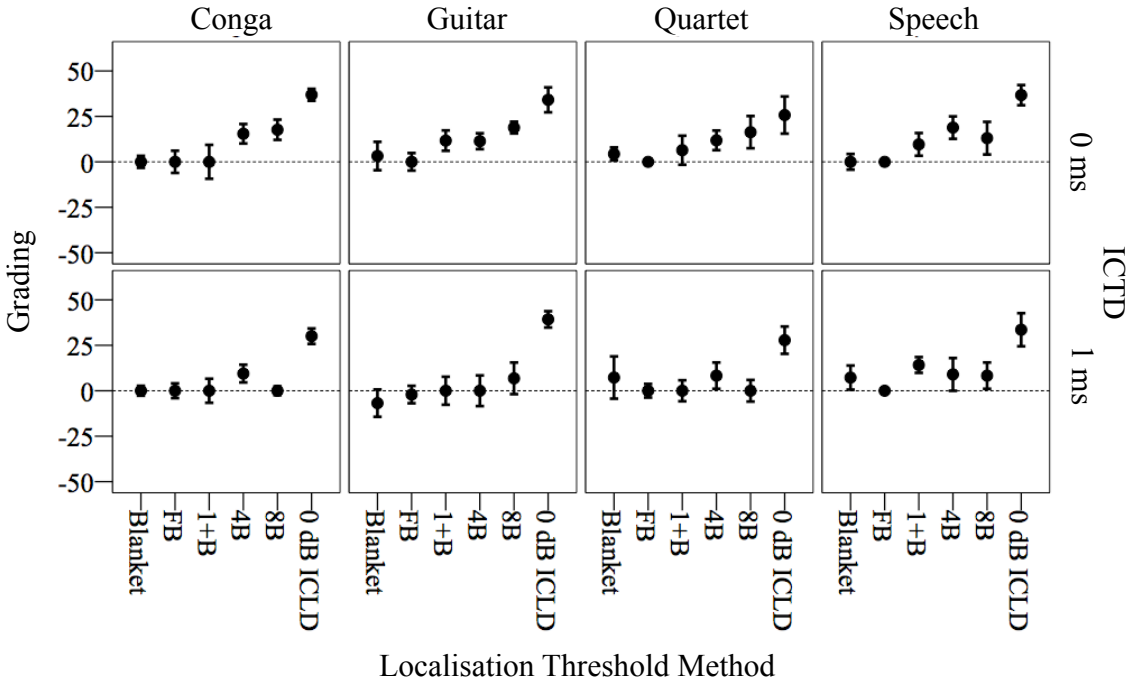

Fig. 5.11: Medians and associated notch edges showing the effect of localisation threshold method on perceived fullness.

It should be noted that, with respect to the Wilcoxon results for the other localisation threshold methods, the use of the Bonferroni correction resulted in a high number of apparent Type-II errors. As a result, the analysis considered the data when the correction was not applied. For 0 ms, significantly different pairs were identified for all sources. For the conga, guitar and quartet, significant differences were identified between FB and 8B (p = 0.026, 0.028 and 0.037 respectively). Also, for the speech the grading for 4B was significantly higher than those for FB (p = 0.005) and blanket (p = 0.007). Each of these results shows good agreement with the notch edges. It should be noted however that Fig. 5.11 indicates further significant differences, including between 4B and both FB and blanket for the conga and between 4B and FB for the guitar. Nevertheless, it is clear that, in general, the median perceived fullness was greater for 4B and 8B than for blanket and FB and sometimes this difference was significant.

For the 1 ms condition, the notch edges suggest that there was no significant effect for the guitar or quartet. For the speech, 1+B and FB appear significantly different from one another, whilst 4B and FB appear

significantly different for the conga. These results were generally confirmed with the Wilcoxon test. For the conga, the FB and 4B conditions were significantly different from one another (p = 0.005), whilst no significantly different pairs were identified for either the guitar or quartet. For the speech, the results suggested that there were no significantly different pairs. This could be a Type-I error, as there is clearly no overlap between the notch edges 1+B and FB. Therefore, localisation threshold method can be concluded to have had an effect on the speech source.

5.5.3 Discussion

A primary aim of the present thesis was to determine the most salient perceptual effects of vertical interchannel crosstalk. As was discussed earlier, two of the more apparent effects are increases in both source elevation and loudness. The results of the present experiment are able to build on this by showing that the perception of both VIS and fullness also increase. For both attributes, the increase in their perception when source presentation changed from main layer only to vertically oriented phantom image with 0 dB ICLD applied was significant for all the stimuli tested. Further, when the localisation threshold was applied to the height layer, some significant increases in both VIS and fullness were identified with respect to the main layer only condition, although this was somewhat dependent both on the ICTD and on the localisation threshold method. Based on this, the null hypothesis that vertical interchannel crosstalk does not have an audible effect on the main channel signal can be rejected. Equally, it is also possible to reject the null hypothesis that the perceptual effects of vertical interchannel crosstalk are the same for all localisation threshold methods. These results agree with Lee [2011], in that vertical interchannel crosstalk would have an audible influence on the main channel signal, even when the localisation threshold is applied.

5.5.3.1 VIS

That the perception of VIS increased as a result of vertical interchannel crosstalk was somewhat expected given both the literature and the informal observations following Experiment One. It is interesting to note, however, that variations in perceived VIS were not consistent for the different localisation threshold methods tested. This result shows partial agreement with what was hypothesised. Despite this, it was initially thought that, excluding the 0 dB condition, the greatest increases in perceived VIS would be observed for the 8B and 4B conditions, as these required the least attenuation of the direct sound in the height layer. However, rather than this being the case, it can be seen from the results that the median gradings for the 8B and 4B methods were not consistently higher than those for FB and blanket. This result would seemingly indicate that the perception of VIS was not solely related to increased amplitude of the direct sound in the height layer with respect to that in the main layer. Despite this, as there appears to be no consistent pattern with the results, further study would be needed to ascertain the reasons for the present data. What can at least be said is that there was generally an increase in perceived VIS for the majority of localisation threshold methods tested with respect to the main layer only condition.

With respect to ICTD, it can be seen from the results that there was little effect for the guitar, speech and quartet sources. Conversely, gradings for the conga were significantly higher for the 0 ms ICTD for FB, 1+B and 4B, compared to those for 1 ms. Additionally, although not significant, the median gradings for 8B and blanket were also higher for the 0 ms ICTD. That the presence of small ICTDs can reduce the perception of VIS for a conga source has previously been demonstrated by Tregonning and Martin [2015]. In that study, conga and female speech sources were presented to subjects from the frontal channels of an Auro 9.1 configuration. The main layer consisted of C, L and R loudspeakers, whilst the height layer consisted of HL and HR loudspeakers. The test stimuli were presented simultaneously from both layers, with an ICTD applied to the height layer with respect to the main. The results showed that, for the conga source, an ICTD of 5 ms resulted in a narrower perceived VIS compared to when the ICTD was 0 ms. This somewhat mirrors

the results observed in the present study, although in this case the ICTD was notably smaller and no centre channel was present in the main layer.

A potential reason for a reduction in the perceived VIS of the conga in the presence of an ICTD, as observed both in the present study and by Tregonning and Martin [2015] is as follows. Informal listening conducted by the author showed that, when presented as a vertically oriented phantom image, the perception of VIS is related to the pitch height effect. In other words, the higher frequencies within the source are perceived as being physically higher in space than are the lower frequencies, which gives rise to the perception of a vertically spread image. Incidentally, a similar suggestion was made following a study conducted by Sundaram and Kyriakakis [2005]. In addition, as was discussed previously, the spectral energy for the conga predominantly lies in the 0.6-2 kHz range, with notable peaks around 0.7 and 1.5 kHz. When the pattern of comb filtering for a 1 ms ICTD is considered (Fig. 5.12) it can be seen that there are notable cuts around 600 Hz and 1.8 kHz. If consideration is again given to how VIS is perceived for a vertically oriented phantom image the following can be suggested. The presence of comb filtering resulted in attenuation at the extremes of the frequency range at which the energy for the conga is predominantly concentrated. Therefore, based on the pitch height effect, the simultaneous attenuation of both highest and lowest frequencies within the signal would inevitably make the source appear much narrower vertically. That such a result was only observed for the conga source in the present study might be related to the audibility of comb filtering for the other sources. The transient characteristics and relatively narrow bandwidth of the conga would arguably make comb filtering more audible compared to the speech, guitar and quartet sources, which each occupied a greater bandwidth and were all more continuous in nature. This, however, would have to be studied further. It should also be noted that the present hypothesis does not explain why the effect of ICTD on the perception of VIS was significant for the conga for FB, 1+B and 4B but not 8B and blanket.

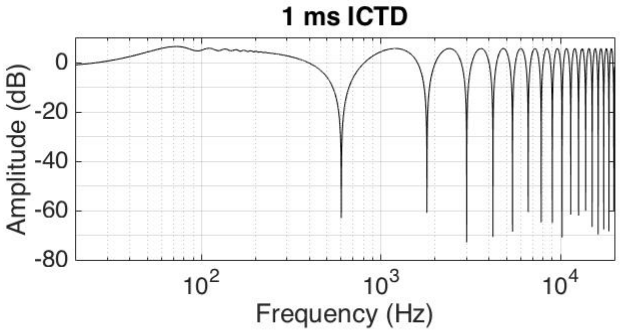

Fig. 5.12: Comb Filtering Pattern for 1 ms ICTD.

5.5.3.2 Fullness

The increase in perceived fullness as a result of vertical interchannel crosstalk can be approached both from the perspective of loudness and from the perspective of increased levels of low frequencies in the height channels. It can be seen from the results that, when the ICTD was 0 ms, a general pattern emerged for the median perceived fullness for each localisation threshold method. The grading for FB was usually the lowest, increasing progressively for 1+B, 4B and 8B. It should be noted that this was not maintained for the speech source, with the median grading for the 8B method being lower than 4B and equal to 1+B. It should also be noted that the differences here were not always significant. What is indicated, however, is that the perception of fullness increased with increases in the amplitude of the signal in the height layer when the ICTD was 0 ms. This result was somewhat expected, and indeed was hypothesised, as an increase in amplitude would inevitably make the low frequencies within each source more audible, which would increase perceived fullness.

However, as already mentioned, increases in perceived loudness are not the only way in which the results can be explained. It is also apparent that the median gradings for 0 ms generally coincided with less attenuation of the low frequencies in the height layer. For the FB condition, for example, all the low frequency bands were attenuated somewhat, whilst for the other band reduction conditions the 63, 125, 250

and 500 Hz bands were all untouched. It would appear then that the presence of more low frequencies contributes to the perception of increased fullness. This seems logical given that the resultant attenuation in the height layer was somewhat similar for the 1+B and blanket conditions and yet the median grading for the former condition was higher for all four sources (although this difference was not significant). It should however be noted that, because increases in low frequency content in the height layer naturally coincided with an increase in amplitude, it is not possible to say with certainty the exact cause of the perception of increased fullness observed in the present study.

It is also interesting to note that the aforementioned trend for 0 ms was not observed for 1 ms. Instead, the balance of gradings appeared to be somewhat more random. In addition to this point, there were also fewer significantly different increases in perceived fullness with respect to the main layer only condition for the 1 ms ICTD compared to 0 ms. It would appear then that the presence of an ICTD caused both a reduction in overall fullness (although this was not necessarily significant) and a breakdown in the relationship between both amplitude and low frequency content and the perception of increased fullness. It would seem logical that this result is related to the perceptual effects of comb filtering. As can be seen in Fig. 5.12, a 1 ms ICTD results in a notable notch in the spectrum in the range of 400-800 Hz. Any attenuation in the region would be sufficient to reduce the perceived fullness of the main channel signal, which agrees with the results obtained. However, it should be noted that this would require further study.

5.5.3.3 The Relationship Between Perceived Fullness and VIS

As the perception of both VIS and Fullness generally increased as a result of vertical interchannel crosstalk, it was decided to conduct an analysis into the strength of the relationship between the two attributes. Fig. 5.13 shows a scatter plot of the gradings that each attribute was given in the test. From the line of best fit, it would appear that the relationship between the two attributes was positive. In order to test this further a bivariate correlation test was conducted. The results of this analysis showed that the interaction between

perceived VIS and fullness was significant (p = 0.000). Further, the results of a Pearson's Correlation test showed that this interaction was both positive and had a large effect (r = 0.54). It should be noted that this analysis showed good agreement with the scatter plot shown in Fig. 5.13.

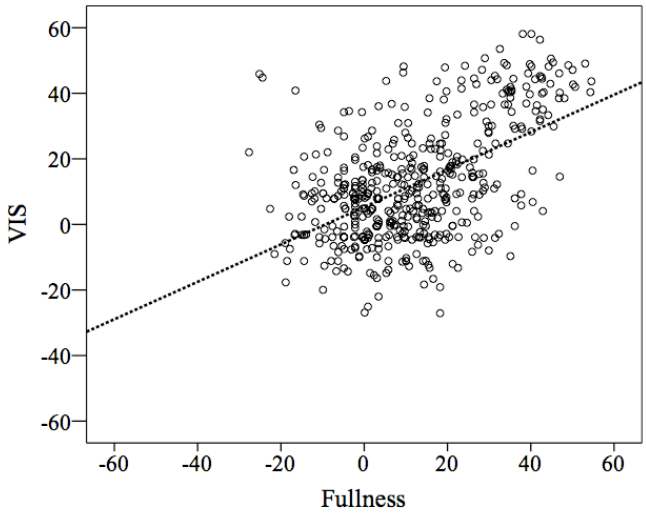

Fig. 5.13: Scatter plot showing the relationship between the gradings for perceived VIS and perceived fullness.

It would appear then that an increase in the perception of VIS corresponded with an increase in the perception of fullness. One of the more apparent reasons for this correlation is the effect of loudness. As has already been discussed, an increase in the amplitude of the signal in the height layer would increase the perception of VIS, whilst also making sources appear more full due to increased audibility of the source's low-frequency energy. However, it should also be noted that the perception of each attribute did not necessarily increase with perceived loudness, most notably in the presence of a 1 ms ICTD in the height layer. Therefore, the effects of comb filtering may also have had some influence, with attenuations in the spectrum in the low frequency range making sources appear thinner, whilst also decreasing the perception of VIS based on the pitch-height effect. It is clear, however, that the relationship between VIS and fullness needs to be studied further, most notably in the absence of the loudness differences caused as a result of the differing localisation threshold methods applied to the height layer.

5.6 PART FOUR: THE SUBJECTIVE PREFERENCE OF LOCALISATION THRESHOLDS

From Parts One-Three of the present experiment, it is known that the most salient perceptual effects of vertical interchannel crosstalk are increases in source elevation, loudness, VIS and fullness. Further, with respect to the latter two attributes, it is also known how their perception is influenced by the different methods of applying the localisation threshold derived in Chapter Four. The aim of the final part of the present experiment is to determine which of the derived methods are the most preferred by subjects and why.

5.6.1 Test Method

Within the literature, there are generally two broad approaches to preference testing for audio. The first of these involves pairwise comparisons between all of the stimuli under test, where, for each trial, subjects indicate which of the two stimuli is more preferred. This method has been used in numerous studies including those of Martens et al. [2006], Trevo et al. [2014] and Fazenda et al. [2015]. Arguably, the primary benefit of the pairwise comparison method when conducting preference tests is its simplicity for subjects. However, there is a notable drawback with respect to the analysis of data. As the data obtained is from a series of A/B comparisons, the results in their raw state are not appropriate for statistical analysis. Instead, it is first necessary to process the data. One of the key ways this can be achieved is with the 'OptiPt' Matlab function developed by Wickelmaier and Schmid [2004], which was the approach taken by Martens et al. [2006]. An alternative method was used by Trevo et al. [2014], who developed a matrix based on how many times a given stimulus was preferred with respect to another.

