Investigations into the Perception of Vertical Interchannel Decorrelation in 3D Surround Sound Reproduction

Christopher Gribben

Applied Psychoacoustics Laboratory (APL)

School of Computing and Engineering

University of Huddersfield, UK

May 2018

A thesis submitted to the University of Huddersfield in partial fulfilment 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 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, trademarks, and any and all other intellectual property rights except for the Copyright (the "Intellectual Property Rights") and any reproductions 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

The use of three-dimensional (3D) surround sound systems has seen a rapid increase over recent years. In two-dimensional (2D) loudspeaker formats (i.e. two-channel stereophony (stereo) and 5.1 Surround), horizontal interchannel decorrelation is a well-established technique for controlling the horizontal spread of a phantom image. Use of interchannel decorrelation can also be found within established two-to-five channel upmixing methods (stereo to 5.1). More recently, proprietary algorithms have been developed that perform 2D-to-3D upmixing, which presumably make use of interchannel decorrelation as well; however, it is not currently known how interchannel decorrelation is perceived in the vertical domain. From this, it is considered that formal investigations into the perception of vertical interchannel decorrelation are necessary. Findings from such experiments may contribute to the improved control of a sound source within 3D surround systems (i.e. the vertical spread), in addition to aiding the optimisation of 2D-to-3D upmixing algorithms.

The current thesis presents a series of experiments that systematically assess vertical interchannel decorrelation under various conditions. Firstly, a comparison is made between horizontal and vertical interchannel decorrelation, where it is found that vertical decorrelation is weaker than horizontal decorrelation. However, it is also seen that vertical decorrelation can generate a significant increase of vertical image spread (VIS) for some conditions. Following this, vertical decorrelation is assessed for octave-band pink noise stimuli at various azimuth angles to the listener. The results demonstrate that vertical decorrelation is dependent on both frequency and presentation angle – a general relationship between the interchannel cross-correlation (ICC) and VIS is observed for the 500 Hz octave-band and above, and strongest for the 8 kHz octave-band. Objective analysis of these stimuli signals determined that spectral changes at higher frequencies appear to be associated with VIS perception – at 0° azimuth, the 8 and 16 kHz octave-bands demonstrate potential spectral cues, at ±30°, similar cues are seen in the 4, 8 and 16 kHz bands, and from ±110°, cues are featured in the 2, 4, 8 and 16 kHz bands. In the case of the 8 kHz octave-band, it seems that vertical decorrelation causes a 'filling in' of vertical localisation notch cues, potentially resulting in ambiguous perception of vertical extent. In contrast, the objective analysis suggests that VIS perception of the 500 Hz and 1 kHz bands may have been related to early reflections in the listening room.

From the experiments above, it is demonstrated that the perception of VIS from vertical interchannel decorrelation is frequency-dependent, with high frequencies playing a particularly important role. A following experiment explores the vertical decorrelation of high frequencies only, where it is seen that decorrelation of the 500 Hz octave-band and above produces a similar perception of VIS to broadband decorrelation, whilst improving tonal quality. The results also indicate that decorrelation of the 8 kHz octave-band and above alone can significantly increase VIS, provided the source signal has sufficient high frequency energy. The final experimental chapter of the present thesis aims to provide a controlled assessment of 2D-to-3D upmixing, taking into account the findings of the previous experiments. In general, 2D-to-3D upmixing by vertical interchannel decorrelation had little impact on listener envelopment (LEV), when compared against a level-matched 2D 5.1 reference. Furthermore, amplitude-based decorrelation appeared to be marginally more effective, and 'high-pass decorrelation' resulted in slightly better tonal quality for sources that featured greater low frequency energy.

ACKNOWLEDGEMENTS

Firstly, I would like to express many, many thanks to my supervisor, Dr Hyunkook Lee, for all his help and advice over the years. Without his guidance and experience, I feel it would have been much harder to navigate through such a comprehensive project, and I am incredibly grateful for all the time that he has spent working with me.

Thanks to my parents for the visits, phone-calls and sustained support throughout; to my brothers, Michael and Dominic, for the catch-ups that were a welcome distraction; and to Ria, for her unending patience and words of encouragement, particularly in the closing stages(!).

To all the members of the Applied Psychoacoustics Laboratory (APL) at the University of Huddersfield, thank you for participating in the listening tests and sharing your fascinating work – seeing what everyone is working on and how it all relates together inspires me no end. Particular thanks go to Rory Wallis, Tom Robotham, Dale Johnson, Mark Wendl, Ben Evans, Jade Clarke, Oli Larkin, Maksims Mironovs and Connor Millns, for their help with various things and insightful discussions (work-related or otherwise). Further thanks to anyone else who sacrificed their time to sit a listening test for me over the years, it was honestly greatly appreciated.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

ACRONYMS

2D	Two-dimensional
3D	Three-dimensional
A/D	Ambience to Direct Sound SPL Ratio
ASW	Apparent Source Width
B/F	Back-Front Energy Ratio
BRIR	Binaural Room Impulse Response
C	Centre Loudspeaker
CF	Complementary Comb-Filter Method
DAW	Digital Audio Workstation
dB	Decibel
ER/D	Early Reflection to Direct Sound Energy Ratio
F/B	Front-Back Energy Ratio
FFT	Fast-Fourier Transform
FIR	Finite Impulse Response
FL	Front Left
FR	Front Right
HIS	Horizontal Image Spread
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function
Hz	Frequency in Hertz
IAC	Interaural Cross-Correlation
IACC	Interaural Cross-Correlation Coefficient
IACF	Interaural Cross-Correlation Function
ICC	Interchannel Cross-Correlation
ICCC	Interchannel Cross-Correlation Coefficient
ICCF	Interchannel Cross-Correlation Function

ICLD	Interchannel Level Difference
ICTD	Interchannel Time Difference
IFFT	Inverse Fast-Fourier Transform
ILD	Interaural Level Difference
ITD	Interaural Time Difference
kHz	Frequency in Kilohertz
KP	Kendall Phase Method
L	Left Loudspeaker
LAeq	Equivalent Continuous Sound Level
LED	Light Emitting Diode
Ls	Left Surround Loudspeaker
LEV	Listener Envelopment
LF	Lateral Energy Fraction
LUFS	Loudness Units relative to Full Scale
MAA	Minimum Audible Angle
MAD	Median Absolute Deviation
PR	Phase Randomisation Method
R	Right Loudspeaker
RL	Rear Left
RR	Rear Right
Rs	Right Surround Loudspeaker
RMS	Root Mean Square Energy
SPL	Sound Pressure Level
STFT	Short-time Fourier Transform
TQ	Tonal Quality
VIS	Vertical Image Spread
ZA	Zotter Amplitude Method
ZP	Zotter Phase Method

0 INTRODUCTION

0.1 Background to the Research

Over recent years, three-dimensional (3D) loudspeaker reproduction has been of much interest in the audio industry. With the advancement of 3D audio comes the problem of the how best to provide content for it. A practical approach would be to 'upmix' existing or legacy 'two-dimensional' (2D) content, in order to generate new signals for the additional channels. Upmixing is the process of taking source signals of a fixed channel count, extracting information from them (typically the ambient part of the signal(s)), then using this to create new auditory feeds. To ensure that the new signals are incoherent (uncorrelated), a process called decorrelation can be applied, so that the signals sound sonically similar, yet are perceptually different (Kendall, 1995). If coherent (correlated) signals were used during upmixing (i.e. duplicating the original signal), it would result in a focused auditory phantom image, due to a phenomenon known as summing localisation (Blauert, 1997). In other words, interchannel decorrelation is necessary to maintain or achieve the spatial impression that is desired from multichannel reproduction.

Without some decorrelation of signals, it is likely that the sound scene would sound very unnatural and unstable, particularly when the listener is located away from the 'sweet-spot'. This is due to destructive comb-filtering and phase cancellation that can occur when multiple coherent signals are not strictly time-aligned or in-phase. Phenomena such as the precedence effect can also have an impact on auditory imaging; that is, where two similar signals from different locations are slightly misaligned (> 1 ms) and the auditory event is only heard from the location of the earlier signal (Blauert, 1997; Litovsky, Colburn, Yost & Guzman, 1999). Consequently, the ideal listening position or sweet-spot when presenting coherent signals is the point where all signals are in total alignment (i.e. a very restricted area). On the other hand, if these signals were decorrelated, the acceptable sweet-spot would likely increase due to a reduction of the cancelling effects that occur when correlated signals are out-of-phase.

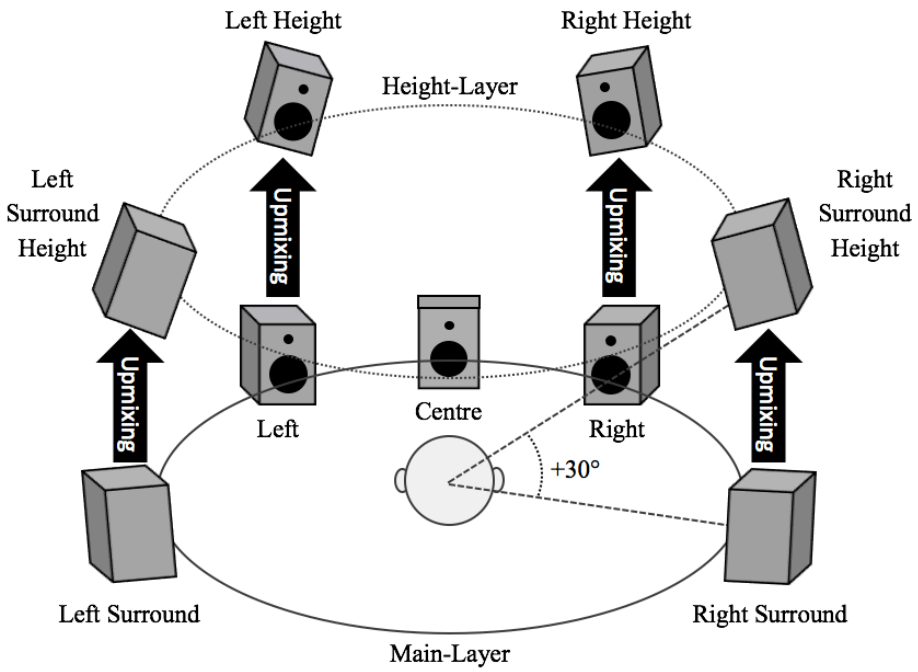

Figure 0.1 Diagram of the potential 2D-to-3D upmixing processing, using 5.1 Surround and Auro-3D 9.1 (Auro Technologies, 2015a) as an example.

Many algorithms based on the concept of ambience extraction and interchannel decorrelation have already been proposed for upmixing from two-channel stereophony (a pair of left and right loudspeakers) to 5.1 Surround (Irwan & Aarts, 2002; Jot & Avendano, 2003; Avendano & Jot, 2004; Li & Driessen 2005). However, consideration must now be placed on applying similar principles in the third dimension. The vertical aspect of commercial 3D systems is commonly achieved by introducing height-channel loudspeakers (the height-layer) above main-channel loudspeakers (the main-layer), where the main-layer is often based on 5.1 or 7.1 Surround. An example of this is Auro-3D 9.1 (Auro Technologies, 2015a), where four height channels are positioned directly above the Left (L), Right (R), Left Surround (Ls) and Right Surround (Rs) main-layer channels of 5.1 Surround (at an elevation angle of +30°). This is represented in Figure 0.1 above, where the upwards arrows indicate potential upmixing from main-layer channels to their respective height-channels. If the principles of the existing upmixing algorithms are followed, this would mean vertically decorrelating extracted ambient information from the L, R, Ls and Rs channels. However, the effectiveness of such a process is yet to be formally tested.

The understanding of interchannel decorrelation perception in the horizontal domain is well established. Blauert (1997) states that when two partially coherent (decorrelated) signals are presented from a spaced pair of left and right loudspeakers, a decrease of the interchannel cross-correlation coefficient (ICCC) (the similarity between the loudspeaker signals) directly relates to an increase of perceived horizontal spread for the phantom image. This is represented on the left of Figure 0.2 below, where it is seen that an ICCC of 1.0 (full correlation between the signals) results in a focused phantom image in the centre of the loudspeakers – then with decorrelation, the image is extended towards the loudspeaker positions. This effect has been demonstrated both over loudspeakers and headphones, where the degree of ICCC relates directly to the extent of the image – that is, as ICCC decreases, the horizontal image spread (HIS) increases almost linearly (Zotter & Frank, 2013; Blauert & Lindemann, 1986b).

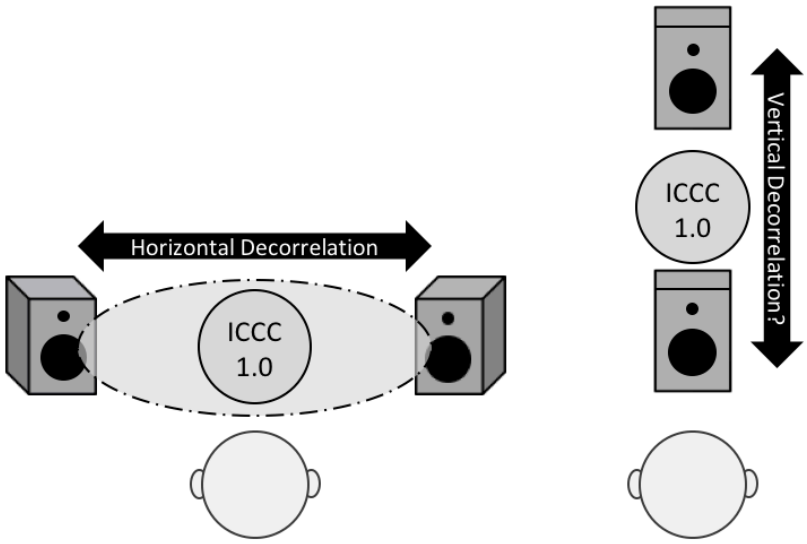

Figure 0.2 Visualisation of horizontal and (potential) vertical interchannel decorrelation perception. ICCC 1.0 is full correlation between the two loudspeaker signals, producing a narrow auditory phantom image. (Left – Horizontal Image Spread (HIS); Right – Potential Vertical Image Spread (VIS))

The effect of horizontal decorrelation is due to the direct influence the ICCC has on the interaural cross-correlation coefficient (IACC) (the degree of similarity between the ears) – where a lower IACC is shown to be associated with apparent source width (ASW) within a reverberant environment (Hidaka, Beranek & Okano, 1995). This relationship between IACC and ASW is

thought to be dictated by early lateral reflections (< 80 ms) and has been investigated extensively in the context of concert hall acoustics (Barron & Marshall, 1981; Hidaka et al., 1995). The greater the decorrelation generated when early reflections are summed at the ear, the greater the sense of ASW – studies have shown that an increase of this effect relates to both listener preference and the quality of concert halls (with IACC being proposed as a good indicator of concert hall quality) (Schroeder, Gottlob & Siebrasse, 1974; Hidaka et al, 1995).

Given that the perception of horizontal decorrelation is well-known, it is of interest to observe whether a similar (if any) effect occurs when decorrelating in the vertical domain. A potential perception of vertical interchannel decorrelation is illustrated on the right of Figure 0.2, based on the known perception of decorrelation horizontally. Here it is hypothesised that an ICCC of 1.0 may result in an elevated phantom image between the two loudspeaker positions, with decorrelation increasing the vertical image spread (VIS) both upwards and downwards. Previous studies have already demonstrated that a phantom image can be perceived between two vertically-arranged loudspeakers, when the loudspeaker signals are coherent (correlated) (Pulkki, 2001; Barbour, 2003; Mironovs & Lee, 2016) (Section 2.2.2). However, it is yet to be seen whether decorrelating these signals has any effect on VIS, which is the focus of the experiments described in the present thesis.

It is understood that the perception of HIS from horizontal decorrelation is strongly dependent on interaural differences, relying on the ears being spaced along the horizontal plane. In the case of vertical decorrelation in the median plane (0° azimuth), no such cues would be available. Therefore, if an effect from vertical decorrelation were to be perceived in this instance, it would likely rely on other cues. Research into vertical localisation along the median plane demonstrates that high frequencies (> 3 kHz) are particularly important for accurate elevation perception, where spectral cues are obtained from high frequency filtering at the pinna (Roffler & Butler, 1968a; Hebrank & Wright, 1974). It is possible that the perception of vertical decorrelation may also depend on such cues, where high frequency spectral notches could be altered by multiple uncorrelated signals arriving at the ear. As a result, it is important that the present

study assesses a potential frequency-dependency of vertical decorrelation. Furthermore, 3D surround sound systems often feature vertically-arranged pairs of loudspeakers at multiple positions around the listener, which is likely to have an impact on both the interaural level difference (ILD) and interaural time difference (ITD); therefore, it is also necessary to consider the azimuth angle presentation of vertical decorrelation during the scope of this project.

With the advancement of 3D surround sound systems, it is expected that interchannel decorrelation in the vertical plane will be utilised for many applications. In addition to multichannel upmixing (as described above), this could potentially include controlling the vertical extent of a sound source, as well as the parametric coding of multichannel audio (where spatial information, such as ICCC, is gathered at the encoding stage, then reproduced at the decoding stage). The recent MPEG-H standard (for 3D audio coding) indicates the use of all-pass filter decorrelation for synthesis of ICCC cues when decoding, which suggests that vertical interchannel decorrelation may already be in use (Murtaza et al., 2015). From this, it is considered that a fundamental study into the perception of vertical decorrelation is necessary to improve our understanding – findings from which may contribute to the optimisation of 3D audio applications, such as, the parametric coding of multichannel audio and 2D-to-3D upmixing.

0.2 Research Questions

From the above background, the following research questions are proposed:

1. What is the perceived effect of vertical interchannel decorrelation?
2. Which cues allow for the perception of vertical interchannel decorrelation?
3. Is the perception of vertical interchannel decorrelation frequency-dependent?
4. Does the presentation angle have an impact on vertical interchannel decorrelation?
5. Is vertical interchannel decorrelation effective in 2D-to-3D upmixing applications?
6. How might 2D-to-3D upmixing by interchannel decorrelation be optimised?

Since there are many scenarios where vertical interchannel decorrelation might be useful, it is necessary to limit the scope of the thesis. As a result, the primary focus of the following investigations is on 2D-to-3D upmixing from 5.1 Surround to Auro-3D 9.1 (Auro Technologies, 2015a), as illustrated in Figure 0.1 above. Auro-3D 9.1 features four vertical pairs of loudspeakers at azimuth angles of ±30° and ±110°, where the main-layer loudspeakers are positioned at ear height, and the height-layer loudspeakers are elevated directly above at +30° to the listening position. Taking this into account, the effects of vertical interchannel decorrelation in the present thesis are observed for a +30° elevation angle between a main-channel loudspeaker and a height-channel loudspeaker. Similarly, the azimuth presentation angles assessed in the present thesis are also related to the Auro-3D 9.1 format (±30° and ±110°). A third azimuth condition of 0° (i.e. the median plane) is also considered, as it is thought that this will reveal important cues for the perception of vertical interchannel decorrelation (similar to those that have been identified for vertical localisation in the median plane).

0.3 Thesis Structure

Chapter 1 of the present thesis examines literature regarding the spatial perception of audio – that is, how sound is located and perceived, both as an auditory source, as well as the impression of sound from reflections. The first section looks at the localisation of a point source in the horizontal, vertical and depth dimensions. Following this, the inherent extent of sound and its dependencies are considered. The next section looks at the spatial impression of sound within an enclosed space, which typically relates to concert hall acoustics. Objective measures for quantifying spatial attributes are then described towards the end of the chapter.

In Chapter 2, the spatial control of audio within loudspeaker reproduction systems is explored. The first section describes both 2D and 3D loudspeaker formats that are in use today. The next section looks at controlling the position of an auditory phantom image by 'panning', both horizontally and vertically. Following this, controlling the extent of a phantom image by interchannel decorrelation is described, along with various approaches for achieving this. Lastly, existing multichannel upmixing algorithms are discussed, as well as considerations towards the use of upmixing for 3D surround sound reproduction.

Chapter 3 describes a two-part subjective experiment in which horizontal and vertical interchannel decorrelation are compared, using the exact same test stimuli. Three pink noise frequency bands were assessed ('Low', 'Middle' and 'High'), each with various degrees of decorrelation, in order to observe the frequency-dependent effect on both horizontal image spread (HIS) and vertical image spread (VIS). The decorrelation technique used is known as complementary comb-filtering (Lauridsen, 1954; Schroeder, 1958) – different parameters for this method were also assessed during testing, to provide some insight into its operation. The findings are then discussed, taking into consideration the results from both experiment parts.

Chapter 4 consists of two subjective experiments that assess the relative and absolute grading of VIS by vertical interchannel decorrelation, where the test stimuli were octave-band (with centre frequencies of 63 Hz to 16 kHz) and broadband pink noise. Stimuli were presented from

three different azimuth positions (0°, ±30° and ±110°), based on the Auro-3D 9.1 loudspeaker format (Auro Technologies, 2015a). For the relative testing stage, two decorrelation methods were compared with the same interchannel cross-correlation coefficients (ICCCs), in order to observe the ICC effect on VIS at different frequencies. During the absolute test, the spatial positions of the upper and lower VIS boundaries were recorded independently for different conditions. The absolute results supplement the relative test results by revealing the inherent extent of VIS for each frequency band, as well as indicating the absolute change of VIS that occurs when vertical interchannel decorrelation is applied.

In Chapter 5, the test stimuli from the subjective experiments in Chapter 4 have been objectively analysed. The stimuli were convolved with two sets of impulse responses: the first being anechoic head-related impulse responses (HRIRs) from the MIT KEMAR database (Gardner & Martin, 1994), while the second were binaural room impulse responses (BRIRs) captured from the loudspeakers in the listening room. Firstly, spectra of the HRIR-convolved stimuli are inspected for potential spectral cues. Following this, interaural cross-correlation coefficients (IACCs) are calculated for both the HRIR- and BRIR-convolved stimuli. Lastly, the effect of room reflections is considered through analysis of the BRIR signals.

Chapter 6 considers the frequency dependent results of Chapters 4 and 5 by assessing the vertical interchannel decorrelation of high frequencies only. Two experiments are described that assess the effects of 'high-pass decorrelation' – the first looked at the impact on VIS, and the second at the perceived tonal quality (TQ) of stimuli. Complex ambient sound sources were tested to provide a practical context (e.g. upmixing), where the stimuli consisted of high-pass decorrelated conditions with varying high-pass cut-off frequencies. The results are then discussed with respect to the practical implications that the findings produce.

Chapter 7 investigated various approaches of 2D-to-3D upmixing by vertical interchannel decorrelation, where both broadband decorrelation and high-pass decorrelation are considered (based on the findings of Chapter 6). Similar to Chapter 6, two experiments are described that

test for different auditory attributes – the first being listener envelopment (LEV), while the second was TQ (as with Chapter 6). This is followed by detailed objective analysis of the bin-auralised stimuli signals, looking at octave-band RMS energy, the ratio of octave-band energy between the front and back, octave-band IACC and potential spectral cues.

In Chapter 8, a summary of the preceding chapters is presented, featuring the basic experimental method and key findings in each case. Following this, final conclusions and practical implications are discussed with respect to the findings. Lastly, proposed further work is described, which intends to expand on the conclusions drawn from the present thesis.

0.4 Original Contributions

The present thesis provides the following original contributions to knowledge:

- A direct comparison between horizontal and vertical interchannel decorrelation using the same stimuli indicates that the effect of vertical decorrelation (between one loudspeaker at ear height and another elevated by 30°) is weaker than horizontal decorrelation (between two loudspeakers with a base angle of 60°) (Chapter 3).
- Both relative and absolute changes of vertical image spread (VIS) are perceivable from vertical interchannel decorrelation between a loudspeaker at ear height and one elevated by 30° – this effect was observed at azimuths of 0°, ±30° and ±110° (Chapter 4).
- Changes to VIS by vertical interchannel decorrelation are frequency-dependent at octave-band level, with the greatest change seen at higher frequencies (Chapter 4).
- The degree of VIS for broadband and high frequency signals relates to vertical interchannel cross-correlation (ICC), where VIS increases as ICC decreases (Chapter 4).
- The azimuth angle of loudspeakers to the listener has a significant impact on the frequency-dependent perception of vertical interchannel decorrelation (Chapter 4).
- Spectral and interaural cues potentially contribute to the perception of vertical decorrelation, as identified through objective analysis of stimuli signals (Chapter 5).
- The vertical interchannel decorrelation of high frequencies alone can significantly increase VIS between a loudspeaker at ear height and one elevated by 30° – this effect was observed at azimuths of 0°, ±30° and ±110°. Reversely, the vertical decorrelation of low frequencies is not necessary for increasing VIS at the same angles (Chapter 6).

0.5 Publications

0.5.1 Conference Papers

Gribben, C., & Lee, H. (2014). The Perceptual Effects of Horizontal and Vertical Interchannel Decorrelation Using the Lauridsen Decorrelator. Presented at the *136th Convention of the Audio Engineering Society*. Paper Number 9027. (Chapter 3)

Lee, H., Gribben, C., & Wallis, R. (2014). Psychoacoustic Considerations in Surround Sound with Height. Presented at the *28th Tonmeistertagung – VDT International Convention*.

Gribben, C., & Lee, H. (2015). Towards the Development of a Universal Listening Test Interface Generator in Max. Presented at the *138th Convention of the Audio Engineering Society*. Engineering Brief 187. (Appendix A)

Gribben, C., & Lee, H. (2017a). The Perceptual Effect of Vertical Interchannel Decorrelation on Vertical Image Spread at Different Azimuth Positions. Presented at the 142nd Convention of the Audio Engineering Society. Paper Number 9747. (Chapter 4)

0.5.2 Journal Papers

Gribben, C., & Lee, H. (2017b). A Comparison between Horizontal and Vertical Interchannel Decorrelation. *Journal of Applied Sciences, 7*(11), 1202. (Chapter 3)

Gribben, C., & Lee, H. (2018). The Frequency and Loudspeaker-Azimuth Dependencies of Vertical Interchannel Decorrelation on Vertical Image Spread. *Journal of the Audio Engineering Society*, (accepted May 2018). (Chapters 4 and 5)

1 THE SPATIAL PERCEPTION OF AUDIO

Interchannel decorrelation in the horizontal domain is often seen as a way to control the width of an auditory image (Zotter & Frank, 2013). In a two-channel stereophonic loudspeaker setup (Left-Right), two partially coherent signals effectively mimic the arrival of decorrelated reflections at either ear, replicating the sensation of actual or apparent source extension in lateral directions. This effect is only achievable through the ongoing detection of interaural cross-correlation (IAC) within the auditory system, relying on a spaced pair of ears along the horizontal axis (Blauert, 1997). Since no such interaural process is available in the median plane, any perceived effect of vertical interchannel decorrelation is likely to depend on other perceptual cues, for instance, spectral filtering from the pinna (Hebrank & Wright, 1974).

In order to investigate vertical decorrelation fully, it is first important to understand how sound is perceived in space, particularly with relation to vertical localisation and vertical elevation effects. Another aim of the study is to assess the effects of vertical interchannel decorrelation from positions surrounding the listener, as would be the case in 3D surround sound systems – therefore, an understanding of the interaural cues that dictate horizontal localisation is also required. Consequently, this first chapter of the present thesis details the localisation mechanisms of the human hearing system (both horizontally and vertically), the inherent extent of sound, spatial impression of sound within an enclosed space, and objective measures that can be used to quantify spatial attributes.

1.1 Localisation of an Auditory Event

The occurrence of a sound can be defined by two positions in both space and time – these are the position of the object (source) generating the sound, referred to as the sound event, and the perceived location of the sound by the listener, known as the auditory event (Blauert, 1997). Auditory space is the entire area in which a sound or auditory event is possible – the term space here is considered in a mathematical sense, where a distance between a set of points (events) can be defined by spatial coordinates. Since the position of both events can always be placed within space, the perception of sound has been referred to as 'spatial hearing', with the idea that it is ever-present (Blauert, 1997). In other words, there is never 'non-spatial hearing', given that a sound source and the consequent auditory perception always feature some positional information. Considering this, the following sections reflect on the different auditory cues that allow for both horizontal and vertical localisation of auditory events within space.

1.1.1 Horizontal Localisation

In the horizontal domain, sound localisation is believed to be dictated by the 'duplex theory', in which the ear input signals are analysed for both interaural time (phase) difference (ITD) and interaural level difference (ILD) by the human auditory system, as the source signal arrives independently at the two ears (Rayleigh, 1907). When a source navigates laterally (horizontally) around the head with increasing azimuth, the ITD and ILD differences between the two ears also increase. It is thought that an azimuth angle of around 90° invokes the maximum difference of both ITD and ILD. This is due to an increased distance and hence time-delay between the ears (affecting ITD), as well as an increase of head-shadowing at high frequencies for the signal in the contralateral (far) ear which reduces the level (thus increasing ILD). As suggested, the relationship between ITD and ILD is frequency-dependent. In general, the localisation cues for lower frequencies come from differences of phase (ITD), whereas high frequency lateralisation is dictated by differences in level (ILD) (as discussed further in the following sections).

1.1.1.1 Interaural Time Difference (ITD)

With interaural time difference (ITD), the distance between the two ears causes a delay in arrival of the same source signal when the source is not located in the median plane (i.e. off-centre). ITD can be perceived if the phase difference of two signals is below half the wavelength (180°) of the frequency signal (Howard & Angus, 2017). This indicates that ITD is frequency dependent, with a greater sensitivity at lower frequencies where the wavelength is greater. The auditory system has the ability to assess the ITD of individual spectral components within an ear input signal – a phenomenon demonstrated by Sayers (1964) and Toole and Sayers (1965) with tonal click stimuli presented over earphones. Furthermore, Mohamed and Cabrera (2008) demonstrate that ITD can contribute to lateralisation at very low frequencies, with a significant effect seen for 1/3-octave-band stimuli down to 31.5 Hz when presented over headphones.

When phase is delayed above 180° (greater than half the frequency wavelength) it can lead to the perception of two separate auditory events in both space and time. Howard and Angus (2017) suggest a common distance between the ears to be 18 cm, calculating that the maximum time-delay at full lateralisation (+90°) is around 673 μs, which takes into account the path around the head. This results in a maximum frequency where direction can be determined by ITD as 743 Hz i.e. when half the wavelength (180°) is equal to the maximum ITD. As frequency is increased above this point, the relative maximum displacement from the median plane decreases rapidly (Blauert, 1997), that is, full lateralisation at 90° by ITD becomes difficult. It is thought that ambiguities of phase difference can be resolved through head movements for frequencies up to around 1.5 kHz (Moore, 2012); however, lateralisation of higher frequencies by ITD above this point is near impossible. An example given by Moore (2012) explains that a 4 kHz sinusoid tone has a cycle period of 250 μs, and if this tone were presented laterally at 90° azimuth, the ITD would be greater than two whole cycles of the 4 kHz tone (assuming the ITD at 90° is 673 μs), resulting in an ambiguity between signals.

Mills (1958) investigated the minimum audible angle (MAA) for differences in azimuth angle perception with the presentation of tonal stimuli – that is, the smallest lateral degree change at which a change in location is perceived. It was found that tones between 250 Hz and 1 kHz had an MAA of 1° from in front (0° azimuth), which increased rapidly (worsened) as frequency increased from 1 kHz to 1.5 kHz. This suggests a greater ambiguity within the region of 1-1.5 kHz in terms of using ITD cues to determine location. Furthermore, as the azimuth angle increased around the head, the MAA also increased rapidly to a point of being indeterminate at +90° (where ITD is greatest). These results demonstrate the inability of ITD to operate at wide azimuth angles, indicating its primary use as an accurate localisation mechanism for low-mid frequencies from in front of the listener. Furthermore, experiments by Wightman and Kistler (1992) appear to show that the ITD localisation cues at lower frequencies have a dominance over the ILD cues at higher frequencies in broadband stimuli.

The role of ITD discussed so far relates to the fine phase structure of sinusoid signals. Bernstein and Trahiotis (2002) conducted an experiment where high frequency signals were amplitude-modulated, so that the envelopes of sound imitated a low frequency sinusoid wave. When modulated by 64 Hz and 128 Hz, the localisation accuracy of the modulated high frequency signal was comparative to its respective low frequency equivalent (a 63 / 128 Hz sine tone) – this was the case for all three modulated signals tested (4 kHz, 6 kHz and 10 kHz). The results here suggest that ITD may also be used to determine the location of high frequencies, however, it is reliant on envelopes of sound, rather than the fine phase structure.

1.1.1.2 Interaural Level Difference (ILD)

In the case of ILD, the level difference between the two ears is caused by a 'head-shadowing' effect that impacts the signal path of higher frequencies to the contralateral (far) ear. Frequencies of longer wavelength (lower in pitch) are able to travel directly through and diffract around the head with minimal shadowing. Howard and Angus (2017) state a cut-off for the head-shadowing effect can be calculated as the frequency where one third of the wavelength $(1/3\,\lambda)$ is

the distance between the two ears. For a head where the ear spacing is 18 cm, they calculate this cut-off frequency point to be 637 Hz, where only frequencies above are affected by head-shadowing. Furthermore, Grantham (1984) demonstrated that the smallest ILD perceivable to the human auditory system is around 1-2 dB. Their results also seemed to indicate a slight insensitivity around 1 kHz for all subjects tested; that is, slightly more ILD was required to perceive a difference for a 1 kHz stimulus.

Kietz (1953) suggests a level difference of 15-20 dB causes a complete lateral image shift to one side, when audio is presented over headphones. However, this is not the case for auditory sources presented in the free field. The head-shadowing effect is highly frequency dependent in terms of spatial localisation, with greater levels of ILD experienced at higher frequencies. Feddersen, Sandel, Teas and Jeffress (1958) demonstrate that when a real source is located at a distance of ~2.1 m from the head, the ILD peaks around 20 dB for 5 kHz and 6 kHz tones at 90° azimuth; whereas frequencies between 1.8 kHz and 4 kHz have a peak ILD of around ± 10dB. For lower frequencies, the ILD is reduced even further, and a 200 Hz tone shows almost equal energy in both ears, due to the diffraction of the wave around the head. On the other hand, Brungart and Rabinowitz (1999) show that as the source is positioned closer to the head (0.12 m), some ILD is also observed at lower frequencies. This suggests that ILD and ITD can both contribute to the localisation of sound sources at all frequencies; however, it is dependent on the proximity of the source to the receiver and the type of signal that is being produced.

In the context of the current study, the role of head-shadowing (and ILD) should be considered in terms of the loudspeaker positions within 3D surround sound systems. Previous studies mentioned above investigate the contribution of ILD to the lateral localisation of sound sources at ear height. With the introduction of height-channels in commercial 3D surround sound systems, ILD is also likely to play a role in the localisation of sounds vertically. That is, as a loudspeaker is elevated, the signal path to the contralateral (far) ear becomes less impeded by the head, as illustrated in Figure 1.1 below. This can be demonstrated by calculating the difference of HRTF

response between the two ears (ILD) for both the main-channel and height-channel loudspeaker positions – Figure 1.2 shows this, using anechoic head-related impulse responses (HRIRs) taken from the MIT KEMAR (Gardner & Martin, 1994), where the FFTs were calculated using 4096 FFT-points with 1/6-octave spectral smoothing.

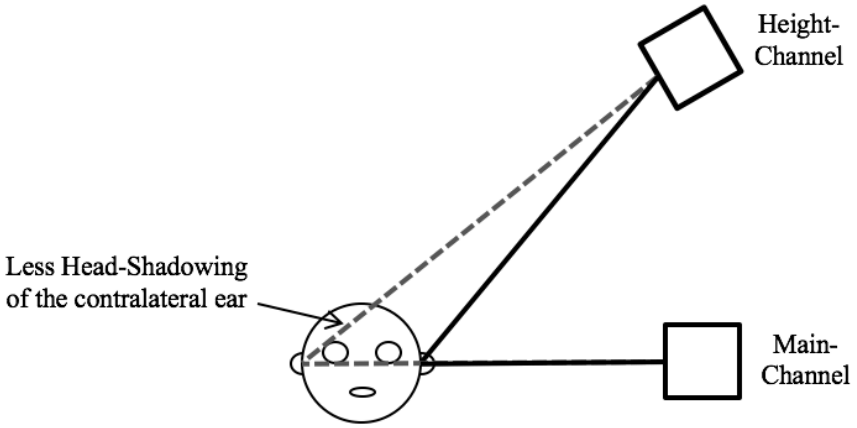

Figure 1.1 Illustrative example of head-shadowing for a main- and height-channel setup

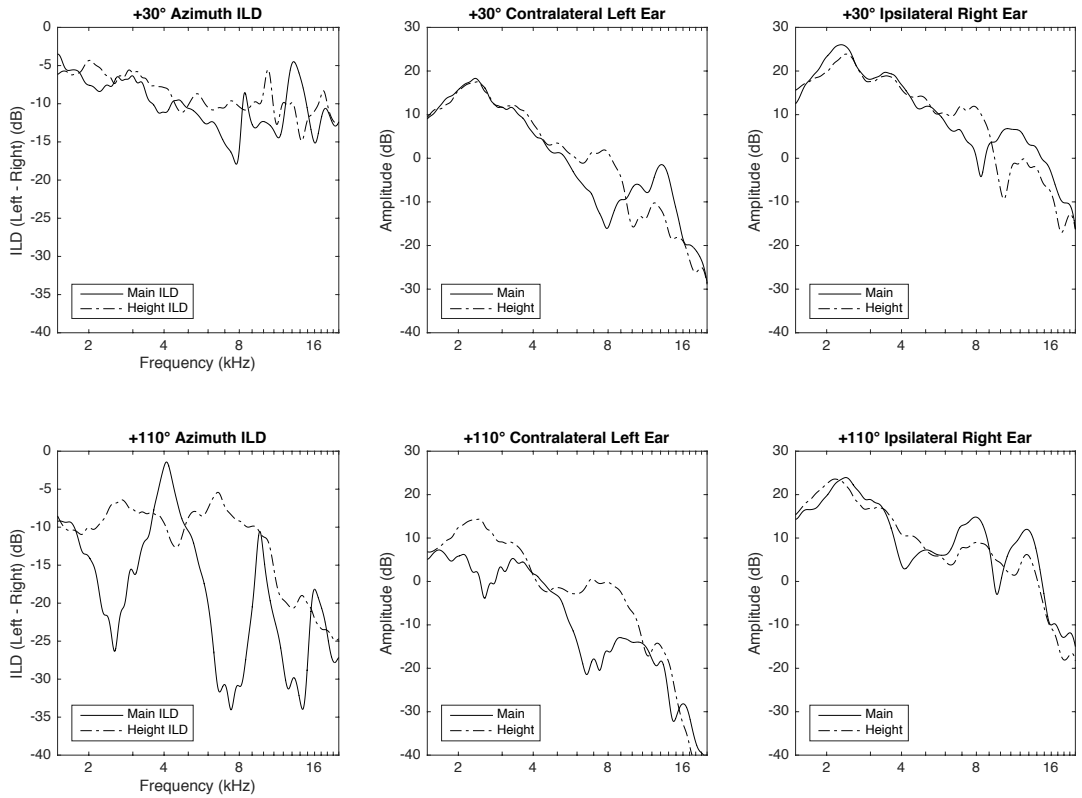

Figure 1.2 Example of the interaural level difference (ILD) between the main-channel and height-channel signals of a 3D surround sound system (based on Auro-3D 9.1 (Auro Technologies, 2015a)).

In the plots above, the first column shows the ILD of the main-channel and height-channel loudspeakers at two different azimuth positions: +30° (top) and +110° (bottom), where the height-channel is elevated by +30° for both. These azimuth angles are the suggested angles of the 3D loudspeaker format Auro-3D 9.1 (Auro Technologies, 2015a), providing a practical context to the discussion. It is seen that the ILD at +30° is broadly similar for both the main- and height-channels (~5-15 dB); however, for the +110° condition where head shadowing is greater, there is a large ILD for the main-channel around 2 - 3.5 kHz, 5-10 kHz and 11-16 kHz (up to ~30 dB). In contrast, the height-channel at +110° shows a considerable reduction of ILD, with a relatively steady ILD of around 10 dB seen up to 10 kHz. The difference of level in the contralateral ear is further reflected in the second column, where the main- and height-channel HRTF responses of the left ear are plotted together for both azimuths. From an azimuth angle of +110°, there is a clear increase of energy for the height-channel condition in the region of 2-4 kHz and 5-10 kHz, presumably due to less head-shadowing.

Considering the literature regarding both ITD and ILD, there is a clear difference between the two in terms of frequency-dependency – where lateral localisation of higher frequencies is generally dictated by ILD, and the localisation of lower frequencies by ITD. Decorrelation techniques tend to work by either manipulating the phase or spectral-amplitude (level) between the two signals (Section 2.3.2), which would suggest that the effectiveness of some decorrelation methods might also be frequency-dependent. Furthermore, an initial observation of the ILD variation between the main- and height-channel conditions indicates that interaural cues should also be considered, particularly when vertically decorrelating signals between main- and height-channel loudspeaker pairs at wide azimuth angles.

1.1.2 Vertical Localisation

Where horizontal localisation relies heavily on interaural differences between the two ear input signals, localising the elevation of an auditory event in the median plane (0° azimuth) is mostly determined by spectral filtering (Roffler & Butler, 1968a). In other words, when sound sources

are presented in the median plane, it results in mostly identical signals at a listener's two ears for all elevations, leaving virtually no interaural cues to aid vertical localisation. To fully explore the perception of vertical interchannel decorrelation, an understanding of these vertical localisation cues is required. Roffler and Butler (1968a) declare that for accurate median plane localisation from the front, the following criteria of auditory stimuli should be met:

1. *The sound must be complex (not a single tone).*
2. *The complex sound must include frequencies above 7 kHz.*
3. *The pinna (outer ear) must be present.*

These prerequisites were initially determined through three experiments into median plane localisation (Roffler & Butler, 1968a). Six stimuli were assessed during the first experiment: 1) broadband noise, 2) low-pass filtered noise (< 2 kHz), 3) high-pass filtered noise (> 2 kHz), 4) high-pass filtered noise (> 8 kHz), 5) a 600 Hz tone burst, and 6) a 4.8 kHz tone burst. Each stimulus was presented from four vertical loudspeakers, positioned with elevation angles of -13°, -2°, 9° and 20° to the listener. The results, displayed in Figure 1.3 below, show that the tones (600 Hz and 4.8 kHz) and low-pass filtered noise (< 2 kHz) have very poor localisation accuracy towards the loudspeaker position. Instead, these auditory events were perceived at more or less the same height, regardless of loudspeaker presentation, demonstrating a general inability to localise singular tones and low frequency sounds. The vertical localisation phenomenon observed here is known as the "pitch-height effect" or "Pratt's effect", where higher pitched tones are perceived consistently higher in space than lower pitched ones (as discussed further in Section 1.1.2.2 below). Roffler and Butler's (1968a) results also show poor localisation of the low-pass stimulus (< 2 kHz), demonstrating that frequencies above the 2 kHz cut-off are particularly important to vertical localisation in the median plane. Having said that, there is also a very slight upward trend for the low-pass stimulus (< 2 kHz) as the height is increased, which would suggest that some weak vertical localisation cues may still exist at lower frequencies – this is possibly influenced by a torso reflection cue, as discussed in Section 1.1.2.4 below.