The second approach makes use of magnitude estimation scaling, which is described in detail in Section 5.5.1. In experiments conducted in order to determine the subjective preference of different loudspeakers,

Olive [2003] provided subjects with an 11-point scale, with values ranging from 0-10, which also included pre-defined anchor points. Subjects were required to directly compare all stimuli when making their preference judgments, with strict instructions on how this should be undertaken. An alternative, and more minimalist, form of magnitude estimation scaling was used by Manor et al. [2015]. Their scale did not include any anchors or numbers, with the only markers on the scale being a '+' at the top of the scale, to indicate 'more preferred', and a '-' at the bottom to indicate 'less preferred'. A further marker was included at the midpoint to represent 'no preference'. The numerical range of the scale was also different, being -1.5 to 1.5. The use of magnitude estimation scales when conducting preference testing is beneficial for a number of reasons. Firstly, such scales are able to show the relative differences in preference between all stimuli. In addition, the data obtained is immediately suitable for statistical analysis without the often-complicated methods of conversion discussed above, which is not the case for the pairwise comparison method.

For the present study then, it was decided to use the magnitude estimation method. This was partly motivated by the increased simplicity with respect to the analysis of data. In addition, this method enabled subjects to compare each of the localisation threshold methods to each other simultaneously, which would not have been possible with pairwise comparisons. A further reason the decision was made related to the overall duration of the respective methods. As there were 48 stimuli, being four sound sources, two ICTDs and six localisation threshold conditions, a pairwise comparison approach would have required around 120 trials (15 comparisons for each sound source with a single ICTD applied). This would have necessitated two sittings per subject to avoid any effects of fatigue. Conversely the magnitude estimation method enabled the test to be conducted in a single sitting consisting of eight trials.

It has already been noted on multiple occasions during the present experiment the potential effects that differences in perceived loudness might have on the results obtained. For the preference test this issue is particularly salient, as subjective preference is often influenced heavily by perceived loudness (i.e. the 'louder is better' principle). Because of this, it was decided to conduct the preference tests in two separate and smaller tests. For the first test the different threshold methods were not level matched. As with the

previous experiments, the main layer only condition was presented at 70 dB LAeq, with the increase in amplitude for the other stimuli dependent on the threshold method used. For the second test, all conditions were matched to 70 dB LAeq. Other than this difference the two tests were identical.

The test interface was created using Max/MSP (Fig. 5.14). The test consisted of eight trials, each one being one of the sound sources (guitar, speech, conga and quartet) with one ICTD (0 and 1 ms) applied to the height layer. For each trial, subjects were presented with six stimuli, one being the main layer only condition and the other five being vertical phantom images with a localisation threshold method (blanket, FB, 1+B, 4B, 8B) applied to the height layer. A slider accompanied each of these stimuli, with values ranging from -50 to 50 in increments of 1. A positive value was to be attributed to a given sound source in the case that it was preferred, whilst a negative value was attributed to disliked sounds. A score of '0' indicated no preference. Subjects were instructed to give a progressively higher rating to a given stimulus the more it was preferred and vice versa. The only labels that appeared on the scale were 'more preferred' at the top (50), 'less preferred' at the bottom (-50) and 'no preference' in the middle (0). As a result of this minimal use of labels, it was necessary to normalise the data as described in Section 5.5.1.

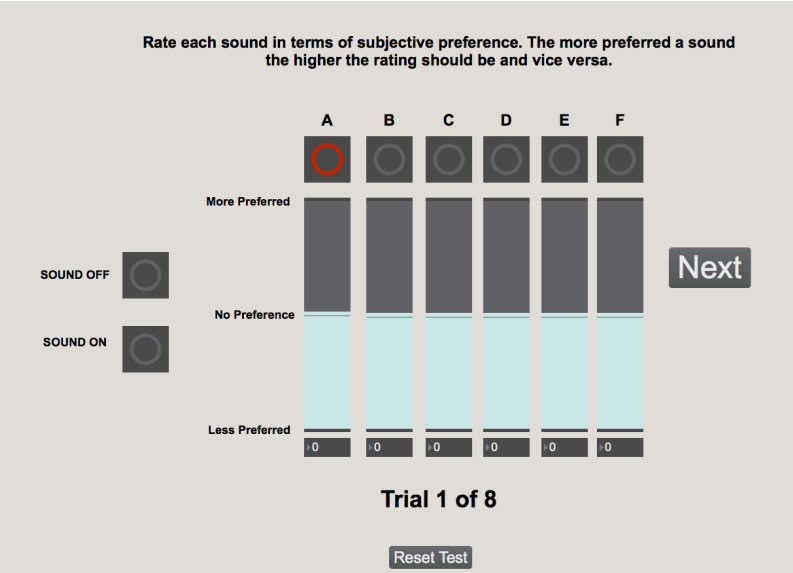

Fig. 5.14: Max/MSP interface used for preference tests.

Alongside the above task, subjects were also asked for each trial to give feedback on the reasons why they liked the highest graded stimulus and also why they disliked the lowest graded stimulus. In order to help with this they were provided with the list of terms and descriptions elicited from the group discussion in Part One. Subjects were instructed to use these terms where possible, however were free to choose other attributes if the provided list did not fit their perceptions. Ten experienced subjects participated in the experiment. The test order was randomised for each subject, as were the order of stimuli within each test.

5.6.2 Data Analysis and Results

Levene and Shapiro-Wilk tests were first conducted on the obtained data in order to determine its suitability for parametric statistical analysis. The results of the Levene tests showed homogeneity of variance, whilst the Shapiro-Wilk test showed that not all scores for each condition featured normal distribution. Therefore, as the assumptions of ANOVA were violated, non-parametric methods were used for the statistical analysis.

5.6.2.1 The Effect of Loudness

Fig. 5.15 shows the median preference gradings for each of the two tests (normal and level matched). The medians have been plotted with notch edges. Consideration of the notch edges alone indicates that the effect of loudness was generally not significant on the gradings given by subjects. The only apparent exception to this looks to be when the speech source was presented from the main layer only with 0 ms ICTD. In that case there is no overlap between notch edges, suggesting statistical significance. This was confirmed with the results of a Wilcoxon test (p = 0.037). According to this analysis there were no other significantly different pairs, which agrees with the notch edges.

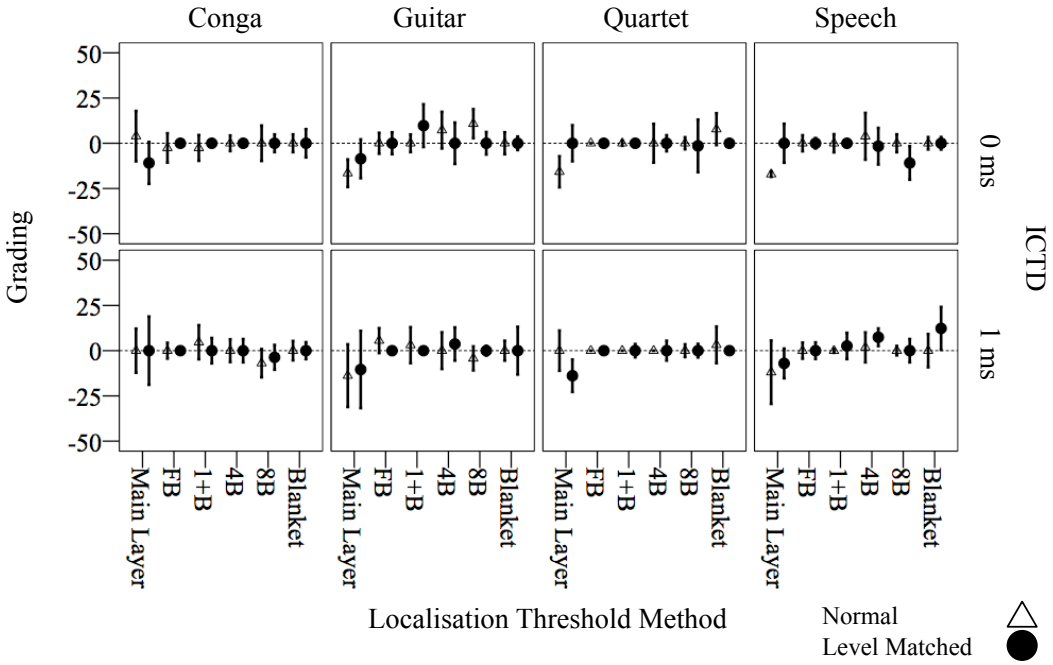

Fig. 5.15: Medians and associated notch edges showing the effect of loudness on subjective preference.

5.6.2.2 The Effect of ICTD

Fig. 5.16 shows the effect of ICTD on the median preference gradings. The medians have been plotted with notch edges. It should be noted that, as the effect of loudness was only significant for one pair of stimuli, the results for each of the two tests have been amalgamated. It can be seen that the notch edges for all cases overlap, which suggests that ICTD did not have a significant effect on the preference gradings. This was confirmed with the results of a series of Wilcoxon tests, which also showed that the effect size, r, was less than 0.3 for all pairs, suggesting a small effect.

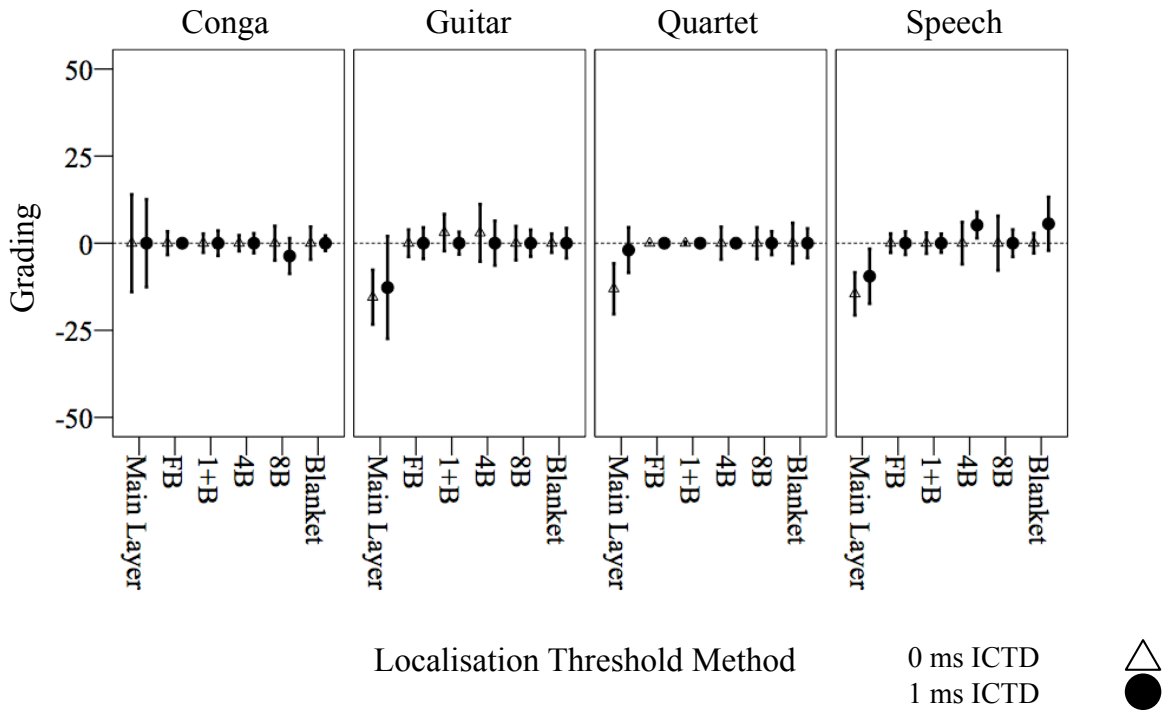

Fig. 5.16: Medians and associated notch edges showing the effect of ICTD on subjective preference.

5.6.2.3 The Effect of Sound Source

The effect of sound source is shown in Fig. 5.17. The medians have been plotted with notch edges. As the effect of ICTD was found to not be significant, the results for each ICTD have been amalgamated. From the notch edges, it would appear that the effect of sound source was not significant for any of the localisation threshold methods. However, for the main layer only condition it is noticeable that the overlap between the notch edges for the conga and both the speech and guitar is small. Despite this, a Friedman test conducted on the data showed that the effect of sound source was not significant for any case ($p > 0.05$ for all). A Wilcoxon test was conducted on the results for the main layer only condition to further determine if there were any significantly different pairs (Bonferroni correction applied). The results showed that the interactions between the conga and guitar ($p = 0.133$) and the conga and speech ($p = 0.203$) were not significant. In addition, the effect size, r, indicated a small effect (<0.2 for both). Therefore it can be concluded that the effect of sound source on the preference gradings was not significant.

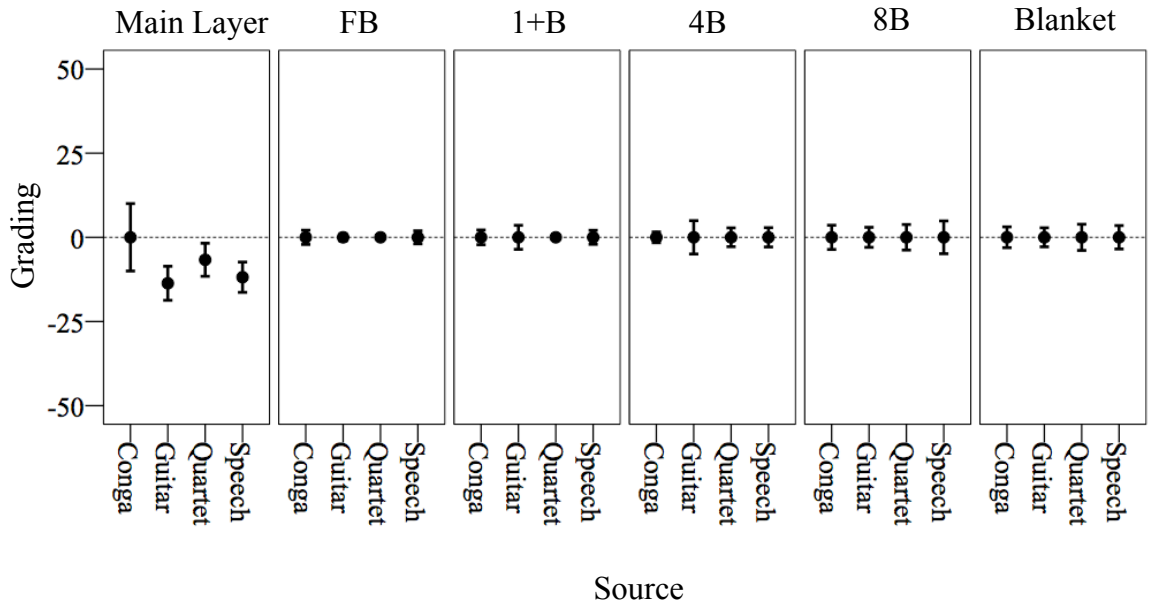

Fig. 5.17: Medians and associated notch edges showing the effect of sound source on subjective preference.

5.6.2.4 The Effect of Localisation Threshold Method

The effect of localisation threshold method on the median preference gradings is shown in Fig. 5.18. The medians have been plotted with notch edges. As its effect was found to be non-significant, the data for each sound source has been amalgamated. The overlapping notch edges suggested both that there was not a significant difference in preference between each of the threshold methods and further that each method was preferred significantly more than the main layer only condition. A Friedman test was conducted on the data in order to verify this. The results showed that the effect of localisation threshold method was significant (p = 0.000). A Wilcoxon test was conducted to determine which pairs were significantly different from one another (Bonferroni correction applied). The result showed principally that there were no significant differences in preference between the localisation threshold conditions. Instead, the only significant differences were between the main layer only condition and 1+B, 4B and blanket (p = 0.000 for all). It

should be noted that the Wilcoxon test results did not show a significant difference between the main layer condition and either FB (p = 0.075) or 8B (p = 1.000). This, however, would appear to be a Type-II error as a result of the use of the Bonferroni correction, as there is clearly no overlap between the notch edges for these stimuli. It can therefore be concluded that the main layer only condition was rated as being significantly less preferred with respect to the conditions whereby a localisation threshold was applied to the height layer. The threshold conditions themselves were rated consistently and not significantly different from one another.