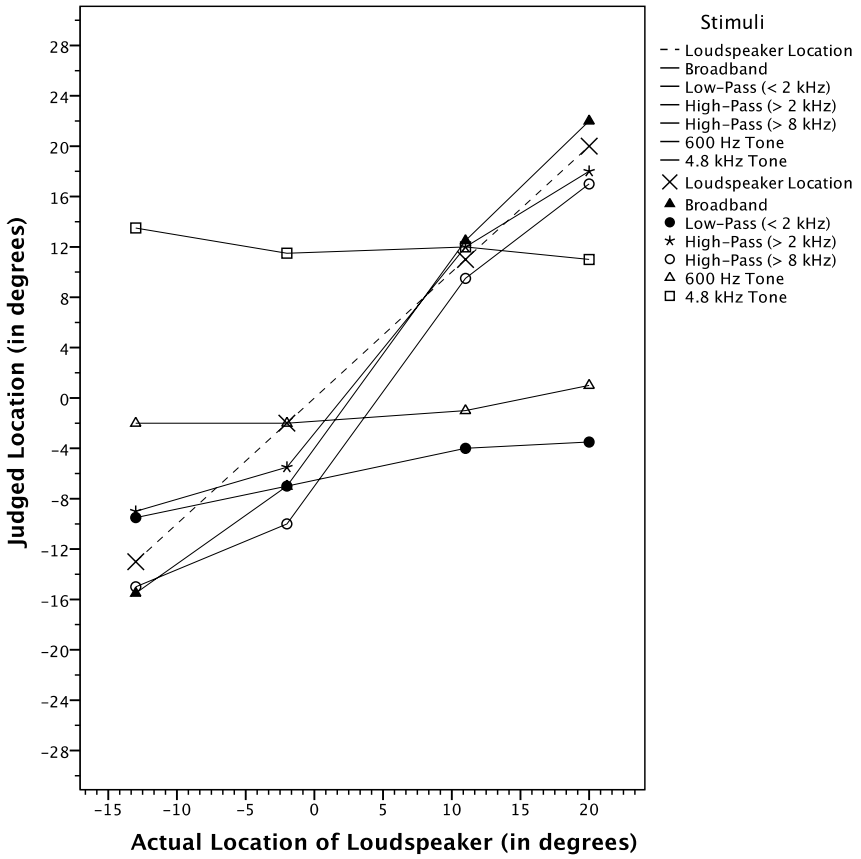

Figure 1.3 Mean judged location at four vertical loudspeaker positions on the median plane (-13°, -2°, 9° and 20°) for tonal and noise auditory stimuli (after Roffler & Butler, 1968a).

A second experiment in the same study (Roffler & Butler, 1968a) aimed to determine the frequency cut-off at which high frequencies become useful for vertical localisation. In this test, noise stimuli with 13 low-pass filter cut-offs (ranging from 500 Hz to 12 kHz) were assessed at the same four heights. It was found that vertical localisation was accurate when the low-pass filter cut-off was 8-12 kHz; however, a 7 kHz cut-off resulted in very poor vertical localisation (i.e. when all frequencies above 7 kHz were excluded). This indicates that spectral activity above 7 kHz is particularly important to vertical localisation in the median plane, hence the prerequisite stated above. Roffler and Butler (1968a) went on to assess the localisation of broadband and high-pass filtered (> 8 kHz) noise, both with and without the pinna present, demonstrating the importance of pinna reflections for localising elevated sources. With the pinna present, localisation was accurate, and without, both stimuli were perceived below ear level (-13°)

for all height presentations (-13-20°). Reasons for Roffler and Butler's (1968a) results (and the subsequent criteria that were set) are explored further in the following sections.

1.1.2.1 Pinna Spectral Filtering

The pinna refers to the entire external element of the human hearing system, before leading into the ear canal. The acoustical function of the pinna provides a linear filter for arriving sound, with the transfer function of the filter depending on the distance and direction of the sound source. The directional frequency response at the ear is referred to as the head-related transfer function (HRTF), which also includes the ITD and ILD cues discussed in Section 1.1.1 (when a source is presented off-centre). These frequency and direction dependent manipulations are what allows for the localisation of a sound source within space. In particular, temporal and spectral filtering at the pinna provides cues for the vertical localisation of a sound when no inter-aural cues are present i.e. in the median plane (Blauert, 1997). There are slight interaural differences between the ears at high frequencies due to asymmetry, which is also thought to contribute to vertical localisation (Searle, Braida, Cuddy & Davis, 1975).

Gardner and Gardner (1973) investigated the occlusion of pinna cavities by testing broadband and 1/2-octave-band noise (with centre frequencies of 2 kHz, 3 kHz, 6 kHz, 8 kHz and 10 kHz). Subjects were asked to identify the loudspeaker that they perceived as reproducing the stimulus signal, with the results reported on an error index of wrong judgements. As occlusion of the pinna was reduced, identification of the active loudspeaker improved for all bands except the 2 kHz 1/2-octave-band – for the 2 kHz band, identification was poor even with the pinna present. In particular, the ability to identify the active loudspeaker was closely comparable between the broadband stimulus and both the 8 and 10 kHz 1/2-octave-bands. This supports Roffler and Butler's (1968a) hypothesis that frequencies above 7 kHz are important for localisation along the median plane. On the other hand, Gardner and Gardner's (1973) results also show relatively accurate responses for the 3 kHz and 6 kHz 1/2-octave-bands, seemingly in contrast with Roffler and Butler (1968a). However, Cabrera et al. (2005) comment that the vertical localisation

accuracy actually improved for the > 2 kHz high-pass condition in Roffler and Butler's (1968a) study (Figure 1.3). In other words, when frequencies between 2-8 kHz were also included (in addition to those above 8 kHz), subjects could more accurately identify the source position, in comparison to the > 8 kHz condition.

The results of Gardner and Gardner (1973) clearly suggest that frequencies below 7 kHz may also contribute to vertical localisation, potentially as low as 3 kHz. Hebrank and Wright (1974) investigated this observation through three experiments, the first of which assessed the vertical localisation of high-pass and low-pass filtered noise stimuli from the median plane, in order to determine the frequency-band(s) where spectral cues exist. The high-pass cut-off frequencies were 3.8, 5.8, 7.5, 10.0, 13.2 and 15.3 kHz, and the low-pass cut-offs were 3.6, 6.0, 7.5, 8.0, 10.3, 12.0, 14.5 and 16.0 kHz. It was found that frequencies below 3.8 kHz and above 16 kHz did not contribute to the vertical localisation. The implication that the spectral cues for vertical localisation lie between 3.8 kHz and 16 kHz is in support of the findings by Gardner and Gardner (1973). Hebrank and Wright (1974) followed up the first experiment by testing various high-pass, low-pass, band-pass and band-stop filtered noise conditions in the median plane, to investigate and understand these spectral cues further. A summary of the results are as follows:

- Low-pass filtered noise with cut-off frequencies between 3.9-8.0 kHz was perceived in front. As the cut-off increased, a slight elevation was observed.
- A low-pass cut-off frequency of 10.3 kHz was perceived from above.
- Noise with a high-pass cut-off at 10.0 kHz tended to be perceived behind, whereas high-pass cut-offs of 13.2 and 15.3 kHz were perceived in front.
- 1/2-octave-band noise with centre frequencies of 4.0-7.2 kHz and 14.5 kHz were perceived in front, whereas 8.1-9.1 kHz were perceived above.
- Noise with spectral notches (band-stop filters) between 7.4-10.8 kHz was perceived in front, while notches of 12.0-17.8 kHz resulted in above perception.

The above results suggest that the perception of height and 'aboveness' is specifically related to frequencies around 8.1-9.1 kHz, whereas localisation towards the front seems to be related to frequencies between 12.0-17.8 kHz. When the higher frequencies are excluded (12.0-17.8 kHz notch conditions and low-pass filter condition with a cut-off at 10.3 kHz), the noise stimuli were all perceived from above. Given that the band-pass conditions with centre frequencies of 8.1-9.1 kHz were also perceived above, this indicates the dominance that height cues in this region (8-10 kHz) can have over lower frequency spectral cues (3-7 kHz). Going back to Roffler and Butler's (1968a) prerequisites of vertical localisation at the beginning of Section 1.1.2, it is stated that accurate localisation in the median plane requires frequencies above 7 kHz. The results here suggest that this remains partly true – it is apparent that vertical localisation cues do exist below 7 kHz, however, they appear to be weaker in comparison to those at higher frequencies, where it is seen that frequencies around 8.1-9.1 kHz seem to take dominance in elevation perception. The association with frequencies around 8 kHz and the perception of 'aboveness' relates to Blauert's (1969/70) work with directional bands, which is discussed further in Section 1.1.2.3 below.

Morimoto, Yairi, Iida and Itoh (2003) further demonstrated the dominance of high frequencies in vertical localisation. Two filtered white noise sources were reproduced simultaneously from spaced positions on the vertical plane, one with a high-pass filter (> 4800 Hz) and the other with a low-pass filter (< 4800 Hz). Subjects were asked to localise the source position, with the results showing a consistent trend towards the position of the high-pass filtered condition. This effect was also investigated by Ferguson and Cabrera (2005), with varying positions of tweeter (high frequency) and woofer (low frequency) signals along a vertical plane. When the two signals were reproduced synchronously, it was shown that the high frequencies of both noise and musical stimuli were localised at the tweeter quite well, whereas localisation of the low frequency woofer signal was consistently poor. Both these studies further suggest that high frequencies have a localisation dominance over lower frequencies, when both are presented simultaneously and in synchrony.

The literature above emphasises the importance of high frequencies in vertical localisation, particularly in the median plane where interaural differences are minimal. In general, it is accepted that the localisation cues for elevation lie above 3 kHz, with a particular contribution from those around 8 kHz. In the present study, it is important to consider the influence of these spectral cues in the median plane, and explore whether such cues exist for the perception of vertical interchannel decorrelation. The literature also demonstrates how the localisation of high frequencies can be independent to those of lower frequencies, with strong directional cues for some frequency bands (i.e. 8 kHz from above). With this in mind, and in the context of the present study, it is necessary to understand the phenomena in which different frequencies are perceived from different directions, which may cause an inherent spread of sound in the vertical domain (e.g. 'directional bands' and the 'pitch-height' effect).

1.1.2.2 Directional Bands

For some bands of frequencies, the perceived localisation of the auditory event is independent of the source position under certain conditions. To demonstrate this, Blauert (1969/70) conducted experiments in an anechoic chamber, assessing the localisation of 1/3-octave-band noise burst stimuli (centre frequencies: 125 Hz to 16 kHz) from five different loudspeaker conditions surrounding the listener. Subjects were asked to localise each noise stimulus within one of three regions on the median plane: 'Front' (345° to 45°), 'Above' (45° to 135°) and 'Behind' (135° to 195°) (where 0° is directly in front of the listener at ear height). If over 50% of subjects were in agreement of the perceived source location of a band (when the responses for each loudspeaker condition were combined), a directional band was consequently defined. From the results, it was found that the bands associated with the 'Front' position were around 500 Hz and 4 kHz; bands around 1 kHz were perceived from 'Behind'; and the 8 kHz band was heard 'Above'. The perception of the 8 kHz band from above is in agreement with the results from Hebrank and Wright's (1974) investigations into pinna spectral cues.

A similar experiment was conducted recently by Wallis and Lee (2015a), in which a greater resolution of response areas was offered to the listener. The regions consisted of eight 45° segments that fully surrounded the listener in the median plane, and were named as follows: 'Front Low', 'Front', 'Front High', 'Above', 'Back High', 'Back', 'Back Low' and 'Below'. Four centre frequencies were assessed, namely within the key directional bands of Blauert's (1969/70) work, which were 500 Hz, 1 kHz, 4 kHz and 8 kHz. For each centre frequency, tests were conducted on its related sinusoid tone, 1/3-octave-band and octave-band, presented both in bursts and continuously. All stimuli were then reproduced by a single loudspeaker in the front and the rear positions during independent trials. In general agreement with Blauert's (1969/70) findings, the results show that the 1 kHz 1/3-octave-band burst stimulus was mostly localised in the 'Rear' region and the 4 kHz 1/3-octave-band from an elevated frontal position, however, neither were observed with the same consistency. A difference between the two methodologies appears to be the restriction of head movements – in Blauert's (1969/70) study, subjects' heads were clamped in place to ensure minimal movement, whereas with Wallis and Lee (2015a), subjects were only instructed to face forward and remain still. It is possible that even the slightest of head movements could have helped to resolve front-back confusion of directional stimuli, in particular for the 1 kHz 1/3-octave-band from the 'Front' and the 4 kHz 1/3-octave-band from the 'Rear'. This is seen with an increased frequency of responses towards the actual source in the direction opposite to that of the known directional band (e.g. 30% for 1 kHz in 'Front' and 45% for 4 kHz at the 'Rear'). If this were the case, it suggests that there might be no perceptual benefit of utilising directional bands in a practical context, as the effect would likely break down considerably under normal listening conditions. On the other hand, the 8 kHz 1/3-octave-band was consistently localised from the 'Above' regions in Wallis and Lee's (2015a) experiment, as with the studies of Blauert (1969/70) and Hebrank and Wright (1974). This provides further evidence that the perceptual effect of elevation with the 8 kHz band is particularly robust – supporting the importance of this high frequency region to vertical localisation, and potentially vertical decorrelation.

1.1.2.3 The 'Pitch-height' Effect

The 'pitch-height' effect (or 'Pratt's effect') is another elevation phenomenon (not dissimilar to directional bands), where tones of a higher frequency are generally perceived higher in space than those of a lower frequency. The effect was first experimentally observed by Pratt (1930), who tested five sinusoid tones at the following octave intervals: 256 Hz, 512 Hz, 1024 Hz, 2048 Hz and 4096 Hz. When the tones were presented from the same height, the average perceived elevation of the 256 Hz tone was 0.7 m, whereas the average elevation of the 4096 Hz tone was 2 m, clearly demonstrating the extent of the effect. These findings were supported by Trimble (1934) who tested nine tones from 500 Hz to 3950 Hz, of which the results also demonstrated a correlation between pitch and height in front of the listener.

Some years later, Roffler and Butler (1968b) confirmed and expanded on the initial findings of Pratt (1930) and Trimble (1934), establishing that the pitch-height effect remained when listeners lay on their backs with the loudspeakers arranged above, as well as when they lay on their side, with respect to elevation above their head orientation. The same effect was also evident among congenitally blind persons and 4- and 5-year-old children in Roffler and Butler's (1968b) work, indicating that the phenomenon is not fundamentally caused by visual cues, nor a semantic association between a high pitch and the height of elevation (though it is accepted that these associations may contribute to enhancing the effect). Roffler and Butler's (1968b) study concludes with support of a notion by Pratt (1930), stating that every tone has an intrinsic spatial character relating to height and depth on a pitch-continuum, and this character is present prior to forming any associative relationships.

More recently, Cabrera and Tilley (2003) examined the vertical localisation of band-limited noise in the median plane, using octave-bands with centre frequencies of 125 Hz, 500 Hz, 2 kHz and 8 kHz, in addition to broadband and low-pass filtered (3 kHz cut-off, -24 dB/oct) pink noise. Stimuli were presented as 200 ms bursts at two loudness levels (84 and 64 phons) from five elevated positions (0°, ±7.9° and ±15.6°). In the experiment, subjects were asked to

define the upper, lower, left and right boundaries of the auditory image, with vertical locations determined from the average centre point between the upper and lower image boundaries. This may have had an impact on the accuracy of responses, since exact locations based on the focal point of the sound were not recorded independently – it is possible that the actual focus differed from the calculated centre of the overall perceptual spread.

Cabrera and Tilley's (2003) results in Figure 1.4 below show that only the broadband pink noise was accurately localised to the centre of all the source positions, with the octave-band noise stimuli demonstrating the pitch-height effect. Of the octave-band stimuli, the 8 kHz band showed the most accurate localisation, particularly for the highest source position (+15.6°). In conjunction with the relatively poor low-pass pink noise scores, these results further support the idea that frequencies within this region are important to the vertical localisation of sources. In terms of the pitch-height effect, the 2 kHz, 8 kHz and the low-passed pink noise frequency bands were all elevated above the lower loudspeaker positions. On the other hand, both the 125 Hz and 500 Hz octave-bands were localised just below ear level (0°) for all conditions, regardless of the presentation height. This implies that the pitch-height effect experienced by Pratt (1930), Trimble (1934) and Roffler and Butler (1968b) with single tones can also apply to band-limited stimuli. Furthermore, Cabrera and Tilley's (2003) results indicate that loudness generally had little effect on localisation. The 84 phon presentation produced some slight elevation for the 8 kHz octave-band, presumably due to an enhancement of the directional band effect discussed in Section 1.1.2.2 above; and 84 phon also slightly improved localisation accuracy towards the source position for the 125 Hz band.

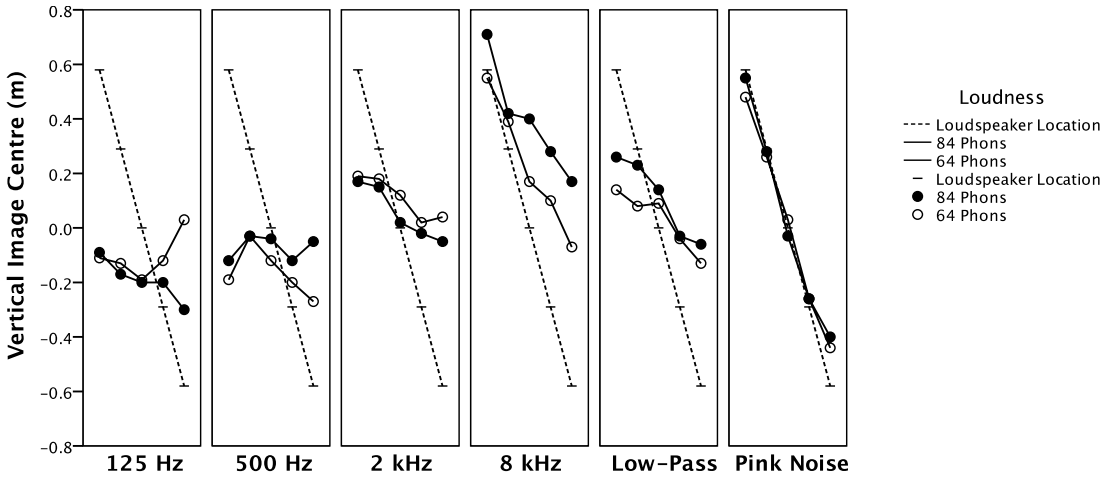

Figure 1.4 Mean vertical localisation results recorded at 5 source position heights, demonstrating the pitch-height effect between octave-band stimuli (after Cabrera and Tilley 2003).

In a following experiment, Ferguson and Cabrera (2005) observed the vertical localisation of two signals concurrently within a vertically-arranged tweeter and woofer pair – high and low frequency band-limited noise were presented simultaneously as 200 ms bursts from the tweeter and woofer respectively, assessing various frequency cut-offs between the two signals. The results showed that, when both signal bursts were synchronous (i.e. time-aligned), the high frequency signal was localised at the tweeter position, whereas the low frequency signal was consistently observed around ear level, regardless of the woofer height. These results are in agreement with Cabrera and Tilley's (2003) 'pitch-height' findings, and demonstrates that the localisation of different frequency bands can somewhat be achieved when presented simultaneously. Ferguson and Cabrera (2005) also found that when the signals were slightly asynchronous (i.e. when a 100 ms delay was applied to the woofer signal), localisation of the woofer signal improved if it featured frequencies up to 1 kHz, but not for the 500 Hz cut-off condition. This seems to support observations made by Algazi, Avendano and Duda (2001), where an independent vertical localisation cue is present around 700 Hz, as generated by a torso reflection that creates a comb-filter effect at the ear input signals (discussed in Section 1.1.2.4 below). It would appear that when both low and high frequency signals are presented synchronously, this localisation cue around 700 Hz may become masked by a dominance of the high frequency

content. Evidence of the pitch-height effect was also seen in Wallis and Lee's (2015a) directional band work (described in Section 1.1.2.2 above), where the 4 kHz 1/3rd-octave-band was perceived from an elevated frontal position, further demonstrating the 'pitch-height' effect with band-limited stimuli.

Lee (2016b) applied the notion of the 'pitch-height' effect to a novel method of upmixing 2D audio content to 3D surround sound systems. He obtained the perceived height (location) of octave-band stimuli (with centre frequencies of 63 Hz to 16 kHz), presented as a coherent stereophonic phantom image between a Left and Right loudspeaker pair (with base angle of 60°), and also a frontal height-channel loudspeaker pair (Left Height and Right Height loudspeakers elevated by 30° to the listening position with a base angle of 60°). The upmixing idea is to discretely route each octave-band to either the main- or height-layer loudspeakers, based on the intrinsic spatial location of the different bands, spreading the sound vertically between the two layers (this approach to upmixing has been discussed further in Section 2.4).

For Lee's localisation results of the main-layer presentation (the white boxes in Figure 1.5), a 'pitch-height' effect is observed from 63 Hz to 500 Hz. However, at 1 kHz, the elevation was significantly reduced. Then from 1 kHz to 8 kHz, the pitch-height effect was observed again, before elevation reducing for the 16 kHz octave-band. Lee (2016b) refers to the 1 kHz octave-band as a "reset" point of the pitch-height effect. The height-layer results show a slightly elevated pattern of the pitch-height effect for all octave-bands (shaded boxes in Figure 1.5), with only the 250 Hz, 500 Hz, 8 kHz, 16 kHz and Broadband frequency bands perceived as significantly elevated above their respective main-layer conditions. It is considered that the lower localisation of the 1 kHz octave-band in both presentations may be due to an inherent localisation from behind for this frequency band, as presented in Blauert's directional band theory (1969/70) (discussed in Section 1.1.2.2 above). Several subjects experienced front-back confusion for the 1 kHz band, which would have likely caused confusion and resulted in the lower judgement.

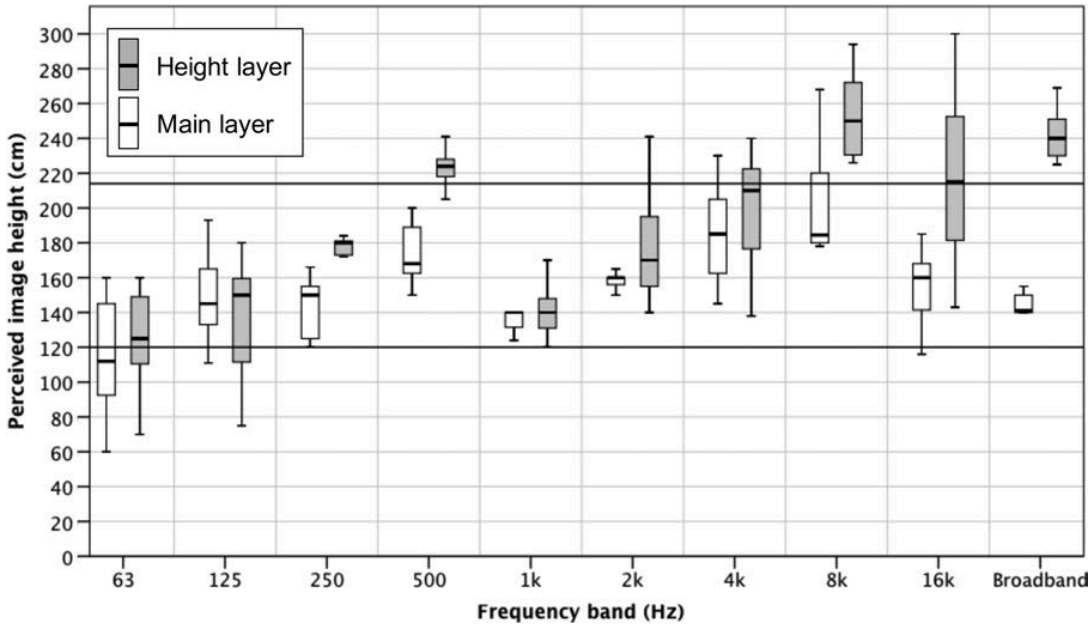

Figure 1.5 Localisation of octave-band stimuli presented as coherent phantom images from both main- and height-layer loudspeaker pairs (left and right) (courtesy of Lee, 2016b).

In comparison to Cabrera and Tilley's (2003) results in Figure 1.4 above, Lee's (2016b) results (Figure 1.5) show the 500 Hz band to be elevated notably higher. He notes that this could be due to the 'phantom image elevation' effect, where a wider base angle between a pair of loudspeakers elevates the coherent phantom image between them. As he hypothesises, this is potentially related to a confusion between interaural cross-talk and torso reflections (discussed further in Section 1.1.2.4 below). Since Lee (2016b) was assessing the localisation of phantom images with a base angle of 60°, compared to Cabrera and Tilley's (2003) experiment where sounds were reproduced by 'real' loudspeaker sources in the median plane, this explanation seems plausible. Another difference between the two studies is the listening environment – Cabrera and Tilley (2003) tested in an anechoic chamber, whereas Lee's (2016b) experiments were conducted in a dry listening room. The inclusion of some room reflections (even slight) may have affected the vertical localisation of sources, particularly with the elevated lower frequency results (125-500 Hz).

The evidence of the 'pitch-height' effect for band-limited stimuli could possibly have an influence on the perception of vertical interchannel decorrelation. The theory behind Lee's (2016b) upmixing method is to spread the frequency bands vertically based on their inherent height (as seen in Figure 1.5). Since it is clear that different bands are localised at different heights when presented from the same position, it is possible that broadband signals may also have a large vertical spread when presented from in front, based on the inherent spatial distribution seen above. If this were the case, then there may not be any perceptual benefit to vertical decorrelation in the median plane (or from other frontal regions); as a result, it is important to test the effects of vertical decorrelation from multiple angles to the listening position, assessing both band-limited and broadband sources.

1.1.2.4 Torso and Shoulder Reflections

In a vertical localisation experiment conducted by Algazi et al. (2001), it was found that elevation could still be perceived with an absence of frequencies above 3 kHz – however, the judgements were generally underestimated from the actual source position. Algazi et al. (2001) hypothesise that this elevation perception is through torso (shoulder) reflections that cause a comb-filtering effect at the ear. This is in contrast with the vertical localisation results discussed in Section 1.1.2.1 above; however, it is thought that the high frequency cues previously discussed (> 3 kHz) take precedence over those at lower frequencies (Morimoto et al., 2003). Binaural analysis of various elevation angles by Algazi et al. (2001) indicate that comb-filter spectral notches (as low as 700 Hz) seem to be associated with elevation. That is, where a greater elevation resulted in a longer delay between the direct signal and torso (shoulder) reflection, causing the frequency of the notch position to decrease. It was observed that similar notches featured in both ears, although the effect was weaker in the contralateral ear. Furthermore, the vertical localisation accuracy improved slightly when the low-pass condition (< 3 kHz) was presented off-centre (presumably influenced by the torso reflections). It may also be that torso cues are easier to detect when there is a secondary ITD cue present (Section

1.1.1.1) – for instance, if the same notches were to occur simultaneously in both ears, the hearing system may assume that the comb-filtering is a feature of the source signal.

An interesting vertical elevation effect investigated by Lee (2017a) has been related to the perception of torso reflections. Known as the 'phantom image elevation' effect, it sees the phantom auditory image (from two coherent signals) elevate upwards in the median plane as the base angle between the two loudspeakers increases. Lee (2017a) hypothesises that the ITD (crosstalk) between both signals arriving at each ear equates to the delay between the direct signal and shoulder reflection for a 'real' elevated source. That is, as a 'real' source is elevated upwards in the median plane, the delay between the direct sound and shoulder reflection at the ear increases. Likewise, as the loudspeaker base angle is increased, the ITD between the ears increases at a similar rate for each loudspeaker signal (as discussed in Section 1.1.1.1). When the signals from both loudspeakers are coherent, it results in a confusion of the interaural crosstalk, mistaking the secondary signal in either ear for a reflection from the torso. If this is indeed the cognitive process that causes phantom image elevation, it demonstrates that torso reflections may also be important for accurate vertical localisation.

1.1.3 Distance Perception

The distance of a sound source can mostly be determined through environmental factors. Rumsey (2001) indicates that a source far away will have more reverberation in a reflective environment, suggesting that the ratio of direct sound to reverberant energy (D/R) is a useful indicator of distance. Furthermore, the brain is able to broadly determine the size of a space, based on the time of arrival between direct sound and subsequent reflections, which also relates depth perception to spatial impression (discussed below Section 1.3). This relationship between distance and reflective energy is demonstrated by Nielsen (1993), where distance perception is assessed in three settings: an anechoic chamber, a classroom and an IEC listening room. The results show that distance was most accurately identified when reflections were present, where a decrease of D/R corresponded with an increase of distance.

On the other hand, Nielsen's (1993) results suggest that the cues for perceived distance in an anechoic or free field space are largely unreliable. In the anechoic chamber, when the source level was kept constant at the listening position for each distance assessed (1-5 m), the source was generally perceived at the same distance each time, regardless of the actual distance. This suggests that some environmental context or reference is required to make accurate distance judgements. In a free field scenario, Blauert (1997) states that intermediate source distances of 3-15 m are determined by the sound pressure level (SPL) of the source. For every doubling of distance in a free field, the SPL falls by 6 dB. As the distance of a source increases above 15 m, higher frequencies become more attenuated than lower frequencies due to absorption along the direct signal path, distorting the source spectrum in addition to the changes of SPL. This absorption and loss of energy is affected by the moisture content of the air and wind speeds, where changes of spectrum above 10kHz (from air absorption) can become audible as little as 15 m from the source. Whereas, distances closer than 3 m have a strong effect on the head-related transfer function, providing spectral changes to the signal, although these changes are different from those experienced at greater distances. As mentioned in the discussion regarding ILD above (Section 1.1.1.2), Brungart and Rabinowitz (1999) found that as the source position decreased in proximity to the head, ILD was affected at all frequencies. This was particularly noticeable for the low frequencies – from a distance of 1 m the low frequency ILD was relatively small (due to diffraction around the head), whereas at a closer distance of 0.12 m, the ILD at low frequencies increased considerably. This low frequency ILD change could provide an additional cue for distance when sources are in close proximity to the head; although it is noted that for such a cue to be effective, the source must be presented from a wide azimuth angle (i.e. where ILD from head-shadowing is present).

When sources are positioned far enough from the head (> 1 m), there appears to be a dependence on reflective energy for distance perception within enclosed spaces. In the context of the present study, it is assumed that typical upmixing applications will use ambient signals to increase the spatial impression of the sound scene (i.e. through a greater coverage of reverberation

from above). Given the dependence of depth perception on the reverberant energy, it may be found that vertical interchannel decorrelation (or 3D upmixing in general) can also increase the perception of distance (or depth) from above. For instance, presenting decorrelated ambient signals from height may mimic the effect of ceiling reflections, potentially giving a greater sense of listener envelopment from a more realistic distribution of sound (Section 1.3).

1.2 The Extent of an Auditory Event

The extent of a sound refers to the apparent volume of its auditory event – that is, the inherent vertical and horizontal spread of the perceived auditory sound image. An aim of the current study is to observe whether vertical spread can be controlled by vertical interchannel decorrelation. In order to do this, it is important to establish the frequency-dependent extent of sound sources in general. For example, it is known that low tones are perceptually broad (Perrott, Musicant & Schwethelm, 1980), therefore, it might be considered that vertical decorrelation is not beneficial for lower frequencies. Furthermore, other factors that affect the extent of a source, such as loudness and duration, have also been reflected on below.

1.2.1 Effects of Frequency on Extent

An investigation by Terrace and Stevens (1962) observed the perceived extent of single tones over headphones, assessing frequencies between 200 Hz and 4 kHz with varying intensities – selection of which was based on previous research conducted by Thomas (1952) (mentioned in Section 1.2.2). It was found that, for the same intensity level, lower frequencies had a larger extent than higher ones, though intensity had an impact on extent as well (see Section 1.2.2). Perrott et al. (1980) also recorded the extent of single tones (with frequencies between 125 Hz and 8 kHz). However, their experiment focused on the perceived width (rather than overall extent) and featured variations of signal duration. All tones were presented at 50 phons over earphones, assessing five time durations for each tone: 0.1, 0.3, 1.0, 3.0 and 10.0 seconds. The results demonstrated that the lowest tone (125 Hz) was perceptually the broadest for each duration, and the highest tone (8 kHz) was the narrowest. An interesting effect was observed where the perceived extent of the tone increased with the signal duration – this is discussed further in Section 1.2.3 below. Both of these investigations suggest that lower frequencies are perceived as broader than higher frequencies of the same intensity.

More recently, Mason, Brookes and Rumsey (2005) also assessed the inherent width of signals over headphones by decorrelating a 3240 Hz amplitude-modulated signal to match the perceived width of critical band signals (centre frequencies: 100 Hz, 200 Hz, 400 Hz, 800 Hz, 1.6 kHz, 3.2 kHz, 6.4 kHz and 12.8 kHz). The horizontal extent of each band was then determined by the required decorrelation of the 3240 Hz signal (see Section 2.3.2 for a general description of decorrelation). Results from their experiment suggest that the 100 Hz signal had the broadest image as it required the greatest decorrelation. A linear decrease was then seen up to the 1.6 kHz and 3.2 kHz signals, where minimal decorrelation was required, followed by a near linear increase of decorrelation again for 6.4 kHz and 12.8 kHz (producing a similar width as the 200-800 Hz signals). This potentially indicates that both low and high frequency bands have inherently broader horizontal image spreads than middle-high frequency bands (1.6 kHz and 3.2 kHz), when presented over headphones. These results differ from the observations of Terrace and Stevens (1962) and Perrott et al. (1980) above, which may be due to the fact that critical bands were tested rather than single tones. The nature of critical bands means that those with a higher centre frequency feature a greater bandwidth than those with a lower centre frequency – as a result, the inclusion of more frequencies for the 6.4 kHz and 12.8 kHz bands may have caused the increased spread that was seen in the results.

Cabrera and Tilley (2003) conducted an assessment of horizontal and vertical extent using five vertically-arranged loudspeakers in the median plane (with elevations of 0°, ±7.9° and ±15.6°). Six frequency conditions were tested: octave-bands with centre frequencies of 125 Hz, 500 Hz, 2 kHz and 8 kHz, as well as broadband and low-pass filtered pink noise (< 3 kHz). The results showed that the lower frequency stimuli (125/500 Hz octave-bands and low-pass filtered pink noise) had a slightly greater horizontal than vertical image spread, whereas the higher frequency stimuli (2 kHz and 8 kHz octave-bands) had a slightly greater vertical than horizontal image spread. In general, as the octave-band frequency increased, the horizontal extent decreased – this is in agreement with the investigations over headphones discussed above (Terrace & Stevens, 1962; Perrott et al., 1980). On the other hand, observing the vertical spread trends, the

greatest spread is still seen for the 125 Hz octave-band. Interestingly, the 8 kHz octave-band also has a greater vertical spread than both the 500 Hz and 2 kHz bands. This effect is likely due to the strong influence of the 8 kHz region that has been identified in vertical localisation experiments (as discussed in Section 1.1.2 above) – torso reflections may have also played a role in the slight increase of the 2 kHz vertical spread over horizontal spread (Section 1.1.2.4; Algazi et al. 2001). For the broadband pink noise stimulus, horizontal and vertical spread was almost identical and similar to that of the 500 Hz, 2 kHz and 8 kHz octave-bands (i.e. smaller than the 125 Hz band). This suggests that the extents generated by higher frequencies may take dominance within broadband signals. The dominance of high frequencies in vertical localisation tasks has also been demonstrated, as discussion in Section 1.1.2.1 (Morimoto et al., 2003; Ferguson and Cabrera, 2005).

1.2.2 Effects of Loudness on Extent

An experiment conducted by Perrott et al. (1980) showed that as the intensity of a 1 kHz tone varied over earphones (between 44-72 dB), so did the perception of the tone's width (where a greater intensity resulted in a greater width). A further trend was observed with the signal duration of the 1 kHz tone, where a shorter duration resulted in a narrow width (as discussed in Section 1.2.3). Terrace and Stevens (1962) assessed the effect of intensity on the perceived extent of single tones, where a similar relationship between intensity and extent was also established. However, unlike Perrott et al. (1980), Terrace and Stevens (1962) assessed the effect of intensity for multiple frequencies (200 Hz to 4 kHz), which were based on equal-volume (extent) contours determined by Thomas (1952). The results validated the equal-volume contours, demonstrating how the extent between tones can broadly be matched by intensity adjustment. For example, the contours suggest that a 1 kHz tone at 90 dB SPL has a similar extent to a 200 Hz tone at 30 dB SPL.

A second experiment by Perrott et al. (1980) used the same 1 kHz stimuli as above, in order to determine what aspects of the sound listeners were using to make their judgement, whilst also

observing whether the changes of extent were vertical as well as horizontal. The experiment assessed whether a 1 kHz test tone was 'softer' or 'louder' than a 1 kHz reference tone, observing varying degrees of intensity and duration. The tones with less intensity were perceived as 'softer', while 'loudness' increased with intensity. These results were similar for all signal duration lengths, suggesting that the perceived loudness has little to do with the 'expanding image effect' (Section 1.2.3). Similar results to the first experiment (on width) were observed for the vertical spread results, where the vertical spread increased as intensity increased (though to a lesser extent than width) – and as with width, this also seemed to be somewhat influenced by the duration of the signal. A reason for a weaker trend with the vertical spread could be related to the presentation method, given that stimuli were reproduced laterally between the two ears over earphones.

In contrast to Perrott et al. (1980), Cabrera and Tilley (2003) assessed the horizontal and vertical spread of stimuli over loudspeakers in the median plane. In their experiment, they assessed five vertically-arranged loudspeakers (elevation angles of 0°, ±7.9° and ±15.6°), looking at the extent of images for all combinations of adjacent loudspeakers – where each centre location featured either one, three or five loudspeakers with incoherent signals in an array. Using the middle centre point (elevation of 0°) as an example: three level-matched combinations of loudspeakers were assessed for 0°, each featuring 1, 3 or 5 loudspeakers of which 0° was the centre, and the location and spread responses were then averaged over these three combinations. Using such an approach somewhat removes the physical size of the loudspeaker as a limiting factor, allowing for the sole assessment of the loudness effect on perceived extent. The results demonstrated that the loudness of the stimuli signals increased both the horizontal and vertical image spread of a sound source for loudness levels of 64 and 84 phons. It remains unclear whether there were any differences in extent based on the number of active loudspeakers. However, it is reported that the subjective ratings were not significantly affected by the loudspeaker number. Since this was the case, it may be found in the present study that vertical decorrelation between two

vertically-spaced loudspeakers also has little effect on increasing vertical image spread, if all stimuli are matched to the same SPL level.

1.2.3 Effect of Signal Duration on Extent

As mentioned above, Perrott et al. (1980) found that the perceived horizontal width of pure tones increased over time, when stimuli were presented over earphones – they refer to this as the 'expanding-image effect'. Signal lengths from 0.1 to 10 seconds were tested, using pure tones between 125 Hz and 8 kHz at octave-band intervals. With relation to pitch, the lower frequency tones were perceived as having a greater extent than the higher tones (similar to the results discussed in Section 1.2.1). As duration of the signal increased, the perceived extent of each signal also increased, with the 10 second 125 Hz tone demonstrating the greatest extent. Change in extent by duration was most apparent for the lowest tones (125 Hz and 250 Hz) – in the case of the 125 Hz tone at 10 seconds the extent was 9.2 inches, which actually exceeded the limits of the earphones used during testing. A following investigation by Perrott and Buell (1982) where tone duration was varied managed to replicate the same observations as above.

1.3 Spatial Impression within Enclosed Spaces

Spatial impression is a general term for the perception of sound within, and the acoustic of, an enclosed environment from the listener's perspective. It is especially related to research in the area of concert hall acoustics, largely in an attempt to identify the characteristics that contribute to a 'good' sounding concert hall, and also how to measure for such attributes (Hidaka et al., 1995). With the advancement of 3D surround sound systems, one aim is to more accurately reproduce the spatial impression of 'real' environments (such as concert halls); in particular, the handling of sound in the vertical dimension and from above the listener. It is thought that vertical interchannel decorrelation may be useful for this purpose, potentially through the process of upmixing ambience into height-channel loudspeakers i.e. reproducing uncorrelated reflections from height. As a result, the following sections look at the common understanding of spatial impression within enclosed spaces, and how it might relate to spatial perception in the present study.

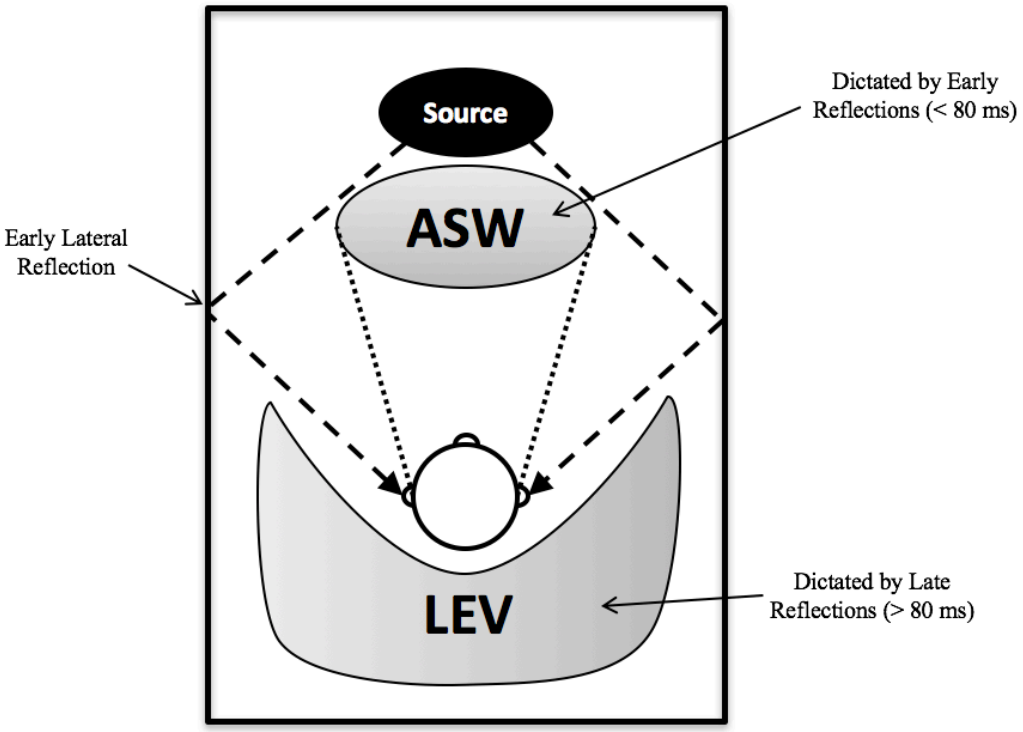

Figure 1.6 Illustrative example of the two attributes that contribute to spatial impression: apparent source width (ASW) and listener envelopment (LEV)

In the past, spatial impression has generally been described as a sense of 'source broadening' (Barron, 1971; Barron & Marshall 1981) and 'spaciousness' (Blauert & Lindemann, 1986a). More recently, it is accepted that spatial impression can be divided into two contributing attributes: apparent source width (ASW) and listener envelopment (LEV) (Bradley & Soulodre, 1995; Hidaka et al., 1995; Morimoto, 2002). ASW relates to the extent of the perceived source image at the performance area, whereas LEV is the impression of the room surrounding the listener – Figure 1.6 above illustrates this. In the remainder of the current section, these terms and related studies have been explored and discussed further.

1.3.1 Apparent Source Width (ASW)

ASW is dictated by the energy of the early lateral reflections (within 80 ms) that arrive at the ears after the direct sound (Rumsey, 2001). This was demonstrated in detail by Barron and Marshall (1981) who assessed the effect of a single pair of symmetrical lateral reflections on the perceived ASW, varying the angle of arrival, intensity, frequency composition and time delay of the reflected signal. Various experiments were conducted to assess these different parameters independently, presenting the direct sound and reflections over loudspeakers in an anechoic chamber. To test the effect of the reflection angle, reflections were produced through symmetrical pairs of loudspeakers with azimuth angles of ±10°, ±40°, ±60°, ±90°, ±140° and ±160°. It was found that as the reflection angle increased up to ±90°, so did the sense of ASW, before reducing again as the azimuth angle increased at the rear. With regard to the delay time, delays between 5-90 ms were presented through a ±40° loudspeaker pair. It was seen that the greatest change of ASW was observed between 5-10 ms, with little change between 10-90 ms, demonstrating that lateral reflection delays as short as 10 ms can have a significantly positive effect on increasing ASW. Given that little change is observed over the entire early reflection region (up to 80 ms), they propose an indicator of ASW as the lateral energy fraction (LF), which calculates the ratio of lateral early reflection energy to the total direct sound and early reflection energy at the listening position (see Section 1.4.1).

Blauert and Lindemann (1986a) also observed the effect of early reflections on ASW. However, the direct sound and early lateral reflections (from ±45° and ±90°) were presented as binaural signals over headphones (as recorded from loudspeakers). The results showed that the greater the spaciousness (ASW) of a stimulus, the more likely it is to be preferred. Blauert and Lindemann (1986a) state that the main contributor to the perception of ASW is the level of early lateral reflections, which is in agreement with the findings of Barron and Marshall (1981). From this, it might be assumed that the greater the lateral reflective energy in a concert hall, the more preferable it is to the listener. Furthermore, it was found that lateral reflections that only contained frequencies below 3 kHz affected the front-back depth of perception rather than ASW; whereas reflections that featured frequencies above 3 kHz gave more prominence to ASW. This finding suggests the importance of high frequency lateral reflections on the perception of ASW in concert halls.