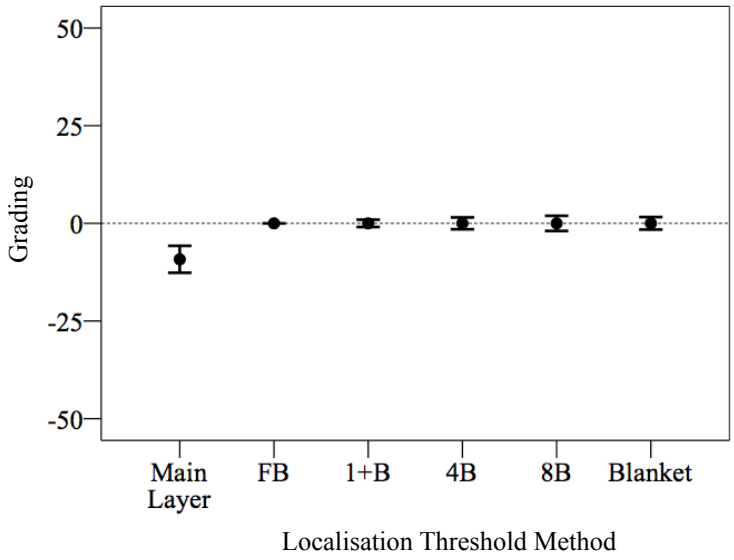

Fig. 5.18: Medians and associated notch edges showing the effect of localisation threshold method on subjective preference.

5.6.3 Discussion

The results of the present study demonstrated principally that subjective preference was not significantly affected by loudness (with the exception of the main layer only condition for speech with 0 ms ICTD), ICTD or sound source. Instead, the only significant effect observed was that the localisation threshold conditions

were considered as being more preferable with respect to the main layer only condition. Further, there were no significant differences between the preference gradings for any of the localisation threshold methods.

In an experiment conducted by Lee [2006] to determine the effect of horizontal interchannel crosstalk on subjective preference, it was noted that subjects tended to make preference judgments based on those attributes that were most saliently affected by the effect. In this regard, the factors contributing most to subjective preference in that experiment were locatedness and source width. From the feedback provided by subjects as to the reasons behind the choice of the most (Table 5.7) and least (Table 5.8) preferred sounds in the present study, there is evidence to support Lee's [2006] suggestion. Here it can be seen that perceptions of both fullness and VIS influenced subjective preference. However, it is also clear that there were other important attributes, such as perceived brightness, naturalness and the sense of envelopment, which also contributed. Because the present study did not consider the weighting for the perception of each attribute, it is not possible to say whether or not the fullness and VIS attributes were the most salient attributes for the determination of preference. Equally, no subject indicated that their preference decisions were influenced by loudness, which was one of the most salient effects of vertical interchannel crosstalk. Therefore, the results of the present study show some agreement with what was reported by Lee [2006], however this would have to be studied further.

Table 5.7: Descriptors for most preferred stimuli in preference tests.

ICTD (ms)	Method	Conga	Guitar	Quartet	Speech
0	Main Layer	Natural (5), less full, bright (3), good locatedness	Good locatedness	Good VIS, natural, balanced, clean	Natural (2), full, good locatedness
	FB	Full	Natural	Enveloping, good horizontal spread	Natural (3), less boomy, enveloping, good locatedness
	1+B	Full	Full (4), bright, enveloping, natural, small VIS		Full, good locatedness, more present
	4B	Natural, present, enveloping (2), good VIS, good horizontal spread	Natural (2), clear, bright (2), full (2), enveloping (2), good VIS (2), good locatedness	Natural (2), full (2), warm, bright, enveloping, good VIS	Natural (4), harder, full (2), good VIS, good horizontal spread
	8B		Good horizontal spread, natural (2), full (2), balanced	Natural, full, good VIS, good horizontal spread, good locatedness	Full (2), natural, bright, present, good VIS
	Blanket	Good locatedness (2), full (2), good horizontal spread	Balanced, natural	Full (4), spacious, natural (2), enveloping (2), balanced	Full, not boomy
1	Main Layer	Natural (4), clear, bright (2), less harsh	Natural (3), bright (2), full, less VIS, less bright, less harsh, good VIS, clean, good locatedness	Bright (2), clear, natural, defined, enveloping, good locatedness (2), enveloping	Natural (3), less boomy, good locatedness, less VIS
	FB		Good locatedness, natural, good horizontal spread	Bright	
	1+B	Full, natural, bright, clear	Balanced, enveloping (2), full, natural		Full (2), good horizontal spread, natural (2)
	4B	Natural (3), full, enveloping, good VIS, good horizontal spread, balanced	Good VIS, balanced (2), natural (3), full	Full, good VIS (2), good horizontal spread (2), enveloping, good locatedness	Good horizontal spread (2), enveloping, full (2)
	8B	Hard, enveloping	Bright, clear	Enveloping (3), good horizontal spread, full	Natural, full, good VIS (2), good horizontal spread
	Blanket	Full, enveloping	Full, enveloping (3), good VIS, good horizontal spread, good locatedness	Good horizontal spread, good VIS (2), natural (3), full, bright, enveloping	Full (4), natural, bright, enveloping, good VIS, balanced

Table 5.8: Descriptors for least preferred stimuli in preference tests.

ICTD (ms)	Method	Conga	Guitar	Quartet	Speech
0	Main Layer	Unnatural (2), thin (4), lacked punch, harsh, distant	Narrow horizontal width (2), distant, too bright (2), thin (2), muffled, unnatural, poor VIS (3), dull (2), distant	Unnatural, thin (2), fewer mids, muffled, dull (2), narrow (2), poor VIS, poor horizontal spread (2)	Thin (7), tinny, distant, dry, boomy, narrow, too bright, dull, poor VIS, too present
	FB	Thin (2)	Dull, thin	Too bright	
	1+B	Too full		Dull	
	4B	Too full, dull, resonant, muddy	Too bright, harsh	Dull, poor locatedness	Poor locatedness, too full, bloated, boomy, poor VIS
	8B	Too full (3), muffled, unnatural, hard	Thin (2), dull (2), poor VIS	Muffled, thin, muddy	Too full (2) thin, distant, dull
	Blanket		Poor locatedness	Boomy, poor locatedness	Too present
1	Main Layer	Too full, thin (4), narrow, harsh, hard	Poor VIS (2), unnatural (3), dull (3), thin (4), narrow, poor horizontal width, lifeless	Dull (2), thin (3), reverberant, poor VIS	Thin (4), dry, poor VIS (2), unnatural (2), narrow, dull (2), distant
	FB		Dull, narrow	Unnatural, dull, poor VIS	Thin, narrow, dull
	1+B	Hard, poor VIS		Thin, poor VIS, narrow horizontal width, dull	Too present
	4B	Hard, harsh, dull, thin, boomy	Unnatural, muffled, bright, thin		Unnatural (2), boomy
	8B	Too full, muffled, dull (2), unnatural	Too bright, harsh, excessive VIS	Unnatural	Too full (2), unnatural, odd VIS, bloated
	Blanket	Harsh, hard	Too bright	Poor locatedness	Too full (2)

A further point made by Lee [2006] was that, if preference was somewhat determined by the most salient attributes, subjects may use their perceptions of how that attribute should sound in order to determine their preference. Indeed, there is evidence in the results of the present study to suggest that this was the case. For example, from the data in Tables 5.7 and 5.8, it can be seen that the 1+B conga (0 ms) was highly regarded by some subjects, who commented that the sound was preferred due to its perceived fullness. However, other subjects perceived the degree of fullness for the same source as being excessive, instead choosing to give the stimulus a low rating. Equally, it can be seen that subjects often cited the perception of a 'good' VIS as being the reason for a high preference grading for the guitar stimuli. Despite this, for other subjects this sensation of VIS was considered as being unnatural, with the main layer condition instead being preferred and the localisation threshold conditions being rated poorly. It should be noted that there were numerous other examples of this across the entire test and this was perhaps the reason why there was no significant difference between the subjective preference gradings for each of the localisation threshold methods. However, it is clear that there is again limited agreement with Lee [2006] as the perceptual weightings of each of the attributes with respect to the preference decisions is not known. Despite this, there is at least evidence that subjective preference as to how some of the more salient effects of vertical interchannel crosstalk should sound influenced the results of the present experiment. This might also be the reason why there was no correlation between subjective preference and the perception of VIS and fullness for each method, as described in Part Three.

A further result of note was that the preference gradings for each of the localisation threshold methods were significantly higher than those for the main layer only condition. This result would seemingly indicate that there is a benefit to the inclusion of direct sounds in the height layer at the localisation threshold. However, it should be noted that this was highly subjective. In Fig. 5.19 the percentage of responses that rated each method as the most and least preferred is shown. Here it can be seen that sources presented using the main layer only condition were the least preferred for 44.4% of trials. However, it should also be noted that in 38.8% of trials a localisation threshold method was considered as being the least preferred, with the highest of these being 8B (12.5%). Moreover, although for 66.3% of trials a localisation threshold was considered as

being the most preferred, 19.4% favoured the main layer only condition. Further to this point, of the six conditions tested, the main layer only condition was rated as the most preferred on the most occasions (4B was the next highest with 18.8%). It can therefore be concluded that, although in general the presence of direct sounds in the height layer at the localisation threshold enhances the perception of music, this is somewhat subjective.

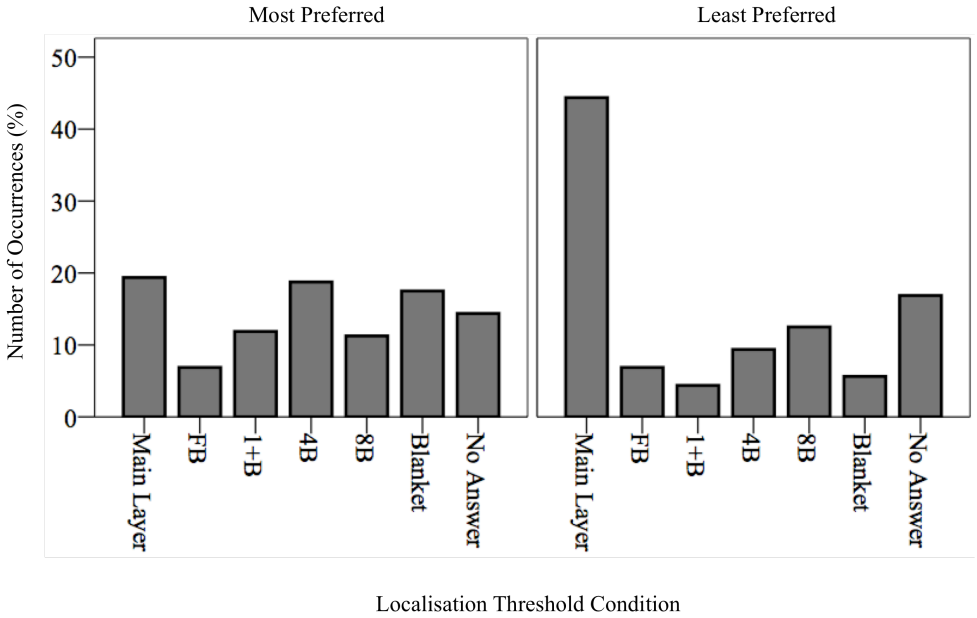

Fig. 5.19: Number of occurrences that each of the localisation threshold methods was rated as being the most preferred for a given trial (left) and the least preferred (right). "No answer" denotes a situation whereby no sound was rated as either most or least preferred.

In addition, although the localisation threshold conditions were considered as being significantly more preferred with respect to the main layer only condition, none of the conditions tested received a positive median grading. Instead, a Wilcoxon signed rank test showed that each condition was not significantly different from the '0' position on the scale (i.e. no preference). This result might mean that, as a whole, subjects generally felt indifferent about their perceptions of each of the localisation threshold methods. Conversely, this result can be taken as being indicative as to the subjective nature of the test. It can be seen from both Table 5.7 and Table 5.8, for example, that almost all of the stimuli were considered as being either

the most or least preferred in at least one of the trials. Equally, from Fig. 5.19 it is clear that no one localisation threshold particularly stood out as being consistently the most preferred by subjects. In this regard the 4B and blanket conditions were rated as the methods that were most often the most preferred and each of these were rated this way for less than 20% of the trials. Therefore, that there was no significant difference between the gradings for each of the threshold methods and the '0' position on the scale might be related to the fact that each method was equally liked and disliked by the subjects as a whole. However, as it is not possible to separate preference for the method from preference for the source itself, such a conclusion is limited until the experiment can be repeated with a much broader range of sound sources.

5.7 Practical Implications

The results of the present study have indicated that vertical interchannel crosstalk enhances the perception of VIS and fullness. Further, such effects are considered as being beneficial when the direct sound in the height layer is reduced to the localisation threshold compared to when it is absent entirely. This naturally has implications for recording and image rendering techniques for 3D audio reproduction. With respect to the former, positioning the height layer of microphones as described in Section 4.1.4.4 will enhance the perceived fullness and VIS of the main channel signal, without affecting its perceived location. Equally, from an image rendering perspective, the results indicate that there are numerous ways to incorporate direct sounds in the height layer. Each of these will have different effects on the perception of fullness and VIS and ultimately it is at the engineer's discretion to decide which method should be used for a given situation. It should also be noted that each of the derived methods in the thesis would affect the perception of other attributes, such as horizontal spread, naturalness and brightness, and each of these would also need consideration when deciding which method to use.

It should be noted that there are a number of limitations with regards to the above conclusions. As was mentioned in Chapter 0, the primary function of the height channels is to enhance the perception of LEV

through the addition of late reflections. However, the presence of such reflections on the perception of different attributes was not considered in the present experiment. It might be, for example, that late vertical reflections enhance the perceived VIS of the main channel signal even before the effects of vertical interchannel crosstalk. Therefore, any additional enhancement as a result of the effect might cause the main channel signal to sound unnatural, which in turn would affect preference. Somewhat related to the above issue was that only a vertical quadraphonic condition was considered for testing. Although representative of the frontal channels of a 3D audio configuration, this method negated the potential effects of the centre and rear loudspeakers in both layers. Arguably, the perceptual effects of vertical interchannel crosstalk, as well as subjective preference, would have to be analysed under such conditions before definitive conclusions could be made.

A further point of note is that the stimuli tested were generally single sources, which were presented as phantom images from directly in front of the listening position, with even the quartet source being a monophonic recording that was fed independently to each channel. In order that the results can be more applicable to recording techniques for 3D audio formats, it is important both that localisation thresholds are analysed away from the median plane and further that stereophonic sound sources are considered. It is clear then that the results of the present experiment provide useful guidelines as to the effects of vertical interchannel crosstalk, how these vary at the localisation threshold and which localisation threshold methods are the most preferred, however further study would be needed to increase the relevance for recording and image rendering techniques for 3D audio systems.

5.8 CONCLUSION

The present experiment was conducted in order to analyse the perceptual effects of vertical interchannel crosstalk, as well as the subjective preference of localisation thresholds. The experiment was conducted in four parts. In Part One, the perceptual effects of vertical interchannel crosstalk were elicited. Conga, speech,

quartet and guitar sources were presented to subjects using a vertical quadraphonic condition in a listening room. Subjects were required to directly compare the main layer only condition to a condition whereby the source was presented as a vertically oriented phantom image with 0 dB ICLD and a delay in the height layer of either 0 or 1 ms. First, subjects sat an individual elicitation test during which they noted the perceived differences between each of the crosstalk and no crosstalk conditions. Following this, a group discussion took place in which each of the terms elicited by subjects were grouped into common attributes. From this part of the experiment a total of 13 attributes were elicited, being horizontal source width, vertical image spread, locatedness, source elevation, fullness, source distance, hardness, brightness, naturalness, envelopment, loudness, richness and resonance. There was the potential that some of these attributes may have arisen purely as a result of difference in amplitude, however it was reasoned that a number of them would have been perceived anyway based on the literature.