In the study of Barron and Marshall (1981), it was seen that the angle of the reflections had a notable impact on ASW, peaking around ±90°; however, there appeared to be little significant difference between reflections arriving at ±90° and those arriving at ±40°, ±60° and ±140°. From this, a recent study by Johnson and Lee (2017) observed the effect of a single lateral reflection, in order to determine the angle at which maximum ASW is reached. The results indicate this angle to be around 39° in front and 134° from behind, suggesting that lateral reflections arriving within this region (39-134°) will produce a similar perception of ASW (i.e. maximum ASW) – as was implied in Barron and Marshall's (1981) results.

Hidaka at el. (1995) present a detailed assessment of another measure for predicting ASW known as the interaural cross-correlation (IAC). It is based on the degree of similarity between the two ear input signals, as dictated by multiple early reflections from different directions, where a greater energy of reflections decreases the correlation. This further emphasises the effect of early lateral reflection energy on perceived ASW, as suggested by Blauert and Lindemann (1986a). The value obtained from the IAC function (IACF) is called the IAC coefficient

(IACC) – calculation of which is described in Section 1.4.2 below. A similar measure was first proposed by Keet (1968) who calculated the cross-correlation between two cardioid microphones facing laterally, rather than the ears, demonstrating that the coefficients obtained linearly related to subjective ASW judgements. Findings from Schroeder et al. (1974) also demonstrated that a lower interaural 'coherence' (IACC) from early reflections corresponded with an increased preference in concert halls. Some years later, Okano, Hidaka and Beranek (1994) observed the relationship between ASW and the IACC for individual octave-bands (with centre frequencies of 125 Hz to 4 kHz). ASW contours were determined from these octave-band measurements, indicating that the IACC for octave-bands of 500 Hz and above strongly aligned with the perceived ASW (shown in (Hidaka et al., 1995)). Based on the results of Okano et al. (1994), and since IACC / ASW appear to be associated with preference (Schroeder et al., 1974; Blauert & Lindemann, 1986a), Hidaka et al. (1995) aimed to refine the calculation of IACC, in order to use it as a general measure of acoustic quality within concert halls. They concluded that the IACC of the direct sound and early reflections (0-80 ms), taken as the average of the 500 Hz, 1 kHz and 2 kHz octave-bands, was the best predictor for ASW and acoustic quality, referring to the measure as $IACC_{E3}$. It was also shown that the general sound pressure level (SPL) of low frequencies contributed to the perception of ASW, suggesting that this measurement should be used in conjunction with $IACC_{E3}$ to improve the prediction. The contribution of low frequencies relates back to the discussion regarding the inherent extent of frequencies in Section 1.2.1, where it is seen that lower frequencies are perceptually broader than higher ones. Hidaka et al. (1995) also compare the results of $IACC_{E3}$ for real concert halls against measurements of LF, indicating that $IACC_{E3}$ better aligned with the subjective judgements of quality made by experienced professionals.

Lee (2012) conducted an assessment of ASW at different source-listener distances within a concert hall, where a decrease of ASW was seen as the distance increased. His results demonstrated that the IACC of early reflections ($IACC_{E}$) also decreased slightly as the distance increased i.e. the ears became less correlated. This is in contrast with the findings of Hidaka et

al. (1995), where it is considered that a lower IACC generally results in a greater ASW. Furthermore, Lee (2012) showed that the strength of early reflections decreased as distance increased, which strongly related to the perception of decreasing ASW. Comparing these results with those of Barron and Marshall (1981) and Hidaka et al. (1995), it would suggest that further source-listener distances lack enough early (lateral) reflection energy to generate an enhanced perception of ASW, despite the decrease seen in IACC. This result implies the importance of the relative early reflection energy in terms of perceiving ASW – similar to the association identified by Barron and Marshall (1981) and Blauert and Lindemann (1986a), where ASW is reliant on sufficient early lateral reflection energy.

Using the known association between IACC and ASW as a foundation, the effect of increasing horizontal width can be achieved by decorrelating a signal between two spaced loudspeakers (left and right). That is, imitating the natural decrease of correlation that occurs between the ears when multiple reflections are summed (with sufficient early reflection energy). A full description of decorrelation and potential techniques for achieving this can be found in Section 2.3.2 of the present thesis. Blauert and Lindemann (1986b) observe the effects of varying signal correlation over headphones, and Zotter and Frank (2013) over loudspeakers, both of which clearly demonstrate that lower degrees of correlation increase the perceived width (horizontal image spread) of the sound. This horizontal decorrelation effect provides the basis for the present study, since it is not currently known whether interchannel decorrelation between a pair of vertically-arranged loudspeakers can be perceived – and if it can, whether it allows for the dynamic control of vertical image spread. Taking the above background of ASW into account, it might be assumed that vertical decorrelation can replicate vertical (ceiling) early reflections and generate a height effect (as considered in Section 1.3.3 below). Having said that, there appears to be a dominance from lateral (wall) reflections on spatial impression, which is likely due to the ears being spaced along a horizontal plane. For vertical decorrelation to be perceivable, particularly in the median plane, it is thought that some reliance on spectral cues would be necessary (similar to those used for vertical localisation (Section 1.1.2.1)).

1.3.2 Listener Envelopment (LEV)

Listener envelopment (LEV) is the sensation of feeling immersed by sound from all around, and is thought to be largely dictated by late reflections (> 80 ms) within reverberant rooms (such as concert halls) (Rumsey, 2001). Over the years, many attempts have been made to define objective predictors of LEV in concert halls. Bradley and Soulodre (1995) found an agreeable trend between LEV and the late lateral sound strength (> 80ms) (GL_{80}^{∞}) (Section 1.4.3), particularly when averaging the GL_{80}^{∞} over octave-bands of 125 Hz to 1 kHz. A study by Lee (2012) found that the sense of LEV decreased as the distance of the listening position from the source was increased in a concert hall. This appeared to be related to a decrease of the total late sound strength (G_{80}^{∞}), rather than the strength of late lateral sound (GL_{80}^{∞}) (as determined by Bradley and Soulodre (1995)). Lee (2012) also found that measurement of Front-Back energy ratio (F/B) of the late reflections (Section 1.4.4) produced a linear relationship with LEV; whereas the interaural cross-correlation coefficient of late reflections ($IACC_L$) demonstrated a weaker trend. These results suggest that while LF (the ratio of overall to lateral sound energy) and IACC are able to predict ASW reasonably well in concert halls (Hidaka et al., 1995), lateral reflections may only account for LEV under particular test conditions (i.e. certain positions / distances within the concert hall).

As with Lee (2012), Morimoto and Iida (1998) demonstrated that LEV could be quantified by the Front-Back energy ratio (F/B) (Section 1.4.4). However, rather than measure F/B in an existing concert hall, stimuli were artificially created with varying degrees of F/B. In their experiment, six loudspeakers were positioned around the listener, spaced at ±45° azimuth. Both the F/B of early reflections (including direct sound) and late reflections were assessed independently. The early reflection stimuli featured delayed signals between 20-65 ms to represent the reflections, and the late reflection stimuli had delayed signals between 80-104 ms, both of which featured three degrees of F/B. It was found that the synthesised F/B stimuli strongly correlated with LEV for both the early and late reflection conditions.

In loudspeaker reproduction, the inclusion of multiple loudspeakers that surround the listener can increase the sensation of LEV (Griesinger, 1999). With a multichannel loudspeaker system, such as 5.1 surround sound (ITU-R, 2012), the additional loudspeakers from the rear can be used to generate (or replicate) the late reflections that are thought to dictate LEV in concert halls i.e. ambient signals. Leading on from the work of Bradley and Soulodre (1995), Soulodre, Lavoie and Norcross (2002, 2003) tested the effectiveness of using the relative late lateral sound strength (GL_{80}^{∞}) to predict the LEV of the reproduced sound field. The results indicated a similar relationship between GL_{80}^{∞} and LEV to that experienced in concert halls, with an average of GL_{80}^{∞} between the 125 Hz to 1 kHz octave-bands showing the strongest correlation (as with Bradley and Soulodre (1995)). They also consider the effect of the listening room acoustic environment on the perception of LEV – the control parameter during their investigation was the relative measurement of early-arriving energy versus late-arriving energy (C_{80}), and note that the lower limit of C_{80} is dictated by the impulse response of the room. Although the energy of late reflections appeared to be the primary indicator of LEV in their investigation, secondary factors of playback level and the reverberation time also contributed somewhat to the results.

George et al. (2010) developed an unintrusive model for predicting LEV in a 5.1 scenario; that is, where LEV is predicted from the source signals being fed to loudspeakers, rather than recording the signals at the listening position. In the model, the greatest features contributing to LEV perception were related to the angular distribution of sound around the listener, which is not dissimilar to the lateral energy importance found by Soulodre at al. (2002). There was also some relationship seen between the perceived LEV and octave-band interaural cross correlation coefficients (IACC). As mentioned in Section 1.3.1 above, IACC is a good indicator of ASW, and Berg and Rumsey (2006) have also reported that LEV may be interpreted by listeners as an extension of horizontal width in some cases – this would support the apparent contribution of IACC to LEV perception. It is likely that the energy of late lateral reflections can also have an impact on the calculation of IACC, which may link both LG_{80} and IACC together in the perception of LEV.

1.3.3 Ceiling Reflections

As demonstrated in the ASW and LEV discussions above, spatial impression is typically dictated by lateral reflections, when sufficient reflective energy is present. This lateral impression is also true of common multichannel surround sound formats, where loudspeakers are positioned horizontally around the listener at ear height, such as, 5.1 and 7.1 surround sound (ITU-R, 2012). One assumed benefit of including height-channel loudspeakers in 3D surround sound formats is to enhance spatial impression from above. Therefore, it is also important to consider the role of ceiling reflections in the perception of spatial attributes. However, little research has been conducted in this area, as it is generally assumed that the two main contributors to spatial impression are ASW and LEV through lateral reflections.

Furuya, Fujimoto, Takeshima and Nakamura (1995) investigated the influence of ceiling reflections on spatial attributes through three controlled experiments. The first looked at the effect of direct sound with a single ceiling reflection arriving at +40° elevation in the median plane, assessing various time-delays (10-80 ms) and two attenuations (-6 and -9 dB) for the reflection signal. The results demonstrated that the vertical image spread (VIS) of the source increased as time delay increased for both attenuations. This suggests a similarity with the perception of ASW in the horizontal plane, in that early reflections can cause a perceptual extension of the image. In the second experiment, multiple reflections were presented from above for two control conditions of lateral energy fraction (LF) (0.3 and 0.5), with each condition featuring five levels of relative 'above reflection energy' to 'direct sound energy' (-∞ to -3 dB). It was found that for both lateral fraction conditions, listener envelopment (LEV) increased as the above reflection energy increased. However, it was concluded that LEV was predominantly controlled by the degree of late lateral energy up to 200 ms (LF_0^{200}) (Section 1.4.3). Findings by Evjen, Bradley and Norcross (2001) somewhat support this, where it is seen that the strength of late lateral energy (GL_{80}^{∞}) corresponds best with LEV for all conditions, when assessing stimuli both

with and without late reflections from numerous directions (including laterally and from above).

Various similar studies (Furuya, Fujimoto, Ji & Higa, 2001; Wakuda, Furuya, Fujimoto & Isogai, 2003; Furuya, Fujimoto, Wakuda & Nakano, 2005; Furuya, Fujimoto & Wakuda, 2008) have since demonstrated that late reflections from above can contribute to the perception of LEV as well – however, it was shown in all that the strength of lateral reflections had the greatest contribution to LEV perception, as was observed in the concert hall study conducted by Bradley and Soulodre (1995). Wakuda et al. (2003) estimate that the contribution of late reflections from behind and overhead to LEV are both 50% that of late lateral reflections; whereas Furuya et al. (2008) suggest that the late reflections from overhead contribute even less at 35% that of the late lateral reflections. On the other hand, Lachenmayr, Haapaniemi and Lokki (2016) state that the contributions of late energy observed by Furuya et al. (2008) may only apply when the energy is relatively consistent from all angles. Using stimuli generated from real concert hall impulse responses, Lachenmayr et al. (2016) found that when a concert hall lacked late reflective energy from the sides and back, there was more reliance on the late ceiling reflections to generate a sense of LEV, as might be expected. However, it is thought that such an issue is not of relevance to 3D surround sound systems, as any imbalance of energy is likely to be addressed at the mixing stage of production.

More recently, Robotham, Stephenson and Lee (2016) performed an experiment looking at the effect of a ceiling reflection on preference, as well as spatial and timbral characteristics. A single reflection was reproduced in the median plane at an angle of +38.62° from the listening position – the reflection signal was delayed by 1.63 ms and attenuated by -2.2 dB with respect to the direct sound, in order to replicate a real reflection from a ceiling height of 1.95 m (based on a source-listener distance of 2 m). The results demonstrated that the inclusion of a vertical reflection was mostly preferable, where the preferred stimuli were positively described as ‘bright’, ‘rich’ and ‘full’ by the subjects. With regard to spatial changes, the most common

descriptor was ‘vertical image shift’, followed by ‘vertical image spread’. This is similar to the suggested effect of a single reflection made by Furuya et al. (1995), where it was reported that vertical spread was also experienced by listeners. Both of these results indicate that reflections from above can indeed affect perception of the vertical image. Furthermore, there was no suggestion of a vertical spread effect in the studies on late reflections and LEV above (Furuya et al., 2001; Wakuda et al., 2003; Furuya et al., 2005; Furuya et al., 2008). It might therefore be assumed that this is only a feature of a single reflection (or early reflections) from the ceiling direction, similar to the effect of early reflections on ASW. Since some vertical image change is apparent when signals arrive from above (Furuya et al., 1995; Robotham et al., 2016), the present study may also establish a similar effect by vertical interchannel decorrelation, potentially through the imitation of decorrelated early ceiling reflections.

1.4 Objective Measures of Spatial Auditory Attributes

Section 1.3 on spatial impression suggests various objective measures for calculating spatial attributes within concert halls. It is considered that many of these can also be applicable to the present study, in order to objectively assess a reproduced sound field generated within a 3D surround sound system. The objective measures discussed above and their calculation are described further over the following sections.

1.4.1 Lateral Energy Fraction (LF)

The ratio of overall energy to early lateral reflection energy has been proposed by Barron and Marshall (1981) as a measure for ASW – this is known as the lateral energy fraction (LF). The LF can be calculated using impulses captured by an omnidirectional microphone and a pressure gradient figure-of-8 microphone at the listening position, where the null point of the figure-of-8 is facing the direct sound source (i.e. rejecting the direct sound and capturing mostly lateral reflections) – see Equation 1.1 below (British Standards Institution, 2009).

$$LF = \frac{\int_{0.005}^{0.08} p_8^2\,(t)dt}{\int_0^{0.08} p^2\,(t)dt} \tag{1.1}$$

where $p(t)$ is the sound pressure level (SPL) of the total sound captured by an omnidirectional microphone, and $p_8(t)$ is the SPL of the lateral early reflections captured by the figure-of-8 microphone. The limits of the total sound energy are set between 0 and 80 ms, covering both the direct sound and early reflections; whereas the limits of the lateral reflections are set between 5 and 80 ms, calculating the energy of the early reflection part only.

Although primarily used to measure ASW, Soulodre et al. (2003) observed the effect of calculating the LF of late reflections (> 80 ms) to predict LEV. The measure was found to be effective in some instances, although it was concluded that the limits of integration should be frequency-dependent. They propose combining the late sound strength (LG) (described in Section 1.4.3)

with the late LF for predicting LEV, which ultimately proved to be more accurate than late LF alone in their study.

1.4.2 Interaural Cross-correlation (IAC)

Interaural cross-correlation (IAC) is a binaural measurement of similarity between the two ear input signals. From the literature in Section 1.3.1, it is known that calculation of the IAC coefficient (IACC) for early reflections (< 80 ms) ($IACC_E$) has a strong relationship with the perceived ASW in a concert hall (Hidaka et al., 1995). This is due to an increase of energy from early lateral reflection energy, causing greater interaural differences when the reflections and direct sound are summed at either ear. ASW has also been shown to relate to preference (Blauert & Lindemann, 1986a), with Hidaka et al. (1995) suggesting that the measurement of IACC is a good indicator of concert hall quality – in this case, Hidaka et al. (1995) propose calculating the average $IACC_E$ of the 500 Hz, 1 kHz and 2 kHz octave-bands ($IACC_{E3}$) (as described in Section 1.3.1). Blau (2002) determines that the difference limen of $IACC_{E3}$ is 0.038 – that is, the smallest change of $IACC_{E3}$ required to generate a just noticeable difference (JND).

The IACC can be calculated using the IAC function (IACF) seen in Equation 1.2 below, with a time lag (τ) between -1 to +1 ms – this takes into account the maximum ITD between the two ears. IACC is taken as the absolute maximum value of the function (Equation 1.3), resulting in a coefficient between 0 and 1, where 1 is full correlation (i.e. identical ear input signals) and 0 is no correlation (British Standards Institution, 2009).

$$IACF_{t1,t2}(\tau) = \frac{\int_{t1}^{t2} p_L(t) \cdot p_R(t+\tau)dt}{\sqrt{\left[\int_{t1}^{t2} p_L^2(t)dt\right]\left[\int_{t1}^{t2} p_R^2(t)dt\right]}} \tag{1.2}$$

$$IACC_{t1,t2} = max\left|IACF_{t1,t2}(\tau)\right| \quad \text{for } -1 < \tau < +1 \tag{1.3}$$

where $p_L(t)$ and $p_R(t)$ are the left and right ear input signals, τ is the time lag, and $t1$ and $t2$ are the limits of the IACC (as described below).

When calculating the IACC of early reflections for ASW prediction, $t1$ is 0 ms and $t2$ is 80 ms of an impulse response signal (Equation 1.2); and when calculating the IACC of late reflections for LEV prediction, $t1$ is 80 ms and $t2$ is ∞ (the total length) of an impulse response signal. For continuous signals (not impulse), IACC can be calculated for the entire signal length, where $t1$ is -∞ and $t2$ is ∞, or as the running average over time (frame-by-frame), where $t1$ and $t2$ represent the frame window limits. Mason, Brookes and Rumsey (2003) determine that an optimal window length for IACC calculation is 50 ms, based on the temporal resolution of the auditory system – as a result, this window length has been used for all IACC calculation in the present thesis. Given the versatility of the IACC for indicating spatial features, it is thought that it may be a revealing measure to use in the current study, particularly when source signals are presented from multiple directions.

1.4.3 Sound Strength (G)

The sound strength (G) is the ratio of the sound pressure at the listening position to the sound pressure at 10 m from the source in a free field. This can be calculated using Equation 1.4 below (British Standards Institution, 2009).

$$G = 10\times \log_{10}\left(\frac{\int_{t1}^{t2} p^2\,(t)dt}{\int_0^{\infty} p_{10}^2\,(t)dt}\right) \tag{1.4}$$

where $p(t)$ is the sound pressure at the listening position captured by an omnidirectional microphone, and $p_{10}(t)$ is the sound pressure at 10 m from the source in a free field, also captured by an omnidirectional microphone.

Typically, $t1$ is set to 0 and $t2$ is set to ∞, in order to calculate the overall sound strength of an impulse (Equation 1.4). However, Lee (2012) has also shown that the early sound strength and late sound strength can correspond with ASW and LEV, respectively, when they are measured at different source-listener distances. In the case of the early sound strength, $t1$ is set to 0 ms and $t2$ is set to 80 ms of an impulse response; whereas for the late sound strength, $t1$ is set to

80 ms and $t2$ is set to ∞ of an impulse response. Bradley and Soloudre (1995) and Furuya et al. (2008) have also suggested that the late lateral sound strength (LG) is strongly associated with LEV (even when above and behind reflections are present, as suggested by Furuya et al. (2008)). For this, they record the energy at the listening position with a figure-of-8 microphone instead, where the null is facing the source and rejecting direct sound (similar to the process of measuring LF).

1.4.4 Front-back Energy Ratio (F/B)

Front-back energy ratio (F/B) has been demonstrated as a potential predictor of LEV in both concert halls (Morimoto & Iida, 1998; Morimoto, Iida & Sakagami, 2001; Morimoto, 2002; Lee, 2012) and multichannel surround sound systems (Morimoto, 1997). It is calculated as the ratio of energy from in front of the listener to that from behind, using Equation 1.5 below (Morimoto et al., 2002).

$$F/B\ energy\ ratio = 10\times \log_{10}\left(\frac{\int_0^\infty f^2(t)dt}{\int_0^\infty b^2(t)dt}\right) \quad (1.5)$$

where $f^2(t)$ is the energy from in front of the listener at the listening position, and $b^2(t)$ is the energy from behind the listener at the listening position.

There is no consensus on how to measure for F/B in a practical situation. However, Lee (2012) used front and rear facing virtual cardioid signals generated from a B-format impulse (captured in a real concert hall), where the results corresponded quite well with the perceived LEV. It might therefore be assumed that a pair of actual cardioid microphones, one facing the front and one facing the rear, could also be used to calculate F/B.

Morimoto and Iida (1998) found that the F/B was able to predict the perceived LEV with both early reflections (including direct sound) and late reflections independently – this suggests that F/B could be used as a general measure of LEV, taking into account the direct sound and all reflections simultaneously. A similar approach may be effective for assessing LEV within 3D

loudspeaker arrays, where the total energy of signals from the frontal loudspeakers is compared to that of the rear loudspeakers. Given the known directional spectral filtering of the HRTF, it would also be beneficial to measure the F/B ratio with binauralised signals, in order to give a true representation of the energy difference experienced by the listener (i.e. including spectral notches and boosts).

1.5 Summary

This chapter discussed various processes of the human auditory system that allow for the spatial perception of audio. A particular emphasis has been placed on the vertical localisation of audio, given that the present study primarily involves signal processing in the vertical domain. Furthermore, the perception of reflections within concert halls (the so-called spatial impression) has also been explored, in order to understand how the upmixing of ambience in 3D surround sound systems could affect spatial attributes, such as, listener envelopment (LEV). The final section of the review focuses on existing objective measures of spatial attributes that might be relevant to the current study.