In Part Two, audibility-grading tests were conducted in order to determine the most salient effects of vertical interchannel crosstalk. The test conditions required subjects to compare each of the crosstalk on and crosstalk off conditions and rate the perceived differences with respect to each of the thirteen attributes elicited in Part One using a 10-point scale. The results of the study showed that the most salient perceptual effects were increases in VIS, source elevation and loudness, which each had audibility indexes greater than 6.0. The most audible attribute relating to changes in timbre was ‘fullness’, whose audibility index was 5.0. That changes in VIS were salient was somewhat expected. This was due both to the presence of a secondary elevated source at an amplitude equal to that of the main channel signal and also because of the loudness increase.

Part Three considered how the most salient perceptual effects were affected when the different localisation threshold methods derived in Chapter Four were applied to the height layer. As it was already clear how source elevation and loudness would be affected, these attributes were not considered. Instead, the test focused on changes in VIS and fullness. Subjects compared the main layer only condition to each of the threshold conditions, as well as the 0 dB condition, and rated the perceived difference in each of the

attributes. The results showed principally that both VIS and fullness increase as a result of vertical interchannel crosstalk. In addition to this, at the localisation threshold variations in both fullness and VIS were audible with respect to the main channel signal, however these were dependent on the source, ICTD and the threshold method itself. It was also shown that there was a positive correlation between VIS and fullness. This result was interpreted from the perspective of loudness changes, with comb filtering also being thought to have had some effect.

Part Four considered the subjective preference of localisation thresholds. The test conditions required subjects to directly compare each of the localisation threshold methods, as well as the main layer only condition, to each other and grade their preference for each one. Alongside this, subjects were required to give feedback as to the reasons why they liked the most preferred sound, and disliked the least, for each trial. Subjects sat two tests for this experiment. For the first, the test stimuli were not level matched, whilst for the second they were. This was to determine whether or not changes in loudness influenced preference. The results showed that the preference gradings were not influenced by loudness, sound source or ICTD. Instead, the only key difference was that the localisation threshold conditions were all graded significantly higher than the main layer only condition. There was evidence that preference was influenced by subject's own opinions with respect to the acceptability of the most salient perceptual effects of vertical interchannel crosstalk, although numerous other attributes were also shown to have influenced their judgments. Also, despite the main layer only condition being given a significantly lower grading than the threshold methods, there were numerous occasions whereby the main layer only condition was more preferred. Also, the lack of significant difference between the threshold methods themselves indicated that the preference judgments were highly subjective.

The results of the experiment ultimately suggest that presence of the direct sound in the height layer at the localisation threshold is beneficial. There are however a number of additional factors that need to be considered, including the effects of reverberation and the use of stereophonic sound sources, before the results can be fully applied to image rendering and recording techniques for 3D audio formats.

5.9 SUMMARY

The present experiment was conducted to determine the most salient perceptual effects of vertical interchannel crosstalk and, further, to analyse the subjective preference of localisation thresholds. The experiment was conducted in four parts. Overall, the results of the experiment have revealed the following:

- The most salient perceptual effects of vertical interchannel crosstalk are increases in loudness, source elevation, VIS and fullness.
- A series of other spatial, timbral and location-based effects are also audible at a much lower level.
- When the localisation threshold is applied, variations in loudness, VIS and fullness remain audible; this is somewhat dependent on the sound source, ICTD and localisation threshold method used.
- Presence of the direct sound in the height layer at each of the localisation threshold methods tested resulted in a significantly higher preference grading than that for the main layer only. This indicates that there is a positive effect of having direct sounds feature in the height layer.
- There was no difference in preference gradings for each of the localisation threshold methods, with preference being highly subjective. It therefore falls to the individual to determine the most ideal method for a given situation.

6 SUMMARY AND CONCLUSIONS

The final chapter of the present thesis summarises the experimental works that have been conducted throughout. First, a summary of each chapter is provided. Following this, the key conclusions arising from the experiment are discussed. The practical applications of the results are also considered, with the chapter concluding with plans for future work.

6.1 CHAPTER SUMMARY

6.1.1 CHAPTER ZERO (INTRODUCTION)

This chapter provided a background into the concept of vertical interchannel crosstalk, which occurs when direct sound is recorded in the height microphones during 3D sound recordings. Vertical interchannel crosstalk has garnered little attention in the literature, with the most salient effects not currently being known. What is known, however, is that the effect can cause the main channel signal to be formed as a phantom image between the main and height layers. Studies have considered the minimum amount of attenuation of direct sound necessary in the height layer for the perceived location of the main channel signal to be unaffected (the localisation threshold). However, such studies have not considered the frequency-dependency of median plane localisation, choosing instead to reduce the amplitude of the direct sound in the height layer evenly across the frequency spectrum (the 'blanket reduction' method). The present thesis aimed to analyse the frequency-dependency of localisation thresholds and determine whether localisation thresholds could be applied through the frequency-dependent manipulation of the direct sound in the height layer (the 'band reduction' method). Additional interest was given to the operation of the precedence effect in the median plane, which, if present, would have meant that sufficient ICTD alone would have prevented

vertical interchannel crosstalk from affecting the perceived location of the main channel signal without ICLD being necessary.

From the above background the following research questions, to be addressed in the present thesis, were derived:

1. How do localisation thresholds vary across the frequency spectrum and is there a sound source dependency for more natural stimuli?
2. Can any evidence be found to support the existence of the precedence effect for vertically arranged loudspeakers?
3. What are the most salient perceptual effects of vertical interchannel crosstalk?
4. How are the most salient effects affected when applying the localisation threshold using the band and blanket reduction methods and which method is more subjectively preferred?

6.1.2 CHAPTER ONE (THE LOCALISATION OF ELEVATED SOUND SOURCES)

Chapter One conducted a full review into human sound localisation, in particular with respect to sound sources incident from the median plane. The chapter began by introducing several general human localisation mechanisms, including the duplex theory of sound localisation and HRTF cues. Following this, sound localisation in the median plane was explored in depth. Notably, the necessity of the spectral cues provided by the pinnae for accurate vertical localisation was explored. The spectral cues relating specifically to elevation perception were also considered, as were the additional localisation cues provided by both head rotations and shoulder and torso reflections. Subsequently, the frequency dependency of median plane localisation was examined. This included a review of directional bands, which primarily describes the relationship between the centre frequency of 1/3-octave bands and the position at which they are localised on the median plane. In addition, the pitch-height effect was considered; this being a phenomenon whereby tonal and band-limited stimuli are localised progressively higher in space as their centre frequency increases.

The chapter concluded with a discussion of the phantom image elevation effect, which describes the increase in perceived phantom image elevation as the base angle between stereophonic loudspeakers increases. Additional discussions with respect to vertical interchannel crosstalk were also presented. They included:

- The phantom image elevation effect might mean that the localisation thresholds might be lower, as the main channel signal would already be elevated with respect to the physical position of the main channel layer.
- It might be possible to apply the localisation threshold by attenuating the directional bands in the height layer that relate to elevation perception (i.e. in the 4-10 kHz region).

6.1.3 CHAPTER TWO (THE PERCEPTUAL EFFECTS OF SECONDARY VERTICAL SOURCES)

In Chapter Two, the perceptual effects of secondary vertical sources were reviewed. This subject was divided into three broad areas; the effect on perceived location, the effect on perceived timbre and the effect on perceived spatial impression. With respect to perceived location, it was considered how a secondary vertical source (the height layer) could cause a direct sound presented form the main layer to be formed as a vertically oriented phantom image at a position intermediate between the two layers; the greater the amplitude of the reflection relative to the direct sound, the greater the elevation. In addition, the effect of ICTD was considered, which primarily comprised a review of the precedence effect. Evidence supporting the existence of this effect in the median plane was presented and discussed. With respect to the effect on perceived timbre, the concept of comb filtering was discussed. This included an analysis of the perceptual effects of both lateral and vertical reflections on perceived timbre, as well as the threshold amplitude of the reflection relative to the direct sound for timbral changes to be audible. The chapter concluded with a review of spatial impression. The concepts of ASW and LEV were first considered, with particular interest being given to the effect of vertical reflections on the perception of each. Following this, a somewhat brief review

of VIS was presented. With respect to vertical interchannel crosstalk, the chapter featured the following discussions:

- As vertical amplitude panning has been shown to be unstable, vertical interchannel crosstalk might have an erratic and unpredictable effect on the perceived elevation of the main channel signal, which might affect the localisation threshold.
- If it were to operate, the precedence effect would mean that the localisation threshold could be applied with sufficient ICTD alone. However, previous research has only showed evidence for localisation dominance and, even then, this has not affected the localisation threshold.
- Timbral colouration of the main channel signal would be the most severe when the vertical interchannel crosstalk signal is similar in amplitude and arrives with a short delay with respect to the main channel signal.
- Colouration effects might be more audible for vertical interchannel crosstalk than they are for horizontal interchannel crosstalk.
- The degree of timbral colouration as a result of vertical interchannel crosstalk might be source dependent.
- From the perspective of spatial impression, the most salient perceptual effects of vertical interchannel crosstalk are likely to be different compared to those for horizontal interchannel crosstalk. The source width increases reported for the latter effect are not expected for vertical interchannel crosstalk, with increases in perceived VIS being more likely.

6.1.4 CHAPTER THREE (THE FREQUENCY DEPENDENCY OF LOCALISATION THRESHOLDS)

Two experiments were described in Chapter Three. The first (Experiment One) examined if there existed a frequency dependency of localisation thresholds. Octave band stimuli, with centre frequencies ranging from 125 Hz – 8 kHz, were presented to subjects from vertically arranged stereophonic loudspeakers located in

front of the listening position in an anechoic chamber. Subjects completed a method-of-adjustment task in which the minimum amount of attenuation of the height loudspeaker necessary for the resultant phantom image position to match that of the same source presented from the main loudspeaker alone (the localisation threshold) was analysed. Delays ranging from 0-10 ms were applied to the height loudspeaker with respect to the main in order to emulate different spacings between the main and height layer of microphones in the context of microphone techniques for recording for 3D audio formats. Key findings from the experiment were as follows:

- Localisation thresholds for octave band stimuli are frequency dependent. Low frequency stimuli required significantly less level reduction than did the mid-high frequencies.
- There was no significant effect of ICTD on localisation threshold for any stimulus.
- ICLD was always necessary for the broadband pink noise source, suggesting that the precedence effect did not operate.
- It might be possible to apply localisation thresholds by manipulating the amplitude of single frequency bands, rather than by reducing direct sound levels as a whole, in the height layer.
- The frequency dependency of localisation thresholds might be related to differences in VIS between the main layer only and phantom image conditions.
- For the broadband pink noise source, the relative strength of the spectral cues from each loudspeaker layer may determine whether the localisation threshold has been met, particularly in the 7-9 kHz region. However, the lack of a localisation dominance effect also indicates the importance of VIS differences.

The second experiment (Experiment Two) was a localisation test, which was conducted in order to determine whether or not there existed a relationship between the perceived elevation of the test stimuli tested in Experiment One and the localisation thresholds that were obtained. The test setup was identical to that used for Experiment One, with the exception that a numbered scale was positioned in front of the listening position. This scale was to be used by subjects to make their localisation judgments. Subjects identified

perceived source elevation for main and height loudspeaker presentation only, as well as for the vertically oriented phantom image positions. The same ICTDs as used in Experiment One were used for this experiment. Key findings were as follows:

- The pitch-height effect is maintained for octave band stimuli presented as vertically oriented phantom images.
- The 1-4 kHz octave band stimuli were significantly affected by ICTD, however the effect was random and inconsistent; the other octave bands were not significantly affected.
- For the broadband source, increasing the ICTD from 0-1 ms caused the resultant phantom image to perceptually move closer to the later loudspeaker. This arguably showed a lack of localisation dominance and was likely related to the effects of comb filtering.
- The perceived elevation of the test stimuli did not correlate with the localisation thresholds derived in Experiment One. This arguably validates the VIS hypothesis for the band-limited stimuli. For the broadband source, the results support the hypothesis that the balance of spectral energy between the main and height layers in the range at which the pinna cues for elevation exist is important.
- No evidence could be found to support the operation of the precedence effect in median plane stereophony.

6.1.5 CHAPTER FOUR (ANALYSIS OF BAND AND BLANKET REDUCTION LOCALISATION THRESHOLD METHODS)

Three experiments were described in Chapter Four. The first (Experiment Three) was conducted as a more thorough analysis of localisation thresholds using the blanket reduction method. Anechoically recorded speech, conga, quartet, guitar and oboe sources were presented to subjects in a listening room using vertical stereophonic and quadraphonic conditions with the height layer delayed with respect to the main by 0, 1 and

10 ms. Subjects completed an adaptive method-of-adjustment task in order to find the localisation thresholds. Key findings from the experiment were as follows:

- The localisation thresholds for the natural sound sources tested were only significantly affected by changes in ICTD. The threshold was -9.5 dB for 0 ms ICTD and -7 dB for 1 and 10 ms.
- The non-significant effect of sound source might be explained by the suggestion that the mechanism that determines whether or not the localisation threshold has been met for natural sound sources is the balance of spectral cues provided by the main and height layers, particularly in the 7-9 kHz region. This is determined primarily by the ICLD and is not affected by the spectral content of the source itself.
- The non-significant effect of presentation method was explained on the basis that increases in ICLD resulted in similar differences in spectral energy in the 7-9 kHz region between the main layer only and phantom image conditions for both presentation methods.
- The results were indicative of a localisation dominance effect of the earlier loudspeaker, although the precedence effect itself was not observed.

In the second experiment (Experiment Four), the frequency dependency of localisation thresholds was explored in a natural listening environment. Burst and continuous octave bands, with centre frequencies ranging from 63 Hz – 16 kHz, were presented to subjects under the same conditions as were tested in Experiment Three. Key findings from the experiment were as follows:

- The frequency dependency of localisation thresholds was maintained in the presence of reflections, albeit the effect was not as strong as was observed in anechoic conditions (Experiment One).

- The effect of frequency was less strong compared to Experiment One. It was suggested that this might be related to comb-filtering effects as a result of floor reflections.
- Differences in the filtering and threshold detection methods might also have caused differences in the results between the two experiments.
- The localisation thresholds obtained were not affected by presentation method. For the octave band stimuli, this might be because the frequency-dependent differences in perceived VIS between the main layer only and phantom image conditions differ little between vertical stereophonic and quadraphonic presentation.
- For the broadband source, the result might have been caused by the same mechanism as was suggested for natural sources following Experiment Three.
- The effect of frequency was dependent on signal duration, being stronger for continuous presentation compared to bursts. This might be because the full-extent of perceived VIS requires time to build up. As such, the duration of the burst stimuli might have been too short to enable frequency-dependent differences in VIS between the main layer only and phantom image conditions to be perceptible. This would also explain why the threshold was higher for the bursts compared to the continuous stimuli.
- ICTD had a random and inconsistent effect on a limited number of stimuli. Significant effects were observed for 4 kHz bursts, 8 kHz continuous and broadband pink noise for both durations. This might be related to the random effects of comb filtering, with the effect being too inconsistent to be indicative of a localisation dominance effect.
- No evidence was found to support the operation of the precedence effect in the median plane.

In final experiment conducted in the Chapter (Experiment Five), a series of methods of applying the localisation threshold using band reduction were proposed. The methods suggested were FB (frequency-dependent attenuation of all octave bands), 1+B (frequency-dependent attenuation of all octave bands at 1 kHz and above), 4B (only the 4 kHz octave band is attenuated) and 8B (only the 8 kHz octave band is

attenuated). A verification test was conducted in which the perceived location each of the natural sound sources when applying each of the band reduction methods, as well as the blanket reduction method, was tested in order to see which were successful at preventing the perceived location of the main channel signal from being affected by vertical interchannel crosstalk. Key findings from the experiment were as follows:

- The blanket reduction thresholds obtained in Experiment Three are sufficient at preventing vertical interchannel crosstalk from affecting the perceived location of the main channel signal. Each of the band reduction methods tested (FB, 1+B, 4B and 8B) were also effective.
- That the 4B method was effective indicates that the balance of spectral energy in the 7-9 kHz region between each layer might not be as important in determining the localisation threshold for natural sound sources as had been suggested previously.
- Vertical interchannel crosstalk has a source dependent effect on the perceived location of the main channel signal. That the localisation thresholds obtained in Experiment Three were not source dependent indicates that the threshold does not necessarily relate to the perceived elevation differences between the main layer only and 0 dB phantom image conditions.
- The oboe source might have been localised based on the pitch height effect. If this is the case, then the balance of spectral energy hypothesis would not explain the localisation thresholds obtained for this source. Instead differences in VIS might be the key mechanism. This might also apply to other natural sound sources.
- No localisation dominance effect was observed; a delay in the height layer instead caused perceived source elevation to increase. Equally, no evidence was found for the precedence effect. It is therefore unclear why less level reduction was necessary in the presence of an ICTD in Experiment Three.
- The phantom image elevation effect might not operate for vertically oriented phantom images.

6.1.6 CHAPTER FIVE (THE PERCEPTUAL EFFECTS OF VERTICAL INTERCHANNEL CROSSTALK)

One experiment was described in Chapter Five (Experiment Six), which was split into four parts. The purpose of this experiment was to determine the most salient perceptual effects of vertical interchannel crosstalk and further to analyse the subjective preference of localisation threshold methods. In Part One, subjects identified the perceived differences between no crosstalk and maximum crosstalk conditions (0 and 1 ms ICTD), with group discussions used to determine a set of common attributes. In Part Two, the audibility of each of the elicited attributes was tested in order to determine those that were the most salient. For Part Three, the perception of each of the most salient attributes was analysed when the different localisation threshold methods derived in Chapter Four were applied to the height layer. Part Four considered the subjective preference of the localisation threshold methods. Key findings from the experiment were as follows:

- The most salient perceptual effects of vertical interchannel crosstalk are increases in loudness, source elevation, VIS and fullness.
- A series of other spatial, timbral and location-based effects are also audible at a much lower level.
- When the localisation threshold is applied, variations in loudness, VIS and fullness remain audible; this is somewhat dependent on the sound source, ICTD and localisation threshold method used.
- Presence of the direct sound in the height layer at each of the localisation threshold methods tested resulted in a significantly higher preference grading than that for the main layer only. This indicates that there is a positive effect of having direct sounds feature in the height layer.
- There was no difference in preference gradings for each of the localisation threshold methods, with preference being highly subjective. It therefore falls to the individual to determine the most ideal method for a given situation.

6.2 CONCLUSIONS

Four research questions were derived at the beginning of this thesis. This section will seek to answer each of them based on the experimental works presented throughout.

1. How do localisation thresholds vary across the frequency spectrum and is there a sound source dependency for more natural stimuli?

   The results of Experiments One and Four showed that localisation thresholds are frequency dependent when octave bands of pink noise are used as the test stimuli. In anechoic conditions (Experiment One), it was found that significantly less level reduction was needed for the low frequency stimuli compared to the mid-high frequency stimuli. In addition, for this experiment the localisation thresholds obtained were not significantly affected by sound source. Discussions were presented that regarded to frequency dependency of localisaiton thresholds to be related to frequency-dependent differences in perceived VIS between the main layer only and phantom image conditions.

   The frequency dependency of localisation thresholds was considered in more detail in Experiment Four. In this case, the experiment was conducted in a natural listening environment and considered a range of variables including frequency, signal duration, ICTD and presentation method. Primarily, it was determined that the frequency dependency of localisation threholds was maintained in a natural listening environment. However, it should be noted that the effect was not as strong as had been reported in Experiment One. It was thought that the presence of reflections, differences in the test method and the use of different filters to create the test stimuli might have been the reasons for this result. There was no effect of presentation method, whilst the effect of ICTD was erratic and indicative of the effects of comb filtering. Signal duration was found to have a significant effect, with the burst stimuli often requiring significantly less level reduction than did the continuous

stimuli. This result was explained based on an apparent 'build up' of perceived VIS over time. Ultimately, the results of this experiment were used to develop a series of band reduction methods for the application of localisation thresholds (FB, 1+B, 4B, 8B) and these were each shown to work in Experiment Five.

Experiment Three considered the effect of sound source on the localisation threshold using the blanket reduction method. The experiment found that, although there were differences in the median threshold between sources, these differences were not significant. Further, when the same thresholds were applied to each source in Experiment Five, it was found that there were no significant differences between each phantom image condition and the main layer only condition. The results therefore agree with those reported by Lee [2011] and Stenzl et al. [2014] in that there is not a sound source dependency of localisation thresholds for natural stimuli. It should be noted that the results of Expeirment Four indicated that different thresholds could be applied to more transient sources compared to those that are more continuous in nature, although this would have to be studied further.

2. Can any evidence be found to support the existence of the precedence effect for vertically arranged loudspeakers?

The precedence effect was condiered either directly or indirectly in Experiments One-Five. In the localisation threshold experiments (One, Three and Four), there was no instance whereby a delay of the signal in the height layer resulted in perceived location matching that of main layer only presentation. Instead, ICLD was always necessary for the localisation threshold to be met. Had the precedence effect operated then, arguably, this result would not have been obtained. Equally, in the localisation experiments (Two and Five) the application of ICTD did not result in perceived source location matching the positon of the main layer loudspeakers, further indicating the non-operation of the precedence effect.

Interestingly, the results of the exeriments also failed to conclusively show the existence of a localisaiton dominance effect for vertically arranged loudspeakers. In Experiment Two, pink noise stimuli were localised progressively higher when delays between 0 and 1 ms were applied to the height layer, which is the opposite of what would have been expected had a localisation dominace effect been present. This result was rationalized based on the presence of additional notches in the spectrum, as a result of comb filtering, that could be mistaken as elevation cues. In addition, delays greater than 1 ms resulted in perceived source location being identical to the 0 ms condition. In Experiment Three it was found that significantly less ICLD was required to meet the localisation threshold when using the blanket reduction method in the presence of an ICTD, with it initially being thought that this was due to a localisation dominace effect. However, the results of Experiment Five showed that, as with broadband pink noise in Experiment Two, delays of 1 ms resulted in perceived elevation judgments being higher than those for 0 ms ICTD. It is therefore unclear what caused the result in Experiment Three. It is at least sufficed to say that the present experimental data does not support the existence of either a localisation dominace effect nor the precedence effect for vertically arranged loudspeakers.

3. What are the most salient perceptual effects of vertical interchannel crosstalk?

Experiment Six (Parts One and Two) found that there were thirteen audible effects as a result of vertical interchannel crosstalk. There were divided into timbral (fullness, hardness, brightness, naturalness, richness and resonance), spatial (VIS, horizontal source width and envelopment), location (locatedness, source elevation and source distance) and loudness (loudness) effects. It might have been the case that some of these were elicited based purely on the loudness differences between the maximum and no corsstalk conditions, however it was reasoned based on the literature that a number of the attributes would likely have been perceived anyway. Subsequent experimentation determined that the most salient perceptual effects of vertical interchannel crosstalk were increases

in VIS, loudness, source elevation and fullness. The other attributes were audible to a lesser degree and were therefore not considered as being salient.

4. How are the most salient effects affected when applying the localisation threshold using the band and blanket reduction methods and which method is more subjectively preferred?

The effects of loudness and source elevation were not considered with respect to this question. Loudness would naturally depend on the localisation threshold method being used, whilst the application of the localisation threshold would mean that there would be no variations in source elevation. Instead, the effect of the different threshold methods on perceived fullness and VIS was considered (Experiment Six, Part Three). The results of the experiment showed that there remained audible variations in both attributes when the localisation threshold was applied although these differences were dependent on numerous factors, including the ICTD and sound source; this makes it difficult to generalise the results. Additionally, it was found that increases in perceived fullness corresponded with increases in perceived VIS. This was explained in two ways. Firstly, it is likely that the result was related to the effects of loudness. However, it might also be the case that comb filtering influenced the perception of each attribute, particularly as the perception of each attribute did not necessarily increase with increases in loudness, especially in the presence of a 1 ms delay.

With respect to preference, Experiment Six Part Four showed that, whilst there was no significant difference between preferences for each localisation threshold method, the application of a localisation threshold was significantly preferred compared to the main layer only condition. It was also interesting to note that there was evidence to suggest that preference judgments were influenced by subject's expectations regarding how some of the most salient effects of vertical interchannel crosstalk should sound, whilst there was also no difference in results when the stimuli were and were not loudness matched. These results, along with those from Part Three, indicate that experimentation is the key when determining which localisation threshold method to use in a given situation.

6.3 PRACTICAL IMPLICATIONS

The discussions following each experiment have considered numerous ways in which the results can be applied in practical situations. For instance, the results of Experiment Three were used to identify that the microphone techniques proposed by Lee [2011], for recording for 3D audio formats, might need to be revised slightly in order to ensure that the perceived location of the main channel signal is not affected by vertical interchannel crosstalk. In this regard, a small increase in the necessary angle between the direct sound and the height microphones was suggested. Equally, the band reduction method was developed based on the results of Experiment Four and this has shown that there are numerous ways to incorporate direct sounds in the height layer without the perceived location of the main channel signal being affected. Each of these would result in differing perceptions of VIS and fullness and it is at the discretion of the engineer to determine which method is appropriate for a given situation.

That there is no 'one size fits all' approach to the application of localisation thresholds leads to the proposition of a plugin for the rendering of 3D images based on the results of the present thesis (example interface shown in Fig 6.1). The features of such a plugin would be as follows. Firstly, it would be important for the user to be able to cycle through each localisation threshold method (FB, 1+B, 4B, 8B and blanket). In each case, the plugin would use an 8^{th} order Butterworth filter to split out the necessary bands in the signal and would mix them into the height layer at amplitudes that would depend on the other settings within the plugin. For example, if a transient source was being used then the user might select the 'Burst' option and this would generally mean less level reduction compared to if the 'Cont.' option (i.e. for continuous stimuli) was selected. Equally, the ICTD parameter would introduce delays for the direct sound in the height layer and would also have a bearing on the amount of level reduction. Further, the 'Routing' options would influence where in the height layer the direct sound is routed to but would not influence the ICLD. Being able to quickly cycle through each method for a given track would enable engineers to determine the method that is best for them and that gives them the most desired outcome. It should be noted that further study

would be beneficial in order to expand on the available parameters but this at least demonstrates in theory how the results of the present thesis could be applied to develop a plugin for the rendering of 3D images.

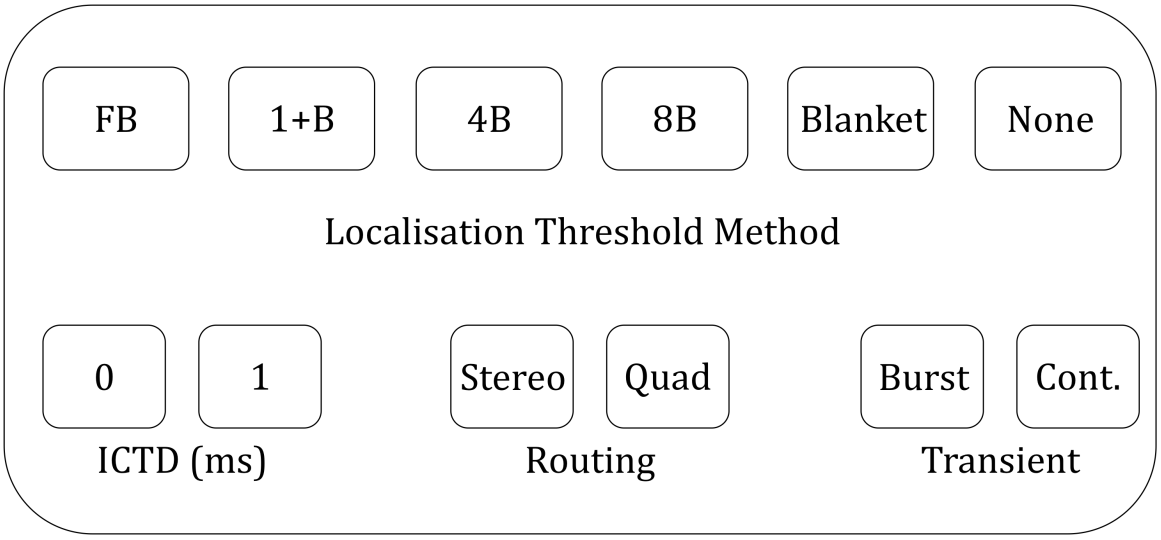

Fig 6.1: Example interface for a 3D image rendering plugin based on the results presented in the present thesis.

6.4 FURTHER WORK

Throughout the present thesis, continual attempts have been made to establish the mechanism that determines the localisation threshold for both complex and octave band stimuli. To reiterate, two hypotheses were proposed with respect to this. The first stated that the balance of spectral cues provided by the main and height layers respectively in the 7-9 kHz range was important. Increases in ICLD will result in the cues provided by the main layer being more dominant, which would ultimately result in the localisation threshold being met. The second hypothesis stated that increases in VIS are perceived when source presentation changes from main layer only to vertical phantom image. These differences could be mistaken as differences in perceived elevation. Increases in ICLD will reduce the difference in perceived VIS between each condition, leading to the localisation threshold being met. In the present thesis, there were a number of

results obtained that arguably rely on the knowing the precise mechanism that governs the localisation threshold in order that they can be explained. This includes the results that the thresholds were not affected by either presentation method or sound source and further that the 4B and 8B conditions were both successful at applying the threshold. Therefore, it is of interest to determine further what the primary mechanism is that determines the localisation threshold and this will help to explain a number of the results presented in this thesis.

In addition, following Experiment Four, it was identified that the perception of the full extent of VIS is not instantaneous for band-limited stimuli presented as vertically oriented phantom images. Instead, the image builds up over the course of a few seconds, with the spread of frequencies within the source being in line with the pitch-height effect. It would be interesting to see if this effect is prevalent for subjects other than the author. Further, assuming this effect does indeed exist, the threshold time for the full extent of VIS to be built up requires attention. This would be interesting from the perspective of pure psychoacoustics and will also be valuable in helping to ascertain why the effect of signal duration was found to be significant for the octave band stimuli tested in Experiment Four.

A further point of note would be to consider the derived thresholds in a more practical setting. For example, on numerous occasions the present thesis has referenced the implications of the results for microphone techniques for recording in 3D audio formats. This, however, is yet to be verified practically. It would be interesting to compare recordings that utilised the methods suggested in the present thesis (i.e. cardioid microphones in the height layer angled at least 105° away from the direct sound) to those that did not. This would help to determine both if the proposed methods are fit-for-purpose and further how distracting the location based effects of vertical interchannel crosstalk actually are. Related to this point is how appropriate the present experimental data is for 3D image rendering techniques. It might be the case, for example, that the presence of reverb influences the perception of VIS and fullness that are introduced as a result of featuring the direct sound in the height layer. This in turn might mean that there is no reason to include direct sounds in the height layer, as the benefits of doing so might be masked. Therefore, the suggestions proposed,

as a result of the experimental data presented, need to be considered in more practical situations in future study.

APPENDIX A: DIRECTIONAL BANDS REVISITED[1]

This appendix describes an experiment that was conducted in order to analyse the directional band phenomenon reported by Blauert [1969]. More specifically, the experiment considered the effects of bandwidth and signal duration, which had seen limited attention previously. In addition, a more refined scale was used in order to analyse the correlation between directional bands and the pitch height effect. This experiment was presented at the 138th convention of the Audio Engineering Society (AES) in May 2015. The below is taken from the manuscript submitted in line with the requirements of the AES.