The first section of the review (Section 1.1) deals with the localisation of auditory events both horizontally and vertically. Section 1.1.1 assesses the known contribution of both interaural time difference (ITD) and interaural level difference (ILD) to the horizontal localisation of sound, termed the 'duplex theory' (Rayleigh, 1907). It is accepted that, in general, the lateral localisation of lower frequencies (< 1.5 kHz) is dictated by ITD, and the localisation of higher frequencies (> 1.5 kHz) by ILD (Moore, 2012). A consideration with regard to ILD in 3D surround sound systems was also demonstrated, where signals from height-channel loudspeakers result in less head-shadowing (ILD) than signals from main-channel loudspeakers, which may impact the perception of vertical decorrelation at wide azimuth angles.

Vertical localisation is investigated in Section 1.1.2, where it appears that high frequency spectral cues are the greatest contributor. It is seen that spectral filtering from the pinna above 3 kHz aids the perception of elevated sound sources in the median plane from the front, with a particular importance of frequencies around 8 kHz (Roffler & Butler, 1968a; Hebrank & Wright, 1974). Elevation cues have also been observed at low frequencies due to shoulder and torso reflections (down to 700 Hz) (Algazi et al., 2001). However, some studies have demonstrated that cues from higher frequencies take precedence over those at lower frequencies (Morimoto et al., 2003; Ferguson & Cabrera, 2005).

With further regard to vertical localisation, elevation phenomena known as the 'pitch-height' effect (Lee, 2016b) and 'directional bands' theory (Blauert, 1969/70) have also been considered in Section 1.1.2. With the 'pitch-height' effect, Lee (2016b) demonstrates that different octave-bands are inherently perceived from different heights when presented from the same location. In general, it is seen that higher frequency bands are perceived higher in space and lower frequency-bands are perceived lower – in the experiments of Lee (2016b), the 8 kHz octave-band was shown to have the greatest elevation effect, which may relate to the role of this region in vertical localisation perception. Similarly, Blauert's (1969/70) directional bands work sees different frequencies being perceived from different directions under anechoic conditions, where 8 kHz tends to be perceived from above. These phenomena and the studies on vertical localisation all demonstrate the importance of frequencies around 8 kHz for elevation perception from the front – it is hypothesised that similar spectral cues are likely to contribute to the perception of vertical interchannel decorrelation.

Section 1.2 considers the inherent spatial extent of audio in terms of frequency, loudness and signal duration. In general, it is seen that lower frequencies are perceived as inherently broader than higher frequencies when presented at the same sound pressure level (SPL) (Terrace & Stevens, 1962; Perrott et al., 1980; Cabrera & Tilley, 2003). In terms of vertical extent, results from Cabrera and Tilley (2003) suggested that higher frequencies were generally perceived as more vertically spread than horizontally spread, whereas lower frequencies tended to be more horizontally spread than vertically spread. It is thought that the enhanced vertical extent of high frequencies may relate to the 'pitch-height' and 'directional bands' effects, particularly when frequencies around 8 kHz are present. Studies also indicated that loudness had a significant impact on perceived extent, where a greater SPL increased the extent of the auditory image – the rate of this change appears to be greater horizontally than vertically, potentially relating to the placement of the ears (Perrott et al., 1980). Lastly, Perrott et al. (1980) also observed an interesting effect where the extent of a tone increased as its duration increased – the cause of this effect remains unclear, and little research has been conducted on the effect since.

Spatial impression (the perception of reflections within an enclosed space) is described in Section 1.3, of which the two main components are apparent source width (ASW) and listener envelopment (LEV). ASW is dictated by early lateral reflections (< 80 ms), whereas LEV is dictated by the late room reflections (> 80 ms) (typically from the lateral direction also) (Rumsey, 2001). Measures of both ASW and LEV are featured in Section 1.4 and summarised below. With regard to reflections from above (the ceiling), it was suggested that a single early reflection in the median plane can contribute to a vertical image shift and an increased sense of vertical image spread (Furuya et al., 1995; Robotham et al., 2016). It is also considered that the greatest contributor to LEV are late lateral reflections from the side walls, with late reflections from above and behind contributing less (Wakuda et al., 2003; Furuya et al., 2008). These studies suggest that some effect of vertical interchannel decorrelation may be perceivable; however, since spatial impression appears to be mostly associated with lateral reflection energy, it may be seen that vertical decorrelation is unable to affect a significant change in 3D surround sound reproduction.

Section 1.4 details some objective measures for predicting spatial impression within concert halls, which may also be applied to 3D surround sound systems. Firstly, lateral energy fraction (LF) is discussed – a measure that has been shown to be associated with ASW when calculated for early reflections (Barron & Marshall, 1981). A second predictor of ASW is to calculate the interaural cross-correlation coefficient (IACC), which assesses the similarity between the two ear input signals, and is mostly dictated by the strength of early lateral reflections – Hidaka et al. (1995) suggest that the measure should be the average of the 500 Hz, 1 kHz and 2 kHz octave-bands ($IACC_{E3}$), which was revealed to be a good predictor of concert hall quality. Sound strength (G) is the ratio of signal energy at the listener position to that at 10 m from the source in a free field – Lee (2012) has shown the early reflection strength and late reflection strength to correspond with ASW and LEV, respectively. Finally, front-back energy ratio (F/B) observes the difference in energy from in front of the listener to that from behind, which is also shown as a reasonable predictor of LEV (Morimoto & Iida, 1998; Lee, 2012).

2 THE SPATIAL CONTROL OF AUDIO IN SURROUND SOUND REPRODUCTION

Chapter 1 of the present thesis assessed the different hearing mechanisms that allow for the perception of audio within space. This current chapter looks at how these mechanisms can be utilised in multichannel surround sound systems to generate positional and size information of auditory objects. The structure of the chapter is as follows. Section 2.1 describes various 2D and 3D loudspeaker formats that are in use today. Section 2.2 explores techniques for controlling the position of a point source between loudspeakers by time and amplitude panning, as well as the effectiveness of these both horizontally and vertically. Controlling the extent of a phantom image is studied in Section 2.3, where interchannel decorrelation is discussed in detail as an effective method of source extension, along with potential approaches for achieving this. There is also consideration of interchannel decorrelation in the vertical domain, which is the focus of the current study; however, little literature currently exists on the subject. Lastly, Section 2.4 looks at multichannel upmixing to high-order loudspeaker formats, which is a potential application of vertical interchannel decorrelation.

2.1 Loudspeaker Reproduction Systems

In order to investigate the application of vertical interchannel decorrelation in 3D surround sound systems, it is first important to gain an understanding of the systems that are currently in existence. This section describes established loudspeaker formats, as well as those proposed more recently, and discusses how sound is perceived when reproduced by them.

2.1.1 Two-Channel Stereophony

The most common loudspeaker reproduction format able to create a spatial sound scene is two-channel stereophony (2.0). This features a pair of left and right loudspeakers positioned in front of the listener, where the auditory image is spread between the boundaries of the loudspeakers. A recommended base angle between a two-channel stereophonic pair of loudspeakers is 60° (the angle between the two loudspeakers from the listener's perspective), positioning each one at ±30° azimuth from the centre position (where 0° is directly in front of the listener) (Figure 2.1) (Rumsey 2001; ITU-R, 2012).

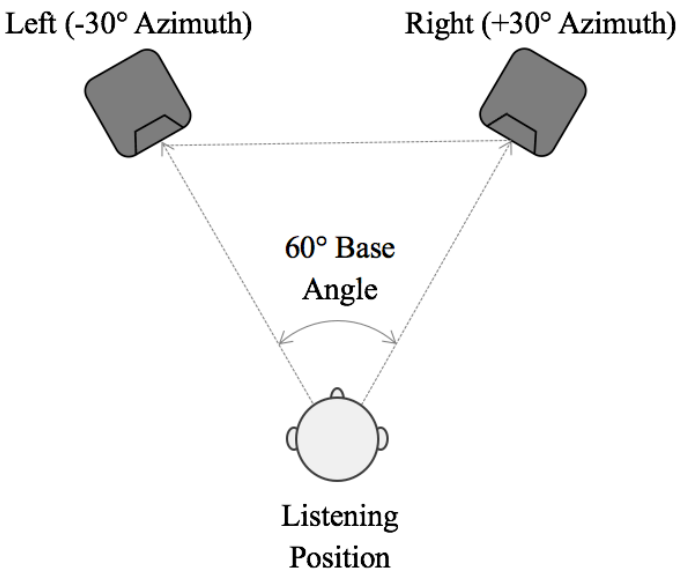

Figure 2.1 Two-channel stereophonic loudspeaker setup.

When coherent (identical) signals are reproduced from the two loudspeakers simultaneously, a phantom auditory image is generated between them, due to an auditory phenomenon known as 'summing localisation' (Blauert, 1997). 'Panning' can then be used to position the phantom image across the horizontal plane, by altering the level and/or time difference between the two coherent signals (as described in Section 2.2.1 below) – a process that is mostly limited to the boundaries of the loudspeakers. The interchannel differences generated by panning rely on the interaural level and time difference localisation cues (ILD & ITD) discussed in Section 1.1.1. Furthermore, multiple phantom image sources can be perceived simultaneously and panned to different locations, which formulates the horizontal sound scene in front of the listener. When the two loudspeaker signals are partially coherent (uncorrelated), the phantom image is perceived as a spread of sound between the two loudspeaker positions. With this, the signals are mimicking the effect that lateral reflections have on the apparent source width (ASW) (Section 1.3.1), by decreasing the level of interaural cross-correlation (IAC). A process called decorrelation can be used to achieve this spread of sound, where the greater the decorrelation (the lower the correlation between the two signals), the greater the perceived spread (see Section 2.3). Performing this process in the vertical domain is the focus of the present study.

2.1.2 5.1 and 7.1 Surround Sound

The most common expansion of two-channel stereophony is 5.1 Surround, which sees four further loudspeakers included alongside the standard 'Left' (L) and 'Right' (R) loudspeakers. These are a 'Centre' (C) channel positioned between L and R at 0° azimuth, two surround channels ('Left Surround' (Ls) and 'Right Surround' (Rs)) behind the listener at ±100-120°, and a 'low frequency effects' subwoofer channel (LFE) (ITU-R, 2012) (see Figure 2.2 (Left)).

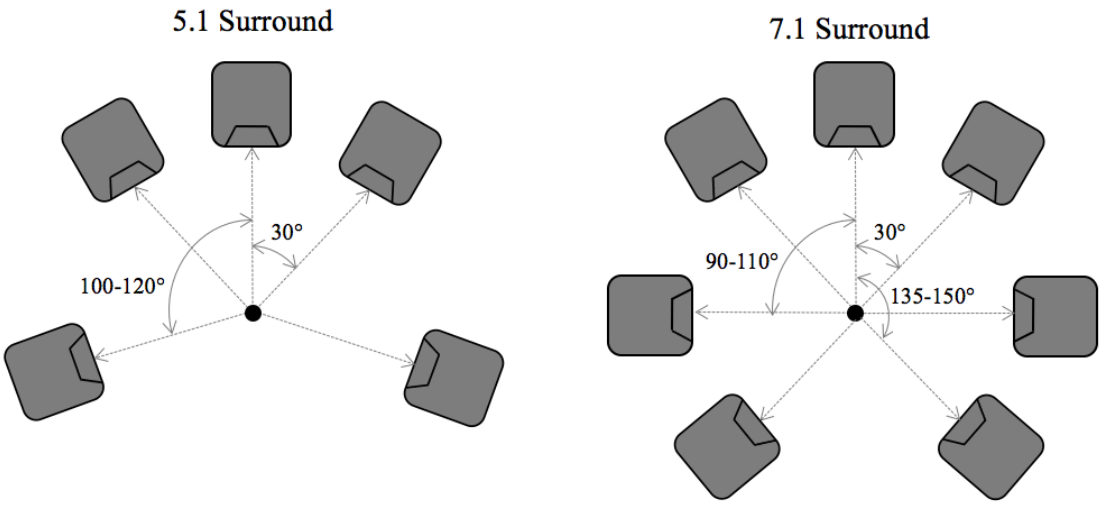

Figure 2.2 5.1 and 7.1 Surround loudspeaker setups (left and right, respectively).

The growth of 5.1 Surround was primarily seen in film and television production, giving the ability for ambience and sound effects to be generated from behind the listener (in the surround channels), while the centre channel is typically used for dialogue (Howard & Angus, 2017). In the present day, commercial music is increasingly being recorded, mixed and released in surround sound formats, providing the listener with a more realistic sense of spatial impression, and ultimately a preferable listening experience. It was seen in Section 1.3 that the two main components of spatial impression are apparent source width (ASW) and listener envelopment (LEV), both of which rely heavily on lateral reflections. Given this, the inclusion of ambient signals in the surround channels supports the generation of such cues, specifically enhancing the perception of LEV. A further extension of 5.1 Surround is to split the two surround channels into four channels spread across the side and rear, known as 7.1 Surround – these consist of two 'wide' channels at ±90-100° and two 'rear' channels at ±135-150°, as shown on the right of Figure 2.2 above (Dolby Laboratories, 2017). In the context of the present study, the phrases '2D formats' and '2D content' refer to 5.1 and 7.1 Surround as described in the ITU-R

Recommendation BS.775-3 (ITU-R, 2012), where the auditory reproduction is limited to the horizontal domain (i.e. at ear height).

An important phenomenon to consider in surround sound reproduction is the "precedence effect" or "law of the first wavefront" (Wallach, Newman & Rosenzweig, 1949; Litovsky et al., 1999). When a pair of similar signals arrive at the ear from two different positions with a delay greater than 1 ms between them, the auditory event is localised at the source position of the earlier signal. This effect has been thought of as a reflection suppression mechanism that aids the localisation of sound sources (Blauert, 1997). If the delay between the signals is increased further, the individual sources will then appear one after the other from their respective positions, with the secondary source perceived as an echo of the first. This upper limit of the precedence effect is known as the "echo threshold", which has been found to differ depending on the type of source. For a click stimulus, the echo threshold is shown to be between 2 and 10 ms (Rosenzweig & Rosenblith, 1950; Thurlow & Parks, 1961), while the threshold increases slightly for noise pulses to around 15 ms (Damaske, 1971). For continuous speech sources, the echo threshold has been found to be as much as 50 ms (Haas, 1972). Through additional exploration of the precedence effect on speech audibility by Haas (1972), it was discovered that the suppression of the secondary signal still occurred even when it was greater in level than the primary signal – this has since been called the "Haas Effect". Haas (1972) demonstrated that the effect was greatest for delays between 10 and 20 ms, where the secondary signal was increased by around 10 dB and the auditory event was still perceived at the primary position. In the context of surround sound reproduction, if the rear signals arrive 1 ms before the front signals (based on the listener's position), the auditory image would be localised from behind. To counter this, the rear signals should be suitably delayed from the front signals (e.g. 10-15 ms), so that the image is localised at the front of the array from multiple listening positions (Avendano & Jot, 2002).

2.1.3 3D Multichannel Surround Sound

For commercial use, 3D multichannel surround sound systems usually incorporate height-channel loudspeakers into a 2D format. For example, the Auro-3D 9.1 format (Auro Technologies, 2015a) is based on the standard 5.1 Surround layout described in Section 2.1.2 above. However, four additional height-channel loudspeakers are also positioned directly above the Left (-30°), Right (+30°), Left Surround (-110°) and Right Surround (+110°) main-layer loudspeakers, at an elevation angle of +30° to the listening position (Figure 2.3). Auro-3D describe the 9.1 format as the "minimal setup for [the] full Auro-3D experience", which can be expanded to 10.1 by positioning another loudspeaker directly above the listener (i.e. the 'voice of god' channel). For larger rooms, such as cinemas, Auro-3D recommend further expansion on 10.1 with an added height-channel at 0° azimuth (Auro-3D 11.1), as well as additional Left Rear Surround and Right Rear Surround main-layer channels at ±150° azimuth (Auro-3D 13.1), based on the 7.1 Surround format described in Section 2.1.2 above.

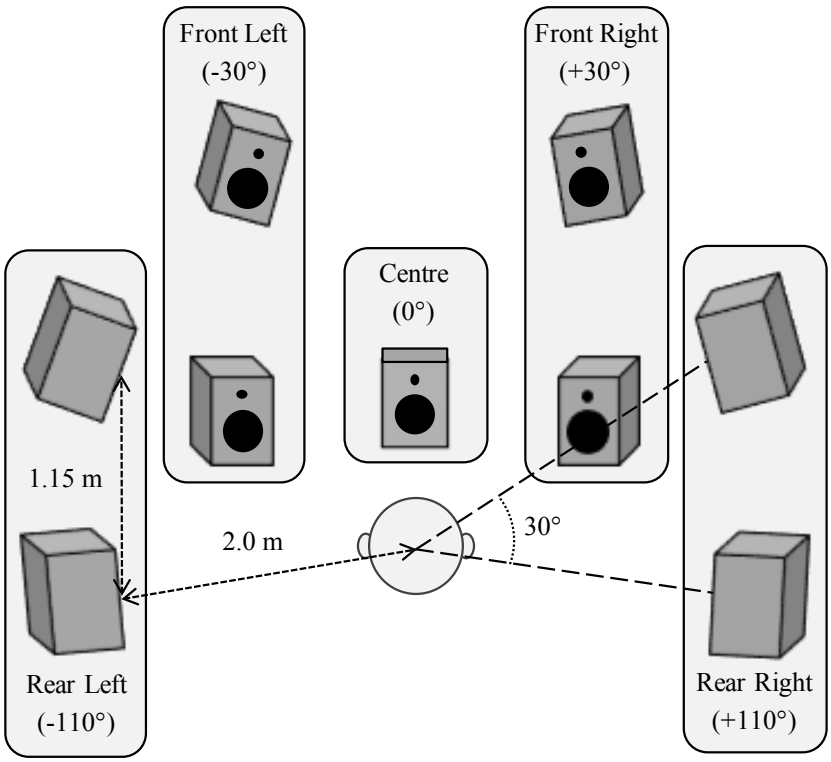

Figure 2.3 Auro-3D 9.1 loudspeaker setup (Auro Technologies, 2015a).

Similar to Auro-3D, other commercial formats that employ height-channels include Dolby Pro Logic IIz and DTS Neo:X. Dolby Pro Logic IIz is an expansion of 7.1 Surround with two additional height-channels above the Left (-30°) and Right (+30°) loudspeakers in front of the listener. DTS Neo:X also has the same two left and right height-channels above the 7.1 standard as Dolby Pro Logic IIz, but features additional front left and right wide channels (±60°) in the main-layer as well, allowing for a broader and potentially more stable frontal image.

The 3D formats mentioned so far are regarded as channel-based systems, where the content is defined by the discrete number of channels. That is, the localisation and level of auditory sources are determined in the post-production and fixed from thereon. A scalable and flexible solution can be found with object-based systems, where sound objects (sources) are encoded independently in the audio files, along with metadata that contains the object's spatial information. The rendering of object-based audio to multichannel systems is dependent on the particular loudspeaker setup of the user, i.e., the position of each audio object is translated from the relative geometric information in the metadata to fit the dimensions of the loudspeaker array. The use of discrete objects also gives more control to the end-user, where they can be muted and potentially panned in real time, similar to the processing of auditory objects within game audio. The main formats to utilise such a system are Dolby Atmos (Dolby Laboratories, 2014), AuroMax (Auro Technologies, 2015b) and DTS:X, all three of which feature both channel and object information. Dolby Atmos is the most prevalent of these object-based formats, and can already be found in many commercial cinemas. Furthermore, object-based content has become standardised over recent years, as demonstrated by the development of the MPEG-H format to include such encoding (Herre, Hilpert, Kuntz & Plogsties, 2014, 2015). The use of object-based systems could make applications such as upmixing to uncommon loudspeaker setups easier, since the ambient signals can also be transported on separate audio streams, alongside the audio objects. This means that there is less of a need to employ ambience extraction methods for upmixing content from 2D to 3D, as described further in Section 2.4 below.

2.2 Controlling the Position of an Auditory Event

2.2.1 Horizontal Panning

As mentioned in Section 2.1.1 above, when two loudspeakers present coherent (identical) signals, a phantom auditory image is formed between the two loudspeakers due to summing localisation. In a two-channel stereophonic system (i.e. a pair of left and right loudspeakers), the interchannel level difference (ICLD) and interchannel time difference (ICTD) can be used to offset the position of the auditory phantom image between the two loudspeakers. This processing works by manipulating the interaural level difference (ILD) and interaural time difference (ITD) localisation cues discussed in Section 1.1.1. The most common approach is to amplitude (or intensity) pan the phantom image, where the ICLD between the channels is adjusted – this is predominantly the process behind panning applications within mixing desks or digital audio workstations (DAWs). Different amplitude panning laws have been defined over the years, in order to maintain constant power when the two outputs are summed at the listening position. Bauer (1961) derives the sine law of amplitude panning, which does not take into account the behaviour of sound waves around the head. It relies on the listener to be positioned directly between two horizontally spaced loudspeakers (i.e. left and right with a base angle of 60°) and facing forward. An improvement on this named the tangent panning law is reported by Bennett, Barker and Edeko (1985), where the curvature of the contralateral loudspeaker signal around the head is considered – this accounts for the listener rotating their head to face the phantom source position and is thought to be a more accurate representation of lateral localisation cues.

Pulkki (1997) centres his vector base amplitude panning (VBAP) method around the tangent panning law. It is a reformulated version of tangent panning that features vectors and vector bases – the use of vectors allows for panning within 3D environments, where multiple loudspeakers can be placed anywhere on a hemi-spherical surface. Pulkki and Karjalainen (2001) conducted a horizontal (2D) panning experiment using VBAP, with subjects instructed to pan

a phantom image to the location of a real loudspeaker source. Broadband and 1/3-octave-band filtered pink noise were assessed, along with 1/3-octave-band filtered pink noise impulse trains. The results demonstrate that localisation of lower frequencies is reliant on ITD, whereas localisation of higher frequencies is dictated by ILD from the amplitude panning. Furthermore, localisation is largely accurate for low and high frequency bands. However, the bands with centre frequencies around 1-2 kHz tended to be underestimated from the real location.