A.0 ABSTRACT

Listening tests were undertaken as part of a comprehensive analysis of directional bands. The effects of frequency, loudspeaker position, signal duration and bandwidth were all considered. The results confirmed the existence of directional bands for 1, 4 and 8 kHz 1/3-octave band bursts. A relationship between pitch and height was also observed. Bandwidth was found to have a variable effect on localisation, depending on frequency, indicating that the spectral cues used in vertical localisation are not of equal bandwidth. Loudspeaker position and signal duration also had some influence on localisation judgments although this was found to be somewhat erratic.

A.1 INTRODUCTION

The localisation of band-limited stimuli in the median plane is a topic that has been researched for many decades. One of the key studies in this field was conducted by Blauert [1969], who used 1/3-octave bands of

[1] Wallis, R. and Lee, H. [2015]: 'Directional Bands Revisited', Audio Engineering Society 138th Convention, Preprint 9278.

noise as the test stimuli. Blauert [1969] concluded that localisation judgments were made independent of the position of the emitting loudspeaker. Instead, localisation was governed solely by the frequency of the stimulus, with certain bands of frequencies being related to specific locations on the median plane. Blauert [1969] described these bands as "directional bands". Among other results, the experimental data revealed a link between both 500 Hz and 4 kHz and front localisation, 1 kHz and the rear and 8 kHz and above. Subsequent experiments by both Hebrank & Wright [1974a] and Asano et al. [1990] have demonstrated that directional bands are related to the spectral cues provided by the pinnae in vertical localisation. In the Hebrank and Wright [1974a] study, for example, it was found that overhead localisation corresponded to a ¼-octave peak in the region between 7 and 9 kHz. Also, Asano et al. [1990] showed that elevation cues existed in the region above 5 kHz.

However, the response method used in Blauert's [1969] experiment was somewhat restricted, with subjects only able to identify whether a stimulus was in front of, above or behind them. Prior to this study, vertical localisation studies conducted by Pratt [1930], Trimble [1934] and Roffler and Butler [1968b] had all identified a relationship between pitch and height (the so-called "pitch-height effect") when tones were used as the test stimuli. The three-region scale utilised by Blauert [1969] did not enable a full analysis of whether the pitch-height effect had been maintained for 1/3-octave bands of noise, although this method allowed a simple and time-efficient experiment. Had a more refined scale been utilised in Blauert's [1969] study then such a correlation could potentially have been identified. It might be, for example, that, rather than both being related simply to "front" localisation, 4000 Hz 1/3-octave band stimuli are associated with the elevated front whilst 500 Hz 1/3-octave bands are localised in a somewhat lower front position. Such a result would provide further agreement with spectral cues.

In addition to the above point, only 1/3-octave bands of noise were tested in Blauert's directional bands experiment. This issue was partially addressed by Itoh et al. [2007], who showed that directional bands were maintained when 1/6-octave bands were used as the test stimuli. Despite this finding, the effect of bandwidth arguably remains largely unexplored. Previous median plane localisation studies [Pratt 1930, Trimble 1934,

Roffler and Butler 1968b, Itoh et al. 2007] have tested the localisation of either tones or octave bands although none considered bandwidth as an independent variable. Moreover, the aforementioned studies were concerned with investigating the localisation of sound sources vertically arranged in front of the listener and therefore utilised an entirely different response method to that used by Blauert [1969]. Due to these contextual differences it becomes difficult to directly compare the effect of bandwidth across each study, let alone determine its influence on directional bands. Perrett & Noble [1995], for example, demonstrated that differences in response method alone have a large bearing on the results of localisation studies. It is therefore necessary to investigate in more detail the effect of bandwidth on directional bands.

In light of the above discussion, listening tests were conducted in a bid to provide a more comprehensive study of directional bands. It was thought that by utilising a more refined scale and a range of bandwidths a more comprehensive understanding of the directional band theory could be provided. Of additional interest in the study were the effects of both signal duration and loudspeaker position on the directional bands perception, for which detailed data have not been reported in the literature. The analysis of multiple variables under the same test method was considered as being valuable in advancing the understanding of median plane localisation.

A.2 EXPERIMENTAL DESIGN

A.2.1 Physical Setup

The listening tests were conducted in the anechoic chamber at the University of Huddersfield. The experiments utilised two Genelec 8040A loudspeakers, one directly in front of the listening position and the other directly behind. The distance from each loudspeaker to the listening position was 1.4 m. Subjects were sat on a height adjustable chair, which was used to ensure that the ear height of all subjects matched the centre point of the speaker cone on each loudspeaker. Surrounding the listening position was an acoustically

transparent curtain. This was utilised so that subjects were not influenced by knowledge of the number of loudspeakers or their position.

A.2.2 Test Stimuli

The stimuli used for the listening tests were created by brick-wall filtering pink noise in to octave bands, 1/3-octave bands and tones, with centre frequencies of 500, 1000, 4000 & 8000 Hz. The choice of frequency was partially motivated by directional bands [Blauert 1969], with the stimuli above 500 Hz being strongly related to specific locations on the median plane. It was thought that this would make directional bands simpler to identify from the experimental data. Additionally, 500 Hz stimuli were utilised for the experiment as this enabled a more broad frequency range to be tested, potentially making any relationship between pitch and height more apparent.

Stimuli were presented to listeners from each loudspeaker individually as either continuous noise (1s onset/offset) or as bursts (200ms duration, 10ms onset/offset, 1 per second) at 75dB LAeq (SPL). The burst stimuli were similar to those used by Blauert [1969], whilst the use of continuous stimuli enabled the difference in localisation of burst and continuous stimuli to be analysed.

Stimuli were arranged and presented in a randomised order to subjects. Each stimulus lasted for a total of ten seconds. This was followed by five seconds of silence after which the next stimulus was played and so forth. This methodology allowed for the 48 stimuli (four frequencies, three bandwidths, two durations, two loudspeakers) to be tested in a twelve-minute timeframe, therefore minimizing listener fatigue.

A.2.3 Subjects

The test group comprised of staff and both undergraduate and postgraduate students from the University of Huddersfield's Music Technology department. 12 subjects participated in the test and they were chosen due to their critical listening experience, with no subject identifying any loss of hearing prior to testing.

A.2.4 Test Method

Prior to the start of each test, subjects were provided with a scale that was to be used when making localisation judgments (Fig. A.1). The entire median plane was divided into eight identical regions; "Front", "Front High", "Above", "Back High", "Back", "Back Low", "Below" and "Front Low". The scale was designed as a more refined version of that used by Blauert [1969], allowing for more apparent differences in the localisation of each stimulus to be identified. The scale also allowed for correlations between pitch and height to be observed. In Blauert's [1969] study, the reason that such a limited and simplistic scale was due to both time constraints and the desire to not overly complicate the experiment. The scale used in the present study was arguably able keep the experiment simple whilst still making the results somewhat more revealing.

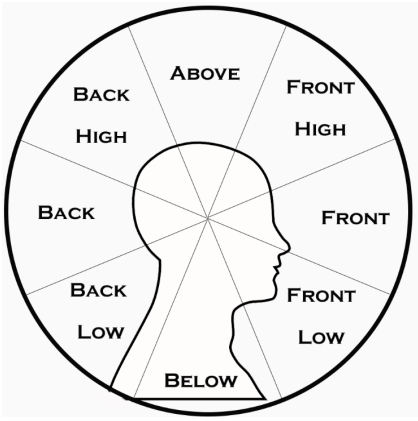

Fig. A.1. Scale used for localisation judgments.

Subjects were required to listen to each stimulus in full, making their localisation judgments only in the silence between stimuli. Localisation responses were not confined to a single region on the scale. Instead, for a given stimulus, subjects were able to respond with as many as regions as they saw fit. This option was given to subjects in a bid to make localisation responses as free as possible. To further remove the influences of any visual bias, subjects were instructed to close their eyes during the presentation of each stimulus. For the duration of the experiment, subjects were required to sit up straight and face forwards, with no head motion permitted. The experimenter sat in with subjects during each test and recorded their responses.

A.3 RESULTS

The results of the listening tests are shown in Fig. A.2. A summary of localisation judgments, separated by the centre frequency of the stimulus, is as follows.

A.3.1 500 Hz

Localisation judgments for the non-tonal 500 Hz stimuli presented continuously from the front show a strong tendency for frontal localisation. 60% of octave band stimuli and 80% of 1/3-octave band stimuli were localised in a frontal region ("Front Low", "Front" or "Front High"). As for tonal stimuli, the tendency for "Front" was less strong, with only 30% perceived in a frontal region. Instead, the tonal stimuli were more predominantly localised in a rear region ("Back Low", "Back", "Back High") (60%). Octave bands also had a small tendency for the rear, with 40% of judgments being in this area, particularly "Back". When the same stimuli were presented from the rear loudspeaker, responses for octave bands and tones were relatively similar. Responses for 1/3-octave bands featured a diminished relationship with the frontal regions, with responses becoming more biased towards the rear (60%).

Judgments for 1/3-octave band bursts from the front loudspeaker were similar to those for continuous presentation; 60% of responses were in a frontal region. However, the number of reversals for these stimuli increased to 30%. Tonal stimuli maintained their perceptual relationship with the rear (50%). Octave bands were predominantly localised in the lower rear, with 70% of responses being in either the "Back", "Back Low" or "Below" regions. For rear loudspeaker presentation, 1/3-octave bands saw a large increase in the number of responses in the rear (80%). These were predominantly focused in the "Back" and "Back High" regions. Octave bands continued to be localised in the rear however judgments were more biased towards the "Back" region (55%). Tonal stimuli were predominantly localised in the frontal regions (60%).

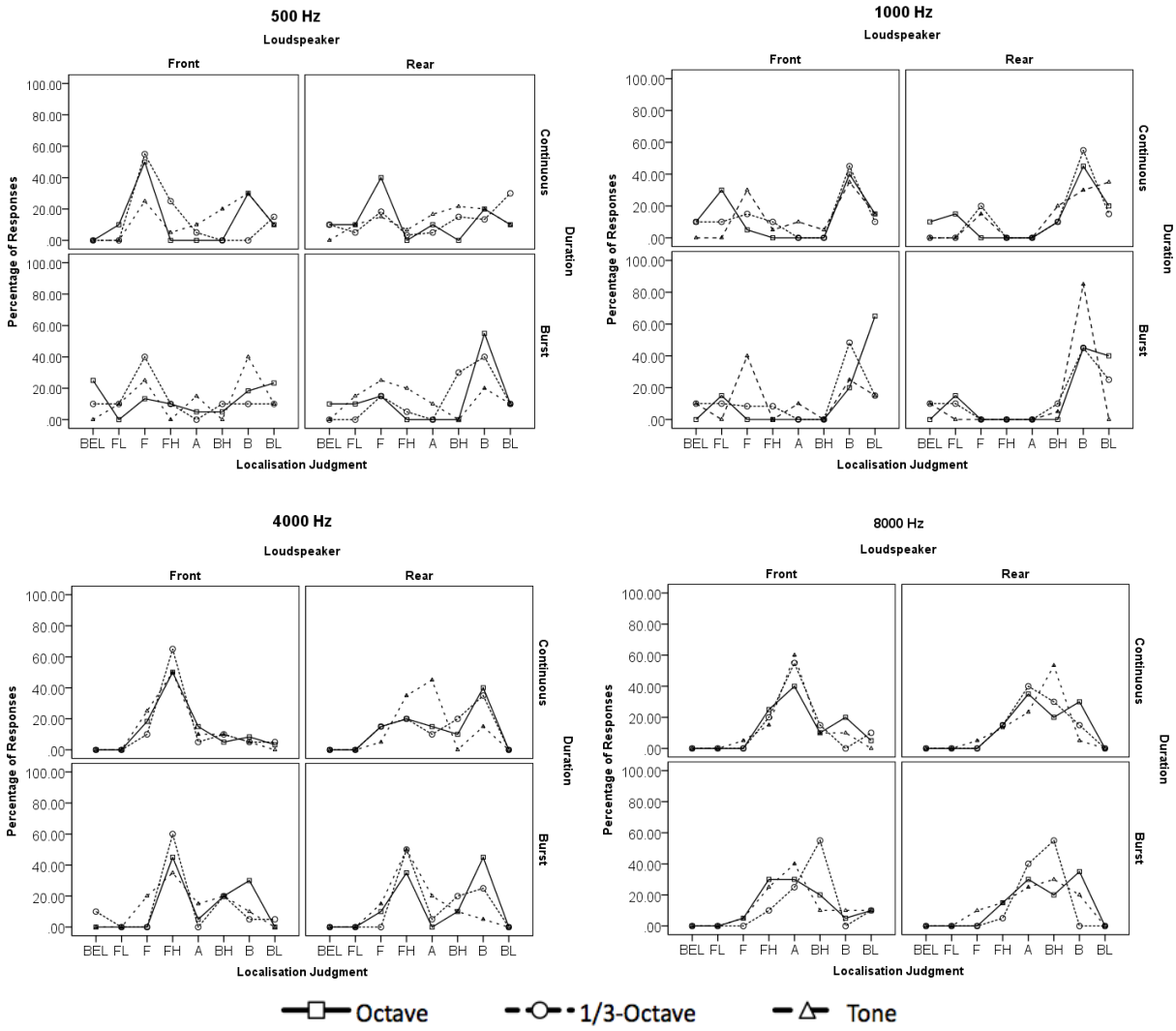

Fig. A.2. Percentage of localisation responses for each region on the test scale.

A.3.2 1000 Hz

For front loudspeaker presentation, localisation judgments for the continuous 1000 Hz stimuli were biased towards the rear, with all three bandwidths having 55% localisation in this region. These judgments were predominantly for "Back". Likewise, all bandwidths featured a good portion of frontal localisation (35% of judgments each). In this case, tonal stimuli were generally perceived as being slightly elevated with respect to the octave bands. The 1/3-octave bands were evenly spread across the frontal region. The relationship between 1000 Hz and the rear was maintained for the continuous stimuli presented from the rear loudspeaker, with judgments becoming more biased to the rear compared with front loudspeaker presentation. Again, localisation was mainly in the "Back" region although for tones the responses were spread across the rear, peaking at "Back Low" (35%).

Non-tonal burst stimuli presented from the front loudspeaker also featured a strong relationship with the rear, with 85% of judgments for octave bands and 65% for 1/3-octave bands being in this region. The relationship was diminished somewhat for the tonal stimuli, where localisation was spread evenly across the front and rear. For rear loudspeaker presentation, the vast majority of stimuli were localised in the rear. The results for octave and 1/3-octave bands remained relatively constant, whilst tonal stimuli featured 90% of responses in this area; 85% of these were in the "Back" region.

A.3.4 4000 Hz

For continuous front loudspeaker presentation, localisation judgments for the 4000 Hz stimuli were predominantly in a frontal region. A large portion of these were in "Front High" (50% for tones and octave bands, 65% for 1/3-octave bands). The number of reversals were minimal for all stimuli. For rear loudspeaker presentation this relationship lessened, with only tonal stimuli maintaining a strong tendency towards the elevated front regions (80%). The number of reversals increased for all stimuli, with octave and

1/3-octave bands having 50% and 55% judgments respectively in these areas. Only 20% of judgments for these stimuli were in the "Front High" region.

For front loudspeaker burst presentation, the tendency towards the elevated front was maintained, particularly for the 1/3-octave bands (60% in "Front High"). This was less strong for other bandwidths, especially the octave band stimuli, which were localised more often in the rear (50%) than the front (45%). The largest difference for rear loudspeaker burst presentation was that the 1/3-octave band stimuli were localised slightly less in the "Front High" region and had an increased tendency towards the rear. Localisation for the other bandwidths remained somewhat similar.

A.3.5 8000 Hz

For continuous front loudspeaker presentation, localisation judgments were focused primarily in the elevated regions ("Front High", "Above", "Back High"). The tendency for "Above" was most strong for the tonal and 1/3-octave band stimuli (60% & 55% respectively), this decreased for octave bands (40%). However, it should be noted that a total of 75% of responses for stimuli of this bandwidth were in the elevated regions. When the same stimuli were presented from the rear loudspeaker the tendency towards elevation was maintained, however responses in the "Back High" region were increased, particularly for the tonal stimuli (55%).

For front loudspeaker burst presentation, the relationship between 8000 Hz and the elevated regions was also apparent although responses for octave and 1/3-octave bands were more evenly spread in the elevated regions. The responses appeared to change little for front and rear presentation. The 1/3-otave band stimuli were strongly associated with the "Above" and "Back High" regions in both cases (80% and 95% respectively). The octave and tonal stimuli maintained their perceptual elevation however localisation judgments became slightly more biased towards the rear for rear loudspeaker presentation.

A.4 DISCUSSION

The results obtained in the present study seem to only partly agree with those presented by Blauert [1969]. In that study, 1000 Hz was related to behind perception, with 4000 Hz being related to the front. In the present study, 1000 Hz 1/3-octave band burst stimuli were consistently localised in the rear of the subject, whilst the results for 4000 Hz show a relationship with elevated frontal localisation. Interestingly, however, when 1000 Hz 1/3-octave bursts were presented from the front, 30% of responses were in a frontal region (the average % of FL, F and FH), which means that 30% of judgments were independent of directional bands. When 4000 Hz 1/3-octave band bursts were presented from the rear loudspeaker, 45% of responses were in a region behind the subject, as opposed to in front. From this, it is suggested that the strengths of directional bands for 1 kHz and 4 kHz were not as great as those reported by Blauert [1969]. In that study, the relative statistical frequency for a 4000 Hz stimulus to be localised in front was 55-75%, depending on stimulus amplitude, while for 1000 Hz the relative statistical frequency for behind localisation was 80-90%.

Additionally, [Blauert [1969] made an association between 8000 Hz and above localisation. The “Above” region in that study was roughly equal to the “Above” region in the present study with some overlap of “Front High” and “Back High”. The present data shows that this is where 8000 Hz 1/3-octave bands were predominantly localised, with a bias towards the elevated rear.

In Blauert’s [1969] study, 500 Hz was found to be predominantly related to frontal localisation, with a relative statistical frequency of between 60 and 70%. The results in the present study were unable to confirm this. Instead, localisation judgments for 500 Hz 1/3-octave band bursts in the present study were biased towards the position of the emitting loudspeaker. Based on Blauert’s [1969] definition, a directional band is maintained only if localisation judgments for a given frequency band are consistent irrelevant of the position of the emitting loudspeaker. Therefore, the present experimental data does not support the existence of directional bands for 500 Hz 1/3-octave band bursts.

Part of the motivation behind the use of a more refined localisation scale than that used by Blauert [1969] was to determine if any correlations between pitch and height could be observed. These results suggest that there existed a somewhat limited correlation between pitch and height; the 500 and 1000 Hz stimuli were generally localised in a non-elevated position, whereas 4000 Hz had a perceptually strong relationship to the "Front High" region. 8000 Hz had strong links to "Above" and "Back High" consistently.

The current results also demonstrate that loudspeaker positioning certainly has some influence on median plane localisation, although the effect is somewhat variable with no consistent results observed. In terms of the effect of loudspeaker positioning on 1/3-octave bands bursts, the results still show general agreement with Blauert [1969] for all frequencies except 500 Hz. It is evident that some of the 1 kHz and 4 kHz results were biased towards the physical direction of the loudspeaker.

This result leads to the following suggestions regarding the effect of bandwidth for burst stimuli. Firstly, as a whole, directional bands appear to operate more strongly for 1/3-octave band bursts than they do for either tones or octave bands. Also, directional bands were observed for all 8000 Hz stimuli with the exception of octave bands. Conversely, for the 1000 Hz stimuli, directional bands were observed for all stimuli except for tones. It therefore appears that the spectral cues used in vertical localisation are not of equal bandwidth. Such a suggestion had been made previously by both Hebrank and Wright [1974a] and Asano *et al.* [1990].

In general, The 1000 Hz and 8000 Hz stimuli were not greatly affected by changes in signal duration. Additionally, there was no regular pattern found for 4000 Hz. However, it is clear that signal duration had a large influence on the localisation of 500 Hz octave bands. When presented as bursts, no directional bands could be observed for the 500 Hz stimuli. Despite this, when presented as continuous octave bands from either loudspeaker, a relationship between 500 Hz and the front became more apparent. This result therefore represents the only experimental data obtained in the present study that shows the relationship between 500 Hz and front localisation described by Blauert [1969]. The reasons for the apparent differences between the present study and Blauert's [1969] regarding directional bands for 500 Hz burst are unclear. What is

suggested from the data obtained in the present study, however, is that directional bands for 500 Hz are dependent both on signal duration and bandwidth.

Of further interest in the study were front-back confusions. An intriguing example of front-back confusions in the present test can be seen in localisation judgments for the 4000 Hz octave bursts presented from the front loudspeaker. Localisation of these stimuli featured front-back confusions for 50% of judgments. The high number of front back confusions in this case indicates that there is no directional band for 4000 Hz octave bursts. This result is line with Morimoto *et al.* [2003] who found that no cues to help resolve front-back confusion existed at frequencies below 4.8 kHz, but disagrees with Asano *et al.* [3], who concluded that the region below 2 kHz is used to help determine front back confusions. This topic requires further research.

A.5 CONCLUSION

The present study was designed as a thorough examination of directional bands. The experimental data obtained using 1/3-octave band bursts tends to agree with Blauert's [1969] findings for 1, 4 and 8 kHz. However, a similar band for 500 Hz could not be observed, with loudspeaker positioning seemingly having an influence on localisation. Additionally, a relationship between pitch and height was identified. This lead to the suggestion that the pitch-height effect and directional bands are both part of the same localisational mechanism, with the response methods utilised in previous studies having a large influence on the results that were obtained.

Bandwidth was found to have a variable effect on burst stimulus localisation. Directional bands were maintained for all bandwidths at 8 kHz with the exception of octave bands. On the other hand, for the 1000 Hz stimuli directional bands were maintained for octave bands and not for tones. This therefore indicated that the spectral cues used in median plane localisation are of different bandwidths. Moreover, in general

directional bands were found to operate more strongly for 1/3-octave bands than for both narrower (tone) and wider (1-octave) bandwidths tested.

The effect of signal duration also varied. For the 1 and 8 kHz stimuli the effect was relatively minor, with differences in signal duration causing small localisation differences. A similar effect was seen for the majority of the 4 kHz stimuli however rear presentation of continuous 1/3-octave bands saw an increase in rear responses, with a more diffuse frontal localisation. Also, continuous octave bands presented from the front had a reduced association with the rear compared to bursts. Interestingly, for the 500 Hz band directional band-like localisation was only observed for continuous octave bands. This lead to the suggestion that both bandwidth and signal duration influence directional bands for stimuli of this frequency.

With regards to loudspeaker position, the results in the present study for 1/3-octave band bursts generally agreed with those reported by Blauert [1969]. However, loudspeaker position did influence localisation of some of the other stimuli. For 1000 Hz continuous tones and octave bands, as well as tone bursts, the strong rear tendency was reduced slightly for front presentation. Additionally, localisation judgments for the 500 and 4000 Hz stimuli were increasingly in the rear for rear loudspeaker presentation. A similar effect was seen for localisation of the 8 kHz stimuli.

REFERENCES

Algazi, V. R., Avendano, C. and Duda, R. O. [2001]: 'Elevation Localisation and Head-Related Transfer Function Analysis at Low Frequencies', *Journal of the Acoustical Society of America*, 109(3), pp. 1110-1122.

Asano, F., Suzuki, Y. and Sone, T. [1990]: 'Role of Spectral Cues in Median Plane Localization', *Journal of the Acoustical Society of America,* 88(1) pp.159-168.

Auro Technologies [2016]: Listening Formats: Auro 3D, retrieved from http://www.auro-3d.com/system/listening-formats

Barbour, J. L. [2003]: 'Elevation Perception, Phantom Images in the Vertical Hemi-Sphere', AES 24th International Conference on Multichannel Audio.

Barron, M. [1971]: 'The Subjective Effects of First Reflections in Concert Halls – The Need for Lateral Reflections', *Journal of Sound and Vibration*, 4(22), pp. 475-494.

Barron, M. and Marshall, A. H. [1981]: 'Spatial Impression due to Early Lateral Reflections in Concert Halls: The Derivation of a Physical Measure', *Journal of Sound and Vibration*, 77(2), pp. 211-232.

Bech, S. [1995]: 'Timbral Aspects of Reproduced Sound in Small Rooms. I', *Journal of the Acoustical Society of America*, 97(3), pp. 1717-1726.

Bech, S. [1998]: 'Spatial Aspects of Reproduced Sound in Small Rooms', *Journal of the Acoustical Society of America*, 103(1), pp. 434-445

Bech, S. & Zacharov, N. [2006]: *Perceptual Audio Evaluation: Theory, Method and Application* (Chester: Wiley)

Beranek, L. L. [2010]: 'Listener Envelopment LEV, Strength G and Reverberation Time RT in Concert Halls', Proceedings of the 20th International Congress on Acoustics.

Berg, J. and Rumsey, F. [1999]: 'Spatial Attribute Identification and Scaling by Repertory Grid Technique and Other Methods', AES 16th International Conference.

Blauert, J. [1969]: 'Sound Localisation in the Median Plane', *Acustica*, 22, pp. 205-213.

Blauert, J. [1971]: '*Localisation and the Law of the First Wavefront in the Median Plane*', Journal of the Acoustical Society of America, 50(2), pp. 466-470.

Blauert, J. and Lindemann, W. [1986]: 'Auditory Spaciousness: Some Further Psychoacoustic Analyses', *Journal of the Acoustical Society of America*, 80(2), pp. 533-542.

Blauert, J. [1997]: *Spatial Hearing: The Psychophysics of Human Sound Localisation* (Cambridge: MIT Press)

Bradley, J. S. and Soloudre, G. A. [1995]: 'The Influence of Late Arriving Energy on Spatial Impression', *Journal of the Acoustical Society of America*, 97(4), pp. 2263-2271.

Bradley, J. S., Reich, R. D. and Norcross, S. G. [2000]: 'On the Combined Effects of Early- and Late-Arriving Sound on Spatial Impression in Concert Halls', *Journal of the Acoustical Society of America*, 108, pp. 651-661.

Bridges, G. S. and Lisagor, N. S. [1975]: 'Scaling and Seriousness: An Evaluation of Magnitude and Category Scaling Techniques', *Journal of Criminal Law and Criminology*, 66(2), pp. 215-221.

Brunner, S., Maempel, H. J. and Weinzierl, S. [2007]: 'On the audibility of comb-filter distortions', Audio Engineering Society 122nd Convention, Preprint 7047.

Butler, R. A. & Belendiuk, K. [1977]: 'Spectral Cues Utilized in the Localization of Sound in the Median Sagittal Plane', *Journal of the Acoustical Society of America*, 61(5), pp. 1264-1269.

Cabrera, D. and Tiley, S. [2003]: 'Vertical Localisation and Image Size Effects in Loudspeaker Reproduction', *AES 24th International Conference on Multichannel Audio.*

Cabrera, D. and Morimoto, M. [2007]: 'Influence of Fundamental Frequency and Source Elevation on the Vertical Localisation of Complex Tones and Complex Tone Pairs', *Journal of the Acoustical Society of America*, 122(1), pp. 478-488.

Cardozo, B. L. [1965]: 'Adjusting the Method of Adjustment: SD vs. DL', *Journal of the Acoustical Society of America*, 37(5), pp. 786-792.

Chun, C. J., Kim, H. K., Choi, S. H., Jang, S. and Lee, S. [2011] 'Sound Source Elevation Using Spectral Notch Filtering and Directional Band Boosting in Stereo Loudspeaker Reproduction', IEEE Transactions on Consumer Electronics, 57(4), pp. 1915-1920.

Cohen, J. [1988]: *Statistical Power Analysis for the Behavioral Sciences* (New York: Lawrence Erlbaum Associates)

De Boer, K. [1947]: 'A Remarkable Phenomenon with Stereophonic Sound Reproduction', *Philips Technical Review*, 9, pp.8-13.

Dimmick, F. L. and Gaylord, E. [1934]: 'The Dependence of Auditory Localisation Upon Pitch', *Journal of Experimental Psychology*, 17(4) pp. 593-599.

Dolby Laboratories [2016]: *Dolby Atmos*, retrieved from http://www.dolby.com/us/en/brands/dolby-atmos.html

Everest, F. A. and Pohlmann, K. C. [2009]: Master Handbook of Acoustics (New York: McGraw Hill)

Freyman, R. L., Clifton, R. K. and Litovsky, R. Y. [1991]: 'Dynamic Processes in the Precedence Effect', *Journal of the Acoustical Society of America*, 90(2), pp. 874-884.

Farnell, A. [2010]: *Designing Sound* (Cambridge: MIT Press)

Fazenda, B., Elmer, L. A., Hargreaves, J. A., Hirst, J. M. and Wankling, M. [2010]: 'Subjective Preference of Modal Control Methods in Listening Rooms', *Journal of the Audio Engineering Society*, 60(5), pp. 338-349.

Ferguson, S. and Cabrera, D. [2005]: 'Vertical Localisation of Sound from Multiway Loudspeakers', *Journal of the Audio Engineering Society*, 53(3), pp.163-173.

Francombe, J. [2014]: 'Perceptual Evaluation of Audio-on-Audio Interference in a Personal Sound Zone System', PhD Thesis, University of Surrey.

Furuya, H., Fujimoto, K., Takeshima, Y. and Nakamura, H. [1995]: 'Effect of Early Reflections from Upside on Auditory Envelopment', *Journal of the Acoustical Society of Japan*, 16(2), pp. 97-104.

Furuya, H., Fujimoto, K., Ji, C. Y. and Higa, N. [2001]: 'Arrival Direction of Late Sound and Listener Envelopment', *Applied Acoustics*, 62, pp. 125-136.

Gardener, B. and Martin, K. [2000]: '*HRTF Measurements of a KEMAR Dummy-Head Microphone*', Retrieved 13 February 2016 from http://sound.media.mit.edu/resources/KEMAR.html

Gardner, M. B. [1973]: 'Some monaural and binaural facets of median plane localisation', *Journal of the Acoustical Society of America*, 54(6), pp. 1489-1495.

Gardner, M. B. and Gardner, R. S. [1973]: 'Problem of Localization in the Median Plane: Effect of Pinnae Cavity Occlusion', *Journal of the Acoustical Society of America*, 53(2), pp. 400-408.

Haas, H. [1972]: 'The Influence of a Single Echo on the Audibility of Speech', *Journal of the Audio Engineering Society*, 20(2), pp. 146-159.

Halmrast, T. [2000]: 'Orchestral Timbre: Comb-Filter Coloration From Reflections', *Journal of Sound and Vibration*, 232(1), pp. 53-69.

Halmrast, T. [2001]: 'Sound Colouration From (Very) Early Reflections', ASA, Acoustical Society of America Meeting, Chicago.

Hanyu, T. and Kimura, S. [2001]: 'A New Objective Measure for Evaluation of Listener Envelopment Focusing on The Spatial Balance of Reflections', *Applied Acoustics*, 62, pp. 155-184.

Hartmann, W. M. [1983]: 'Localisation of Sound in Rooms', *Journal of the Acoustical Society of America*, 74(5), pp. 1380-1391.

Hebrank, J. and Wright, D. [1974a]: 'Spectral Cues used in the Localisation of Sound Sources on the Median Plane', *Journal of the Acoustical Society of America*, 56 (6), pp. 1829-1834.

Hesse, A. [1986]: 'Comparison of Several Psychophysical Procedures with Respect to Threshold Estimates, Reproducibility and Efficiency, *Acustica*, 59, pp. 263-273.

Howard, D. M. and Angus, J. A. S. [2009]: *Acoustics and Psychoacoustics* (Oxford: Focal Press)

Itoh, M., Iida, K. and Morimoto, M. [2007]: 'Individual Differences in Directional Bands in Median Plane Localisation', *Applied Acoustics*, 68, pp. 909-915.