Lee and Rumsey (2013) compared the accuracy of both amplitude and time panning using musical (piano and trumpet) and speech sources. The results showed that a high note on the trumpet (continuous with a fundamental frequency of 922 Hz) was difficult to pan using ICTD, demonstrating that transients and lower frequencies are required for an ICTD approach – this is in accordance with the ILD and ITD duplex theory discussed in Section 1.1.1. On the other hand, the ICLD panning method was relatively effective for all sources, demonstrating that interaural amplitude cues down to below 1 kHz are effective for localisation. For wider azimuth angles, a greater ICLD was required to meet the target angle: between 0° and ±20°, ICLD worked at a rate of 0.425 dB/deg, whereas between ±20° and ±30 it was double at 0.85 dB/deg. From these results, Lee (2017b) developed a novel panning algorithm named perceptually motivated amplitude panning (PMAP), in order to improve on the established tangent panning method (e.g. VBAP). It features coefficient weightings that depend on the target angle, which align with the rate of required ICLD change observed in the previous study by Lee and Rumsey (2013). The use of a scaling factor was also proposed for applying the PMAP algorithm to any base angle between two horizontally-spaced loudspeakers. For a standard 60° base angle, PMAP was shown to perform more accurately than tangent panning – the results for tangent panning tended to be over-estimated for wider angles, confirming the findings of Pulkki and Karjalainen (2001). With a base angle of 90°, PMAP (with a 90° scaling factor) also performed well. However, the 90° tangent panning accuracy was also comparable to PMAP at 90°. These results demonstrate that the use of amplitude panning is not linear in the horizontal plane, and also that one approach is not effective for all loudspeaker base angles.

Thiele and Plenge (1977) observe the amplitude panning of phantom sources to the side of the listener (laterally), which is of particular relevance to the advancement of surround sound systems. Under anechoic conditions, two loudspeakers with a base angle of 60° were positioned with various centre azimuth angles around the head (ranging from 0° to 120°). The stimuli tested were impulses of noise and a speech signal, where the listener was asked to localise the sources for various levels of ICLD. It was found that localisation was least accurate / worst when the loudspeakers were centred around 90° azimuth, with large deviations of data when there was no level difference between the loudspeakers (i.e. a target angle of 90°) – similar results were also seen for a centre angle of 80°. In contrast, when the two loudspeakers are centred around 60° the localisation appears to be mostly accurate, forming an almost linear relationship with the ICLD. From these results, Thiele and Plenge (1977) propose that six loudspeakers should be positioned around the listener to create an "all round effect", using azimuth angles of ±30°, ±90° and ±150°. In the context of the present study, it is also important to assess the azimuth angle dependency of vertical decorrelation. For example, poor phantom image localisation at wide lateral positions may result in a reduction of the perceived effect.

2.2.2 Vertical Panning

For horizontal panning, the location of a phantom image from summing localisation is easily controlled, through adjustment of the ICLD and/or ICTD between the two loudspeaker channels. With the advancement of 3D surround sound systems, it is also necessary to observe the control of a point source within the vertical domain. Pulkki (2001) explores the use of VBAP to pan between two vertically-arranged loudspeakers in the median plane, with one positioned at -15° elevation and the other at +30° elevation. Subjects were asked to localise the position of phantom images with elevated target angles of 0° and +15° against real loudspeaker sources at the same positions. The results demonstrate a reasonable correspondence between the phantom and real source locations. However, the accuracy appears to be rather poor, with the data spread over a range of 20° for both phantom positions. Nevertheless, the results indicate that it

is possible to perceive a phantom image between two correlated sources when presented vertically. This is of interest to the current study, as it supports the idea that partially correlated sources may also be perceived as phantom images between vertically-arranged loudspeakers (as generated by vertical interchannel decorrelation).

Barbour (2003) assessed the vertical localisation of phantom sources in both the median (0° azimuth) and frontal (90° azimuth) planes. For each azimuth angle, phantom images were assessed between three discrete pairs of vertically-arranged loudspeakers. For each pair, the phantom image was generated between a lower main-layer loudspeaker at 0° elevation and three upper height layer loudspeaker positions independently, with elevation angles of +45°, +60° and +90°. The subjects were asked to locate the phantom image of male speech and pink noise that had been amplitude panned with varying degrees of ICLD (ranging from -15 dB to +15 dB). When panning between 0° to +45° in both the median and frontal planes, the middle ICLD values (between ±6 dB) resulted in a great deviation of responses, which increases as the amplitude bias moves slightly towards the height loudspeaker. Vertical panning from 0° to +60° and from 0° to +90° in the median plane also performed similarly, in that large deviations are seen for middling ICLD values. In contrast, vertical panning from 0° to +60° and from 0° to +90° in the frontal plane sees a noticeable improvement of localisation accuracy. This is likely due to changes in ILD from the loudspeakers being presented off-centre – it relates back to the consideration of ILD and height-channels made in Section 1.1.1.2, where it is seen that elevation reduces ILD with respect to the main-layer channel (of which the signal is largely head-shadowed in the contralateral ear, causing a greater ILD). In the context of the present study, these results demonstrate the improvement height-channels off-centre can have on vertical localisation, which may also impact on the perception of vertical interchannel decorrelation.

Mironovs and Lee (2017) assessed the vertical panning of signals between two vertically-arranged pairs of loudspeakers, one at 0° azimuth (the median plane) and the other at +30° azimuth. The two lower main-layer loudspeakers were positioned at ear height, while the upper

height-layer loudspeakers were elevated by +30° to the listening position. VBAP (based on the tangent panning law) (Pulkki, 1997, 2001) was used to pan the stimuli to various angles along the vertical plane. For each azimuth angle, seven target elevation angles were assessed between 0° and +30° at 5° intervals. A total of six stimuli were assessed for each azimuth and elevation combination, consisting broadband pink noise, low-pass filtered pink noise (< 3 kHz), high-pass filtered pink noise (> 3 kHz), birdsong, a tank shot and male speech. In general, the results show that the localisation of most stimuli jumps from near the main-layer loudspeaker up to the height-layer loudspeaker around +10-20°, with little accurate localisation in between – this is largely in agreement with the findings of Barbour (2003). Only the broadband pink noise and high-pass filtered pink noise appear to demonstrate a stable phantom image for the 15° target angle, whereas the low-pass filtered pink noise and birdsong were poorly localised for all angles. These results would suggest that high frequencies within a broadband signal are necessary for the localisation of vertical panning (i.e. not a narrowband of high frequencies as with the birdsong) – this is similar to the requirements of accurate vertical localisation as seen in Section 1.1.2.1 (Roffler & Butler, 1968a; Hebrank & Wright, 1974). Despite the general lack of accuracy towards the target angle, the results still provide some evidence that it is possible to perceive an elevated phantom image by vertical panning, as was the case in previous studies (Pulkki, 2001; Barbour, 2003).

Further demonstration of a vertical phantom image is seen in the work of Wallis and Lee (2015b). They observe the effect of interchannel crosstalk between a main-layer loudspeaker and a height-layer loudspeaker, assessing both octave-band and broadband pink noise in the median plane. It is seen that when the two signals are time and level aligned (ICTD and ICLD = 0), the broadband stimulus and the 2 kHz, 4 kHz and 8 kHz octave-band stimuli are all localised between the two loudspeaker positions (i.e. a vertical phantom image). On the other hand, octave-bands with centre frequencies between 125 Hz and 1 kHz were perceived around or below the main-layer loudspeaker. This also demonstrates the influence of high frequencies in the perception of vertical panning, and potentially their importance in the perception of vertical

interchannel decorrelation – the results here also agree with the notion that vertical localisation is dictated by high frequency content (Roffler & Butler, 1968a).

2.3 Controlling the Spatial Extent of an Auditory Event

Section 2.2 discussed the manipulation of coherent (correlated) signal pairs, in order to position a sound source within a loudspeaker array i.e. the panning of a phantom image by changing the ICTD and ICLD. As mentioned in Section 2.1.1, when the coherence (correlation) decreases between two signals in the horizontal domain (i.e. left and right), the phantom image between the loudspeakers increases in horizontal extent. In two-channel stereophony, this is a replication of the lateral reflections that cause an increase of apparent source width (ASW), by decreasing the interaural cross-correlation coefficient (IACC) (Section 1.3.1). Consequently, the lower the interchannel cross-correlation coefficient (ICCC) (the degree of correlation between the two loudspeaker signals), the lower the IACC, and the greater the horizontal spread that is perceived (Zotter & Frank, 2013). The effect of IACC on spread was formally assessed over headphones by Blauert and Lindemann (1986b) using noise sources. It was clearly shown that the extent of spread corresponded directly with the degree of IACC.

The ICCC can be calculated using the normalised cross-correlation function, similar to the calculation of the IACC (without the time lag), as seen in Equation 2.1 below – where the ICCC is defined as the result from the interchannel cross-correlation function (ICCF). ICCC of an entire signal can be calculated by windowing the signal and averaging the ICCC over time. In the present study, the window length is set to 50 ms for all ICCC calculation, which is based on the optimum window length determined by Mason et al. (2003) for IACC calculation (due to the temporal resolution of the hearing system).

$$ICCF = \frac{\int_{t1}^{t2} s_1(t)s_2(t)dt}{\sqrt{\left[\int_{t1}^{t2} s_1^2(t)dt\right]\left[\int_{t1}^{t2} s_2^2(t)dt\right]}} \qquad (2.1)$$

where $s_1(t)$ and $s_2(t)$ are the two signals being analysed for correlation, t_1 is the lower limit of the window and t_2 is the upper limit of the window.

Blauert (1997) states that two techniques might be used to generate partially coherent or incoherent (uncorrelated) signals:

1) The distortion technique, where distorted signals are created from the original signal.
2) The superposition technique, where another signal is introduced to the original signal.

The processes described over the following sections are considered to be decorrelation by distortion, since this is a more practical approach to achieve decorrelation from existing signals (as would be the case in upmixing applications). Distortion here does not necessarily refer to a degradation of the signal(s) in terms of tonal quality; that is, the decorrelated signals remain largely true to the original signal. The aim of decorrelation is to generate signals that sound sonically similar to the input signal, yet are perceptually different to the human auditory system. This is achieved through very subtle phase and/or spectral-amplitude manipulations that allow the hearing system to perceive two independent signals, which then fuse together to form the broad phantom image. Both phase- and amplitude-based approaches are discussed in Sections 2.3.1 and 2.3.2 below, respectively.

Martens, Braasch and Woszczyk (2004) investigated the ability to spatially discriminate between a monophonic signal and two decorrelated low frequency 1/3rd-octave-band noise signals, featuring centre frequencies between 31-125 Hz. It was found that when interaural differences were minimised, subjects were unable to identify any difference between the two conditions. However, when interaural differences were present, the ability to discriminate between them improved for the 63 Hz band and above. The spatial changes observed were related to listener envelopment (LEV) – the sensation of feeling immersed by sound from all around (Section 1.3.2). This suggests that vertical decorrelation of low frequencies may be ineffective in the median plane (where interaural differences at low frequencies are nil), though if the decorrelated signals were presented off-centre, enhanced spatial impression (e.g. LEV) might be realised. Furthermore, the spatial effect of decorrelation appears to be perceivable down to around 63 Hz – this is broadly in agreement with Griesinger (1999), who suggests that lateral

reflections down to 60 Hz are "vital to world class envelopment". Taking this into account, similarly low bands must also be considered in the present study, to observe whether vertical decorrelation of low frequencies can also contribute to an increased sense of LEV (or spatial impression in general).

2.3.1 Phase-Based Decorrelation Methods

Considering phase-based decorrelation approaches, Kendall (1995) proposed the use of all-pass filters to randomise the phase of frequencies within a signal, whilst maintaining the frequency magnitudes between the input and output. It is suggested that when two signals have differing phase at each frequency, this amounts to a decorrelation between the signals and a spread is perceived when presented between a left and right loudspeaker pair (by decreasing the level of IACC at the ears). A simple way to implement all-pass filtering is by convolving the original monophonic signal with two impulses of random phase and unit magnitude (i.e. short white noise bursts) (Equation 2.2).

$$\begin{aligned} s_1(n) &= x(n) * h_1(n) \\ s_2(n) &= x(n) * h_2(n) \end{aligned} \qquad (2.2)$$

where s_1 and s_2 are the two output signals, x is the monophonic input signal, and h_1 and h_2 are two FIR filter impulses.

To create the all-pass filters, two random number sequences are generated as FIR filter phase coefficients (featuring random values between –π and π), giving an inherent decorrelation between the two filters – the degree of this correlation can then be controlled by a mixing matrix. It is thought that the correlation between the two impulses directly relates to the correlation between the two outputs; however, given the random generation of the number sequences, the actual degree of maximum ICC can vary drastically between each creation of filters. Further details on implementing Kendall's method can be found in Section 4.2.1.3 of the current thesis.

Expanding on Kendall's approach, Potard and Burnett (2004) filter the input signal into three frequency bands before processing with an all-pass filter – 'low' (0-1 kHz), 'medium' (1-4 kHz) and 'high' (4-20 kHz) – in order to give greater control over the effect. Hawksford and Harris (2002), Faller (2006) and Pulkki (2007) propose the use of exponentially decaying noise rather than a uniform burst of energy. To achieve this, Faller (2006) artificially synthesised late reverberation by using a decaying burst length of 400 ms, whereas Pulkki (2007) suggests shorter frequency-dependent decaying noise bursts. Pulkki (2007) used three different decaying burst lengths for separate frequency-bands (100 ms below 400 Hz, 40 ms for 400 – 1300 Hz, and 10 ms above 1300 Hz). Although this approach may improve the issue of transient response and colouration somewhat, it also increases the complexity of achieving and controlling a low level of ICC.

Given the random phase at each frequency, the waveform of the output from all-pass decorrelation can be distorted by significant opposing phase-shifts of neighbouring frequencies (Bouéri & Kyriakakis, 2004). It is this random interaction of phase-shifts that can lead one implementation of the all-pass filter to sound noticeably different from another. As an alternative to phase-shifting singular frequencies (in an attempt to reduce colouration), Bouéri and Kyriakakis (2004) randomly delay each critical frequency-band of the human hearing system (obtained using an "equivalent rectangular bands" (ERB) filter bank), with the limits of the time-delay set proportionally to the frequency wavelength. These delays ranged from around 45 ms for the lowest critical band, to less than 1 ms for the highest critical band. Results from informal testing by Bouéri and Kyriakakis (2004) suggest that an increase of spatial extent is perceived, however, there was still some audible timbral colouration. This technique may be preferred over a more general all-pass filtering approach if a filter-bank is already being implemented in the signal chain. Vilkamo, Lokki and Pulkki (2009) used the same approach in their directional audio coding (DirAC) technique, where informal testing determined that the random delay of ERB frequency bands produced a higher quality output than other decorrelation methods. They

set a minimum delay of 5 ms for all frequency bands, in order to avoid localisation problems from coherent summation of the signals.

Laitinen, Kuech, Disch and Pulkki (2011) expand on the DirAC framework with the addition of transient detection. They also suggest optimal delay limits for the frequency band delay method, stating that for bands below 1500 Hz the maximum delay should be 50 times the cycle time and no more than 100 ms, whereas above 1500 Hz the maximum delay should always be 50 ms – furthermore, the minimum delay should be 10 times the cycle time and no less than 5 ms (similar to that suggested by Vilkamo et al. (2009)). With regard to the transient handling of decorrelators, when transients are processed through decorrelation algorithms the output is 'smeared' by the filter response (e.g. the length of the noise burst), hence the reason for exponentially decaying noise bursts used by Pulkki (2007). For sources such as an applause, multiple transients that are densely populated in time can almost cause a steady noise-like signal, due to a loss of the transient definition from an overlapping of smearing. Laitinen et al. (2011) propose transient extraction to avoid this, which they implement by comparing the instantaneous energy of each frequency band with respect to the long-term average energy of the diffuse stream, where the threshold of detection can be defined independently. A new diffuse stream with compressed transients is then generated for the decorrelation processing; while a stream of the extracted transients is not processed with decorrelation and presented as coherent, in order to avoid the smearing effect. Through formal listening tests, Laitinen et al. (2011) demonstrate that the greater the temporal resolution used for the transient detection, the better the quality – that is, the detection of transients was better (and the end result improved) when calculating the energy more regularly with smaller window sizes.

Both Kuntz, Disch, Bäckström, Robilliard and Uhle (2011) and Penniman (2014) also employ transient detection as part of the decorrelation process. Kuntz et al. (2011) refer to their implementation as a 'transient steering decorrelator', which works on the same principle proposed by Laitinen et al. (2011), in that the input signal is divided into a transient stream and a non-

transient stream based on frequency band analysis (utilising a Quadrature Mirror Filter-bank (QMF)). The non-transient stream is decorrelated by a pair of all-pass filters, whereas the transient stream is decorrelated by applying slight phase shifts to the signal components. Subjective results showed that use of the transient steering decorrelator slightly improves the overall quality in comparison to all-pass decorrelation of the whole signal. Similarly, Penniman (2014) separates the audio into non-transient and transient streams – the non-transient part is decorrelated using the time-delay approach proposed by Bouéri and Kyriakakis (2004) (with a 24-band filter-bank), while the transient part was subject to random amplitude panning (at intervals of 40 ms). Listening tests conducted by Penniman (2014) demonstrated that the transient extraction approach performed slightly better for listener envelopment, in comparison to decorrelation of the whole signal. However, the proposed approach performed worse for 'stability' with a speech sample, given the random amplitude panning of transient information.

Zotter, Frank, Marentakis and Sontacchi (2011) prefer a deterministic approach to phase-based decorrelation – that is, the same filters are used with each implementation and it does not feature a random element. This kind of approach is designed for two-channel decorrelation and is limiting in terms of multi-channel upmixing, as only a single pair of filters can be generated. The proposed algorithm regularly varies the interchannel time difference (ICTD) over the signal's frequency bandwidth, using either finite impulse response (FIR) or infinite impulse response (IIR) all-pass filter structures, where the depth of ICTD modulation is adjustable and relates to the degree of spread. Furthermore, subjective listening tests demonstrated a clear relationship between decorrelation and the perception of horizontal spread, due to a direct relationship with the IACC. A following development by Zotter and Frank (2013) improved the efficiency of the previous ICTD varying filters and also proposed an interchannel level difference (ICLD) amplitude-based alternative, where the spectral-amplitude is alternated between the two channels over the signal's frequency bandwidth (described in Section 2.3.2 below). Both the phase- and amplitude-based versions work on similar delay networks, where the coefficients of each delay tap are derived from Bessel functions of the first kind. Signal analysis of the frequency-

dependent ICTD and ICLD approaches demonstrates that a decrease of ICCC directly relates to a decrease of IACC, which in turn caused an increase of perceived horizontal spread.

In addition to synthesising a broad phantom image, decorrelation is also used in multichannel parametric coding, where the spatial information (ICC) is transmitted alongside the audio for reconstruction during the decoding stage. MPEG Surround (Herre et al., 2008) is one such standard that applies decorrelation at the output to achieve the original ICC of the input signal. The process involves filtering the input signal using a QMF filter-bank, calculating the ICC for each band, then applying the same ICC to the output bands using lattice all-pass filters. MPEG Surround also features an energy adjustment stage following decorrelation, where the output is scaled to match the energy of the input for each of the QMF bands – this helps to retain transient information and avoid audible reverberation from the decorrelation process (Breebaart & Faller, 2007). Recent developments have seen the emergence of the MPEG-H 3D Audio standard, which also makes use of decorrelation within its Spatial Audio Object Coding for 3D system (SAOC-3D) (Murtaza et al., 2015). The algorithm employs the same QMF and lattice all-pass filter structure as MPEG Surround (Herre et al., 2008), however, has the enhanced ability to decode into higher order loudspeaker systems.

2.3.2 Amplitude-Based Decorrelation Methods

Looking at amplitude-based methods, the simplest of these is in the form of frequency panning, where groups of frequencies are alternately panned between the output channels at regular intervals across the spectrum. Lauridsen (1954) first discovered an effect of increasing image spread when summing and subtracting a signal with a delayed version of itself, creating two comb-filtered signals that had opposing spectral amplitude differences. Schroeder (1958) confirmed this 'pseudo-stereophonic' impression of width by reproducing the comb-filtered signals simultaneously from spaced loudspeaker positions – it has since been termed the "complementary comb-filtering" method (Breebaart & Faller, 2007). The Lauridsen decorrelation process

can be seen in Figure 2.4 below, where the monophonic input signal is time-delayed (T) and multiplied by a gain factor (G) before the summation and subtraction stage.

The Lauridsen method is very simple and cost-effective to implement, and the degree of ICC is easily controlled by the gain factor applied to the delayed signal (between 0.0 and 1.0, where 1.0 is maximum decorrelation). Lauridsen (1954) initially made the discovery when delaying the secondary signal by 50-100 ms; however, the later investigations by Schroeder (1958) confirmed that the artificial stereo effect was perceivable with delays as short as 2.5 ms. Irwan and Aarts (2001) suggest a 10 ms time-delay to be optimal through their informal listening, stating that this delay length provided the desired effect of widening, whilst reducing any confusion that can arise with longer delay times. It has also been suggested that the delay should be frequency-dependent to avoid 'doubles' and 'echoes' at higher frequencies i.e. a time-delay that decreases as frequency increases (Engdegård, Purnhagen, Rödén & Liljeryd, 2004; Breebaart & Faller, 2007); however, there has been no formal investigation into the optimal delay time. An example of the Lauridsen decorrelator output spectra when using a time-delay of 10 ms on the entire signal can be seen in Figure 2.5 below, demonstrating the decorrelation of a 500 Hz pink noise octave-band.

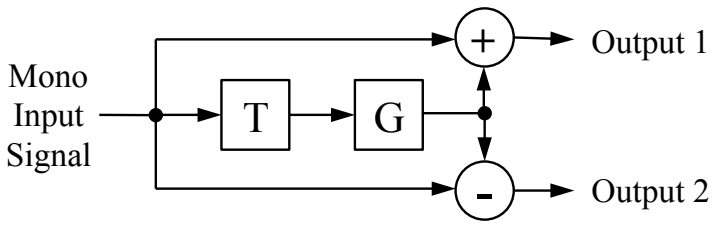

Figure 2.4 Structure of the Complementary Comb-Filter decorrelator (after Breebaart & Faller, 2007).