ITU [1994]: 'Recommendation ITU-R BS.1116-1: Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems', *International Telecommunications Union*.

ITU [1998]: 'Recommendation ITU-T P.800: Methods for Subjective Determination of Transmission Quality', *International Telecommunications Union*.

ITU [2003]: 'Recommendation ITU-R BS.1284-1: General Methods for the Subjective Assessment of Sound Quality'*, International Telecommunications Union.*

Iwaya, Y. Suzuki, Y. and Kimura, D. [2003]: 'Effects of Head Movement on Front-Back Error in Sound Localisation', *Acoustics, Science and Technology*, 24(5), pp. 322-324.

Jo, H., Martens, W. L., Park, Y. and Kim, S. [2010]: 'Confirming the Perception of Virtual Source Elevation Effects Created Using 5.1 Channel Surround Sound Playback', in Proceedings of the 9th ACM SIGGRAPH Conference on Virtual-Reality Continuum and its Applications in Industry.

Jogan, M. and Stocker, A. A. [2014]: 'A New Two Alternative Forced Choice Method for the Unbiased Characterization of Perceptual Bias and Discriminability', *Journal of Vision*, 14(3), pp. 1-18.

Johnson, D., Harker, A. and Lee, H. [2015]: 'HAART: A New Impulse Response Toolbox for Spatial Audio Research', Audio Engineering Society 138th Convention, eBrief 190.

Johnson, T., Gibson, I., Evans, B. and Wendl, M. [2016]: 'An Investigation into Kinect and Middleware Error and Their Suitability for Academic Listening Tests', Audio Engineering Society 140th Convention, eBrief 273.

Kirchner, J. [2001]: Data Analysis Toolkit #1: Graphically Displaying Data Distributions, retrieved from http://seismo.berkeley.edu/~kirchner/eps_120/Toolkits/Toolkit_01.pdf

Kirkeby, O., Seppala, E. T., Karkkainen, A. Karkkainen, L. and Huttunen, T. [2007]: 'Some Effects of the Torso on Head-Related Transfer Functions, Audio Engineering Society 122nd Convention, Preprint 7030.

Kuhn, F. G. and Guernsey, M. [1983]: 'Sound Pressure Distribution About the Human Head and Torso', *Journal of the Acoustical Society of America*, 73(1), pp. 95-105.

Kuhn, F. G. [1987]: 'Physical Measurements and Acoustics Pertaining to Directional Hearing', *Journal of the Acoustical Society of America*, 73, pp. 1.

Kurozumi, K. and Ohgushi, K. [1983]: 'The Relationship Between the Cross-Correlation Coefficient of Two-Channel Acoustic Signals and the Sound Quality', *Journal of the Acoustical Society of Japan*, 74(6), pp. 1726-1733.

Lawless, H. T. [2013]: *Quantitative Sensory Analysis: Psychophysics, Models and Intelligent Design* (Chichester: Wiley-Blackwell)

Lee, H. [2006]: 'Effects of Interchannel Crosstalk in Multichannel Microphone Technique', PhD Thesis, University of Surrey.

Lee, H. [2011]: 'The Relationship Between Interchannel Time and Level Differences in Vertical Sound Localisation and Masking', Audio Engineering Society 131st Convention, Preprint 8556.

Lee, H. [2016]: 'Perceptual Band Allocation (PBA) for the Rendering of Vertical Image Spread with a Vertical 2D Loudspeaker Array', *Journal of the Audio Engineering Society*, 64(12), pp. 1003-1013.

Lee, H. [2017]: 'Sound Source and Loudspeaker Base Angle Dependency of Phantom Image Elevation Effect', Accepted for publication in the Journal of the Audio Engineering Society.

Lee, H., Johnson, D. and Mironovs, M. [2016]: 'A New Response Method for Auditory Localisation and Spread Tests', Audio Engineering Society 140th Convention, e-Brief 240.

Levitt, H. [1970]: 'Transformed Up-Down Methods in Psychoacoustics', *Journal of the Acoustical Society of America*, 49(2), pp. 467-477.

Licklider, J. C. R. [1948]: 'The Influence of Interaural Phase Relations Upon the Masking of Speech by White Noise', *Journal of the Acoustical Society of America*, 20(2), pp. 150-159.

Lieberman, M. D., and Cunningham, W. A. [2009]: 'Type I and Type II Error Concerns in fMRI Research: Re-balancing the Scale', *Social Cognitive and Affective Neuroscience*, 4(4), pp. 423-428.

Litovsky, R. Y., Rakerd, B., Tin, T. C. T. and Hartmann, W. M. [1997]: 'Psychophysical and Physiological Evidence for a Precedence Effect in the Median Sagittal Plane', *Journal of Neurophysiology*, 77(4), pp. 2223-2226.

Lodge, M. and Tursky, B. [1979]: 'Comparisons Between Category and Magnitude Scaling of Political Opinion Employing SRC/CPS Items, *American Political Science Review*, 73(1), pp. 50-66.

Lorho, G. [2005]: 'Individual Vocabulary Profiling of Spatial Enhancement Systems for Stereo Headphone Reproduction', Audio Engineering Society 119th Convention, Preprint 6629.

Martens, W. L., Atsushi, M. and Sungyoung, K. [2006]: 'Investigating Contextual Dependency in a Pairwise Preference Choice Task', Audio Engineering Society 28th International Conference.

Mason, R. and Rumsey, F. [2000]: 'An Assessment of the Spatial Performance of Virtual Home Theatre Algorithms by Subjective and Objective Methods', Audio Engineering Society 108th Convention, preprint 5157.

McGill, R., Tukey, J. W. and Larsen, W. A. [1978]: 'Variations of Box Plots', *Journal of the American Statistical Association,* 32(1) pp.12-16.

Microsoft [2017]: Kinect for Xbox One, retrieved from http://www.xbox.com/en-GB/xbox-one/accessories/Kinect

Middlebrooks, J. C. and Green, D. M. [1991] 'Sound Localisation by Human Listeners', *Annual Review of Psychology*, 42(1), pp. 135-159.

Mills, A. W. [1958] – 'On the Minimum Audible Angle', *Journal of the Acoustical Society of America*, 30(4), pp. 237-246.

Morikawa, D., Toyoda, Y. and Hirahara, T. [2013]: 'Head Movement During Horizontal and Median Sound Localisation Experiments in Which Head-Rotation is Allowed', In Proceedings of Meetings on Acoustics, 19.

Morimoto, M. and Nomachi, K. [1982]: 'Binaural Disparity Cues in Median-Plane Localisation', *Journal of the Acoustical Society of Japan*, 3(2), pp. 99-103.

Morimoto, M. and Aokata, H. [1984]: 'Localisation Cues of Sound Sources in the Upper Hemisphere', *Journal of the Acoustical Society of Japan*, 5(3), pp. 165-173.

Morimoto, M. [2002]: 'The Relation Between Spatial Impression and the Precedence Effect', Presented at the 8th International Conference on Auditory Display (ICAD).

Morimoto, M., Yairi, M., Iida, K. and Itoh, M. [2003]: The Role of Low Frequency Components in Median Plane Localisation', *Acoustical Science & Technology*, 24(2), pp. 76-82.

Moushegian, G. and Jeffress, L. A. [1959]: 'Role of Interaural Time and Intensity Differences in the Lateralization of Low-Frequency Tones', *Journal of the Acoustical Society of America*, 31(11), pp. 1441-1445.

Nyberg, D. and Berg, J. [2008]: 'Listener Envelopment – What Has Been Done and What Future Research is Needed?', Audio Engineering Society 124th Convention, Preprint 7379.

Olive, S. E. and Toole, F. E. [1989]: 'The Detection of Reflections in Typical Rooms', *Journal of the Audio Engineering Society*, 37(7/8), pp. 539-553.

Olive, S. E. [2003]: 'Differences in Performance and Preference of Trained Versus Untrained Listeners in Loudspeaker Tests: A Case Study', *Journal of the Audio Engineering Society*, 51(9), pp. 806-825.

Perneger, T. V. [1998]: 'What's Wrong with Bonferroni Adjustments', *British Medical Journal*, 316, pp. 1236-1238

Perrett, S. and Noble, W. [1995]: 'Available Response Choices Affect Localisation of Sound', *Perception and Psychophysics*, 57(2), pp. 150-158.

Perrett, S. and Noble, W. [1997]: 'The Effect of Head Rotations on Vertical Plane Sound Localisation', *Journal of the Acoustical Society of America*, 102(4), pp. 2325-2332.

Pratt, C. C. [1930]: 'The Spatial Character of High and Low Tones', *Journal of Experimental Psychology*, 13(3), pp. 278-285.

Pulkki, V. [1997]: 'Virtual Sound Source Positioning Using Vector Base Amplitude Panning', *Journal of the Audio Engineering Society*, 45(6), pp. 456-466.

Rakerd, B. and Hartmann, W. M. [1986]: 'Localisation of Sound in Rooms, III: Onset and Duration Effects', *Journal of the Acoustical Society of America*, 80(6), pp. 1695-1706.
Rao, D. and Xie, B. [2005]: 'Head Rotation and Sound Image Localisation in the Median Plane', *Chinese Science Bulletin*, 50(5), pp 412-416.

Rottger, S., Schroger, E., Grube, M., Grimm, S. and Rubsamen, R. [2007]: 'Mismatch Negativity on the Cone of Confusion', *Neuroscience Letters*, 414, pp. 178-182.

Roffler, S. K. and Butler, R. A. [1968a]: 'Factors that Influence the Localisation of Sound in the Vertical Plane', *Journal of the Acoustical Society of America*, 43 (6), pp. 1255-1259.

Roffler, S. K. and Butler, R. A. [1968b]: 'Localisation of Tonal Stimuli in the Vertical Plane', *Journal of the Acoustical Society of America*, 43(6), pp. 1260-1266.

Rosenzweig, M. R. and Rosenblith, W. A. [1950]: 'Some Electrophysiological Correlates of the Perception of Successive Clicks', *Journal of the Acoustical Society of America*, 22(6), pp. 878-880.

Rumsey, F. [2005]: *Spatial Audio* (Oxford: Focal Press)

Rumsey, F. and McCormick, T. [2009]: *Sound and Recording* (Oxford: Focal Press)

Sandel, T. T., Teas, D. C., Feddersen, W. E. and Jefferess, L. A. [1955] 'Localisation of Sound from Single and Paired Sources', *Journal of the Acoustical Society of America*, 27(5), pp. 842-852.

Sato, S. and Ando, Y. [2002]: 'Apparent Source Width (ASW) of Complex Noises in Relation to the Interaural Cross-Correlation Function', *Journal of Temporal Design in Architecture and the Environment*, 2(1), pp. 29-32.

Searle, C. L., Briada, L. D., Davis, M. F. and Colburn, H. S. [1976]: 'Model for Auditory Localisation', *Journal of the Acoustical Society of America*, 60(5), pp. 1164-1975.

Seki, Y. and Ito, K. [2003]: 'Coloration Perception Depending on Sound Direction', *IEEE Transactions on Speech and Audio Processing*, 11(6), pp. 817-825.

Shaw, E. A. G. and Teranishi, R. [1968]: 'Sound Pressure Generated in an External-Ear Replica and Real Human Ears by a Nearby Point Source', *Journal of the Acoustical Society of America*, 44 (1), pp. 240-249.

Simner, R. [1986]: 'An Improved Bonferroni Procedure for Multiple Tests of Significance', *Biometrika*, 73(3), pp. 751-754.

Soloudre, G. A., Michel, C. L. and Norcross, S. G. [2003]: 'Temporal Aspects of Listener Envelopment in Multichannel Surround Systems', Audio Engineering Society 114th Convention, Preprint 5803.

Somervile, T., Gilford, G. L. S., Spring, N. F. and Negus, R. D. M. [1965]: 'Recent Work on the Effects of Reflectors in Concert Halls and Music Studios', British Broadcasting Corporation Engineering Division Research Report No. B-085.

Stenzl, H., Scuda, U. and Lee, H. [2014]: 'Localisation and Masking Thresholds of Diagonally Positioned Sound Sources and Their Relationship to Interchannel Time and Level Differences', in Proceedings of the International Conference on Spatial Audio.

Stone, H. and Sidel, J. L. [2004]: *Sensory Evaluation Practices* (California: Academic Press)

Strutt, J. W. [1907]: 'On Our Perception of Sound Direction, *Philosophical Magazine*, 13, pp. 214-232.

Sundaram, S. and Kyriakakis, C. [2005]: 'Phantom Audio Sources with Vertically Separated Speakers', Audio Engineering Society 119th Convention, Preprint 6614.

Taylor, M. M. and Creelman, C. D. [1965]: 'PEST: Efficient Estimates on Probability Functions', *Journal of the Acoustical Society of America*, 41(4), pp. 782-787.

Theile, G. [2001]: 'Multichannel Natural Recording Based on Psychoacoustic Principles', In *Proceedings of the Audio Engineering Society 19th International Conference*, pp. 201-209.

Thurlow, W. and Parks, T. [1961]: 'Precedence Suppression Effects for Two-Click Sources', Perceptual and Motor Skills, 13, pp. 7-12.

Thurlow, W. R. and Runge, P. S. [1967]: 'Effect of Induced Head Movements on Localization of Direction of Sounds', *Journal of the Acoustical Society of America*, 42(2), pp. 480-488.

Toole, F. [2009]: *Sound Reproduction: Loudspeakers and Rooms* (Cambridge: Focal Press)

Tregonning, A. and Martin, B. [2015]: 'The Vertical Precedence Effect: Utilising Delay Panning for Height Channel Mixing in 3D Audio', Audio Engineering Society 139th convention, Preprint 9469.

Trevo, S., Laukkanen, P., Patynen, J. and Lokki, T. [2014]: 'Preferences of Critical Listening Environments Among Sound Engineers', *Journal of the Audio Engineering Society*, 62(5), pp. 300-314.

Trimble, O. [1934]: 'Localisation of Sound in the Anterior-Posterior and Vertical Dimensions of "Auditory" Space', *British Journal of Experimental Psychology*, 24(3), pp. 320-334.

Wallach, H. [1939]: 'On Sound Localisation', *Journal of the Acoustical Society of America*, 10, pp. 270-274.

Wallach, H. [1940]: 'The Role of Head Movements and Vestibular and Visual Cues in Sound Localisation', *Journal of Experimental Psychology*, 27(4), pp. 339-368.

Wallach, H., Newman, E. B. and Rosenzweig, M. R. [1949]: 'The Precedence Effect in Sound Localisation', *American Journal of Psychology*, 52, pp. 315-216.

Waltz, C. F., Strickland, O. L and Lenz, E, R. [2010]: *Measurement in Nursing and Health Research* (New York: Springer Publishing Company)

Wendt, F., Matthias, M. and Zotter, F. [2014]: 'Panning with height on 2, 3 and 4 Loudspeakers', Proceedings of the 2nd International Conference on Spatial Audio.

Wickelmaier, F. and Schmid, C. [2004]: 'A Matlab Function to Estimate Choice Model Parameters from Paired-Comparison Data', *Behaviour Research Methods, Instruments & Computers*, 26(1), pp. 29-40.

Wightman, F. L. and Kistler, D. J. [1999]: 'Resolution of Front-Back Ambiguity in Spatial Hearing by Listener and Source Movement', *Journal of the Acoustical Society of America*, 105(5), pp. 2841-2853.

Williams, M. [1987]: 'Unified Theory of Microphone Systems for Stereophonic Sound Recording', Audio Engineering Society 82nd Convention, Preprint 2466.

Williams, M. and Le Du, G. [2000]: 'Multichannel Microphone Array Design', Audio Engineering Society 108th Convention, Preprint 5157.

Yost, W.A., Wightman, F. L. and Green, D. M. [1971]: 'Lateralisation of Filtered Clicks', *Journal of the Acoustical Society of America,* 6(20, pp. 1526-1531.