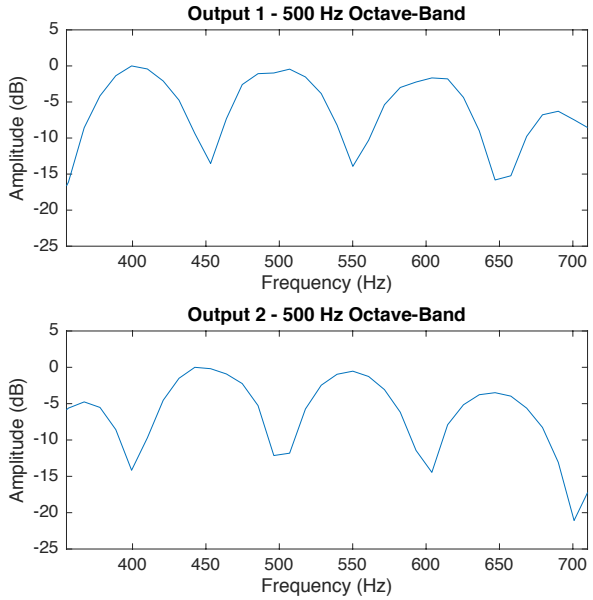

Figure 2.5 FFT plots of the Complementary Comb-Filtered output signals.
(500 Hz octave-band with a 10 ms time-delay and gain factor of 1.0)

Other amplitude-based methods work on a similar principle of frequency distribution between two channels. Fink, Kraft and Zölzer (2015) use two filters with complementary amplitude panning that are more diffuse and less structured than the complementary comb-filter method, as they state that a uniform frequency response sounds unnatural. Faller & Baumgarte (2003) also suggest an amplitude panning approach by randomly modifying the interchannel level difference (ICLD) along the frequency spectrum – this is performed independently between pairs of loudspeakers, and the random ICLDs are time invariant. This randomisation may improve the horizontal spread of sound, rather than rely on a fixed maximum ICLD across all frequencies, as generated by the notches of the comb-filters. Alternatively, Zotter and Frank (2013) propose a multiple delay network, where the gain weightings of each tap are based on Bessel functions of the first kind. This works on a similar principle to the Lauridsen method, where the delay taps for one output are mostly summed with the input, and the delay taps for the other output are subtracted from the input. The result of this is a similar 'frequency-panning' effect between the two outputs to that seen in Figure 2.5 above.

2.3.3 Vertical Decorrelation

The interchannel decorrelation effects discussed above are typically applied between a pair of left and right loudspeakers (i.e. horizontal decorrelation), as a way to increase and control the horizontal image spread (HIS). As far as the author is aware, no formal experiments have been conducted regarding interchannel decorrelation between two vertically spaced loudspeaker sources, as would be the case if it were applied to many modern 3D surround sound formats (Section 2.1.3). In the most similar context, Cabrera and Tilley (2003) used vertically adjacent incoherent noise signals in their vertical localisation and spatial extent experiment (described in Sections 1.1.2.3 and 1.2.1), where subjects were asked to grade the upper, lower, left and right boundaries of the auditory image. A vertically-arranged array of five loudspeakers was assessed in the median plane, with each loudspeaker spaced by 0.28 m, resulting in elevation angles of 0°, ±7.9° and ±15.6°. Decorrelated signals were reproduced through either one, three or five of the active loudspeakers at once, with all conditions level-matched to both 64 and 84 phons. It was reported that the number of vertically-arranged incoherent loudspeaker signals had no significant impact on either the horizontal or vertical extent, when comparing the conditions that have the same centre loudspeaker. In other words, the vertical image spread of one loudspeaker signal at 0° elevation was perceived similarly to that of five incoherent loudspeaker signals at elevation angles of 0°, ±7.9° and ±15.6°, when both conditions were level-matched. The point is made that if a similar array of loudspeakers were arranged horizontally, the number of incoherent loudspeaker signals would have had a more dramatic effect, based on the relationship between the ICCC and IACC horizontally.

Cabrera and Tilley's (2003) study suggests that the effect of vertical interchannel decorrelation might be weak; however, the loudspeaker positions they assessed are not particularly representative of the 3D surround sound formats being established today. As mentioned in Section 2.1.3, the extension of commercial 2D surround sound formats to 3D is typically by way of additional height-channel loudspeakers – these are often elevated by at least +30° to the

listening position, in comparison to the maximum height of 15.6° tested here. Furthermore, only vertical extent in the median plane was assessed, which is also not typical of commercial 3D Surround systems. Given the close proximity of the loudspeakers in the array (0.28 m), the experiment by Cabrera and Tilley (2003) essentially observed vertical extent of multiple real sources adjacent to one another, as opposed to the extent of a phantom image that had been generated vertically by summing localisation. Both of these points indicate that research into vertical decorrelation requires a more practical approach, and investigating the perceptual effect of decorrelation between a spaced pair of loudspeakers at 0° and 30° elevation would be more relevant to the commercial 3D loudspeaker formats currently in use.

Potard and Burnett (2004) also observed the perceived vertical extent of vertically adjacent loudspeakers. Decorrelated signals were presented through two vertical arrays of loudspeakers, ranging from 0-40° and 0-90° in the median plane. It was seen that as the vertical distribution increased (i.e. from 0-40° to 0-90°), the perception of vertical spread also increased. Horizontal arrays of loudspeakers were also assessed during the same experiment, with the results indicating that subjects could discriminate between horizontal and vertical spread from decorrelation. As with Cabrera and Tilley (2003), these results demonstrate that a vertical spread of sound can indeed be perceived in the median plane; however, it remains unrepresentative of a typical commercial 3D loudspeaker system.

2.4 Upmixing to Multichannel Loudspeaker Formats

Upmixing is the process of increasing the channel count of an output, using mostly spatial information provided by the original input signal(s). There are many scenarios where upmixing is useful, for instance, reproducing legacy content on home surround sound systems, as well as the parametric decoding of multichannel audio following transmission. Over the last two decades, many have proposed interchannel decorrelation as an effective method for upmixing and decoding two-channel audio to higher-order surround sound formats (e.g. two-to-five channel upmixing). In this situation, additional diffuse ambient signals are generated by decorrelation and reproduced in the surround channels of a 5.1 system, generating a more enveloping experience for the listener. More recently, these same decorrelation principles have been considered for upmixing, decoding and rendering audio content to 3D multi-channel systems (i.e. with additional height-channel loudspeakers). This section looks at some existing algorithms for upmixing by interchannel decorrelation, as well as those proposed for upmixing to 3D in recent years.

2.4.1 Two-to-Five Channel Upmixing

Two-to-five channel processing refers to the upmixing of a two-channel stereophonic signal (left and right) to 5.1, so that the audio can make effective use of a 5.1 Surround system (Section 2.1). The main upmixing approach is to extract the ambience from the original signal(s) – then decorrelate these ambient signals for reproduction in the 'surround' channels of 5.1. Jot and Avendano (2003) determine that the ambient signal component is generally "weakly correlated and evenly distributed between the channels". An ambience extraction method is employed by Irwan and Aarts (2002) in their upmixing algorithm, who determine a single surround signal by calculating the interchannel cross-correlation (ICC) between the left and right input signals. Since only one ambient signal is defined, interchannel decorrelation is required to reproduce ambience from both surround channels. They propose use of the Lauridsen decorrelator (Lauridsen, 1954) (complementary comb-filtering), suggesting that a time-delay of 10 ms provides

a "compromise between widening and diffuseness", with longer delays amounting to a confusion of the image. Furthermore, to generate a centre channel for 5.1 presentation, Irwan and Aarts (2002) suggest a principal component analysis (PCA) approach (which also contributes to the ambience extraction part). The PCA retrieves the direction of the phantom image and represents it in vector form, from which the gain coefficient of the centre channel is defined. Subjective listening tests demonstrate that the proposed algorithm performs favourably for a listener positioned in the sweet spot as well as off-centre, when compared against two commercial upmixing methods (unnamed). Similar to Irwan and Aarts (2002), Li and Driessen (2005) propose an algorithm that determines the direction of the stereo image using PCA, except prefer to perform this operation on QMF sub-bands rather than the broadband signal. Likewise, ICC between the two input signals is calculated at QMF level to detect the ambient part of the signal. The extracted ambience is then decorrelated using all-pass filters for reproduction in the two surround channels of 5.1.

Avendano and Jot (2004) also propose an upmixing algorithm that extracts ambience from a two-channel stereophonic signal. The ambience extraction method is detailed in a previous article (Avendano & Jot, 2002), where an interchannel coherence index is computed to identify time-frequency regions that feature minimal correlation. After an ambient signal is extracted it is first duplicated, then both signals are processed with a low-pass filter, all-pass filter and a time-delay. The all-pass filter is applied to decorrelate the two ambient signals for reproduction in the surround channels, reducing the likelihood of phantom lateral images and creating a sense of spaciousness. Furthermore, the low-pass filter and delay (typically 10-15 ms) are used to avoid localisation confusion from the precedence effect (Litovsky et al., 1999), ensuring that the frontal image is the focal point of the sound scene.

As mentioned above in Section 2.3.1, decorrelation is often found in parametric multichannel audio coding as well, where multichannel audio is encoded into a single audio channel along with side information (spatial metadata). The side information includes spatial parameters for

the reconstruction of the multichannel audio signals at the decoding stage, which is effectively upmixing from one channel to multiple channels. Faller (2006) discusses the use of decorrelation for the purpose of synthesising the original ICC between multiple channels during the decoding process, where he proposes the use of decaying white noise to model the perception of late reflections. In his algorithm, the ICC spatial cue is updated every 4-16 ms (reducing the reverberation effect), where maximum decorrelation is determined by the length and decay of the filter. Similarly, the standardised MPEG Surround audio coding framework utilises lattice all-pass filters in the QMF domain to synthesise the ICC between the original multichannel input signals (Herre et al., 2008). In recent years, the MPEG-H standard has been developed which works on the same QMF and all-pass filter structure as MPEG Surround; however, it also incorporates object-based coding and the ability to decode audio to 3D multichannel systems (Murtaza et al., 2015).

2.4.2 2D-to-3D Surround Sound Upmixing

It is thought that the same upmixing principles described in Section 2.4.1 might also be applied to 3D surround sound systems. That is, decorrelating ambient signals from 2D content for use in the additional height-channel loudspeakers, in order to generate a more enveloping sense of spatial impression from above. Kraft and Zölzer (2016) propose a 3D upmixing algorithm, albeit not strictly with vertical interchannel decorrelation. It is focused around a mid-side ambience extraction technique that had been described in previous work (Kraft and Zölzer, 2015). Firstly, the azimuth position and panning coefficients of sources are estimated from a two-channel stereophonic input signal – this information allows for the calculation and extraction of both the direct and ambient parts of the signal. The direct part is then panned using VBAP, while the ambient part is decorrelated into the required number of surround channels. The decorrelation method used is the amplitude-based approach by Fink et al. (2015), where random amplitude differences are applied between two signals across the frequency spectrum (see Section 2.3.2). In order to generate the required number of decorrelated signals, a decorrelation

tree structure is proposed. For example (using a multichannel system similar to Auro-3D 9.1 (Auro Technologies, 2015a)), the original ambient signal is decorrelated into two amplitude complementary signals (left and right), then those two ambient signals are decorrelated into two further signals each (left / right front and rear), resulting a total of four decorrelated signals. From this, rather than decorrelate vertically, Kraft and Zölzer (2016) suggest that the four decorrelated ambient signals (Left, Right, Rear Left and Rear Right) are low-pass and high-pass filtered for the main- and height-layers, respectively – this is based on the work of Lee (2016b) mentioned below. Although vertical decorrelation does not feature in this particular algorithm, it certainly indicates the direction in which developments are heading, and it is thought that a deeper understanding of vertical interchannel decorrelation would be useful for this potential application.

A novel 3D upmixing method proposed by Lee (2016b) relies on the 'pitch-height' effect, where higher frequencies inherently tend to be perceived in elevated positions, while lower frequencies are often perceived at or below ear height. This effect and the findings of Lee (2016b) are described in greater detail in Section 1.1.2.3 of the present thesis. Lee (2016b) proposes that frequency bands should be routed to either the main- or height-layer based on their inherent position within space (from the pitch-height effect), in order to generate a natural spread of sound without duplicating or decorrelating signals. An initial assessment by Lee (2016a) in the frontal plane demonstrated that the proposed method generated a slightly greater vertical spread than both all-pass filter and complementary comb-filter decorrelation. These results suggest that vertical decorrelation may be quite ineffective, and that discretely routing different frequency bands to either the main- or height-layer is a better approach for 3D upmixing. Despite this, it is thought that further investigations into vertical interchannel decorrelation would still be beneficial, in order to formally assess the effectiveness of such approaches.

Further to the potential 3D upmixing methods mentioned above, many commercial products (such as home cinema AV receivers) feature proprietary and system specific 2D-to-3D

upmixing algorithms. With the Dolby Atmos system, the algorithm is referred to as the 'Dolby Surround' upmixer. Auro Technologies call their approach the 'Auro-Matic' upmixer, which is also available as a plug-in for rendering 3D audio within a digital audio workstation (DAW). Furthermore, DTS have the 'DTS Neural:X' upmixing algorithm for use with DTS:X systems. All three of these are able to take a two-channel stereophonic input and upmix it to a 3D multichannel surround sound system, presumably using similar upmixing techniques to those described above. Although the processing details are unavailable, some of the algorithms mentioned here are likely to utilise interchannel decorrelation for generating additional ambient signals (potentially for height-channel reproduction). Furthermore, multichannel parametric coding algorithms are also likely to utilise interchannel decorrelation for decoding to 3D surround sound systems (Murtaza et al., 2015). Consequently, it is important to gain a better understanding of how decorrelation is perceived in the vertical domain, as has been investigated in the present thesis.

2.5 Summary

This chapter discussed how sound can be manipulated within loudspeaker reproduction systems – that is, the localisation and spatial impression of reproduced sound sources over loudspeakers. A focus has been placed on the perception of existing spatial techniques and how it applies to surround sound presentation. Section 2.1 provides an overview of loudspeaker formats that are in use, from two-channel stereophony (a pair of left and right loudspeakers) to 3D surround sound systems – where the three main commercial 3D loudspeaker formats at present are Dolby Atmos (Dolby Laboratories, 2015), Auro-3D 9.1 (with various other configurations) (Auro Technologies, 2015a) and DTS:X. It is seen that commercial 3D audio is typically achieved by introducing height-channel loudspeakers above the arrangement of 2D loudspeaker arrays, such as, 5.1 and 7.1 Surround.

In the case of two-channel stereophony, summing localisation occurs when two spaced loudspeakers reproduce coherent (identical) signals, where a 'phantom' auditory image is perceived between the loudspeaker boundaries. This phantom image can be positioned along the horizontal plane by 'panning' (altering interchannel level difference (ICLD) and/or interchannel time difference (ICTD)), as discussed in Section 2.2. The most common approach is to use amplitude panning (ICLD), of which an established method is 'vector base amplitude panning' (VBAP) (using the tangent panning law) (Pulkki, 1997). Studies into amplitude panning between pairs of vertically spaced loudspeakers demonstrate that a phantom image can also be perceived along the vertical plane, however, the localisation of vertical panning angles tends to be largely inaccurate (Pulkki, 2001; Barbour, 2003; Mironovs & Lee, 2017). The apparent perception of a vertical phantom image is of particular importance to the current study, as the main aim is to observe whether the vertical spread of a phantom image can be controlled by vertical interchannel decorrelation. It was further suggested that high frequencies are important to the perception of vertical panning (Mironovs, 2017), which aligns with the findings for vertical localisation in

general (Roffler & Butler, 1968a; Hebrank & Wright, 1974) – it is anticipated that any perceptual effect of vertical interchannel decorrelation would also be reliant on such cues.

Section 2.3 of this chapter focused on controlling the extent of a phantom image by way of interchannel decorrelation. When two signals are decorrelated between a pair of left and right loudspeakers (i.e. the interchannel cross-correlation (ICC) is decreased), a horizontal spread of sound is perceived – this is due to the direct relationship between the ICC and the interaural cross-correlation (IAC) (a known contributor of apparent source width (ASW) in concert halls (Hidaka et al., 1995)). Approaches to achieve a decorrelation of signals can broadly be split into phase-based and amplitude-based. Kendall (1995) proposes the use of all-pass filters, in order to randomise the phase of an input signal, while maintaining spectral unity. Another phase-based approach is suggested by Bouéri and Kyriakakis (2004), where the input is filtered into critical bands and random delays are applied to each band. One particular problem of decorrelation is the processing of transients, which can be smeared by the length of the filter used. Laitinen et al. (2011) demonstrated the use of transient extraction to avoid this, where just the continuous part of the signal is decorrelated which was shown to improve quality. With regard to amplitude-based methods, the simplest of these is complementary comb-filtering (Lauridsen, 1954; Schroeder, 1958; Breebaart & Faller, 2007) – that is, the two signals being decorrelated feature opposing amplitude differences along the spectrum as created by comb-filters. Similar approaches also feature regular amplitude differences (varying ICLD with frequency), which is either implemented randomly (Faller & Baumgarte, 2003; Fink et al., 2015) or through a deterministic delay network (Zotter & Frank, 2013). Vertical interchannel decorrelation is considered towards the end of the section, where it is seen that vertically adjacent loudspeakers with decorrelated signals can be perceived as a vertical image spread (Potard & Burnett, 2004). However, as far as the author is aware, no formal investigation has been conducted into the perception of decorrelation between vertically spaced loudspeakers, or the extent of the vertical phantom image.

The final section of this chapter (Section 2.4) discussed the use of vertical interchannel decorrelation in multichannel upmixing algorithms. When upmixing from two-channel stereophony to 5.1 Surround, the ambient part of the input is generally extracted based on analysis of the ICC. This extracted ambience is then decorrelated between the two surround channels in the 5.1 format to increase listener envelopment (LEV). At present, very few 3D upmixing algorithms have been formally proposed, with those that have preferring a discrete distribution of frequency bands between the main and height loudspeaker layers, rather than vertical interchannel decorrelation (Kraft et al., 2016; Lee, 2016b). However, it is yet to be investigated whether interchannel decorrelation is effective in the vertical domain for such a purpose. Recent multichannel parametric coding standards imply the use of interchannel decorrelation for the decoding of height-channel loudspeaker signals (Murtaza et al., 2015). It would therefore be beneficial to explore this area further, in order to gain a fundamental insight of how decorrelation is perceived in the vertical domain. Applications of vertical interchannel decorrelation are also not limited to 3D upmixing – if it is determined that the vertical extent of a source can be controlled by decorrelation, then this process might also be utilised in the manipulation of objects within object-based 3D surround sound systems.

3 A COMPARISON BETWEEN HORIZONTAL AND VERTICAL INTERCHANNEL DECORRELATION[1,2]

This chapter details two subjective listening tests that have been designed to compare the perceptual effect of interchannel decorrelation in the horizontal and vertical domains, as set out in Section 3.2 below. The first test deals with decorrelation in the horizontal plane between left and right loudspeaker channels (Section 3.3); and the second with vertical decorrelation between a lower (main-layer) and upper (height-layer) spaced loudspeaker pair in the median plane (Section 3.4). All testing was conducted under semi-anechoic conditions and both tests featured the exact same stimuli, in order to gain a fundamental insight of the decorrelation effect. A comparative discussion of the two experiments and objective analysis of stimuli (as well as room signals) are then featured in Section 3.5.

It is well documented in the literature that horizontal interchannel decorrelation between left and right loudspeaker signals relates directly to the perceived width of the horizontal auditory image (Zotter & Frank, 2013). This is due to a strong relationship between the interchannel cross-correlation (ICC) and interaural cross-correlation (IAC). It is known that the IAC coefficient (IACC) is a good indicator of apparent source width (ASW) in concert hall acoustics, as dictated by decorrelated early reflections from lateral directions (Hidaka et al., 1995). In turn, an artificial synthesis of this natural decorrelation controls the horizontal extent of a phantom auditory image between left and right loudspeaker channels. A visual representation of this can be seen on the left in Figure 0.2 of Chapter 0 (Introduction), where the phantom auditory image of the ICCC 1.0 condition (full correlation between the two loudspeaker signals) is narrow and

[1] Gribben, C., & Lee, H. (2014). The Perceptual Effects of Horizontal and Vertical Interchannel Decorrelation, Using the Lauridsen Decorrelator. Presented at the *136th Convention of the Audio Engineering Society*, 9027.

[2] Gribben, C., & Lee, H. (2017). A Comparison between Horizontal and Vertical Interchannel Decorrelation. *Journal of Applied Sciences, 7*(11), 1202.

located directly between the two source positions – decorrelation of the two signals then extends the horizontal image spread (HIS) towards the loudspeakers.

In contrast, very little is known about the psychoacoustic effects of interchannel decorrelation in the vertical domain. Research regarding vertical panning demonstrates that an elevated phantom image is generated between two vertically spaced coherent signals (as represented by the ICCC 1.0 image on the right of Figure 0.2) (Pulkki, 2001; Barbour, 2003; Mironovs & Lee, 2016); however, there has been little investigation into how this phantom image is perceived when the correlation between the two signals is decreased. If the perception of vertical decorrelation were similar to that of horizontal decorrelation, then a decrease of correlation would result in an increase of vertical image spread (VIS) (as proposed in Figure 0.2), possibly mimicking the effect of decorrelated ceiling reflections. It has previously been suggested that a single ceiling reflection can increase the perception of VIS and cause a vertical image shift (Furuya et al., 1995; Robotham et al., 2016), indicating that reflections from above can indeed influence the vertical image. This comparison between horizontal and vertical decorrelation is the focus of the current experiment, where both domains are judged using the same stimuli under identical testing conditions.

Taking into account the perceptual cues of vertical localisation, frequencies above around 3 kHz are important to elevation perception (Roffler & Butler, 1968a; Hebrank & Wright, 1974); it must therefore be considered that any effect of vertical decorrelation might be frequency-dependent as well, potentially with a greater influence from the key high frequency regions related to vertical localisation (specifically around 8 kHz). To observe the spectral impact alone, the effect of vertical decorrelation has been assessed in the median plane where there is little interaural difference, except for that caused by ear asymmetry at high frequencies (Searle et al., 1975). Recently it has been suggested that the decorrelation of ambience signals into height-channels might improve or enhance 3D upmixing algorithms (Kraft & Zölzer, 2016). However, if it is found that interchannel decorrelation is not perceivable in the vertical

domain, then these additional ambient signals might unnecessarily increase the risk of phase cancellation or comb-filtering when the signals combine at the ears.

Considering the above background, the following research questions are proposed:

- Is there a direct relationship between vertical ICC and VIS in the median plane?
- How does vertical decorrelation compare to horizontal decorrelation?
- Is the perception of horizontal and vertical decorrelation frequency-dependent?

In order to answer the above questions, broadband pink noise was filtered into three frequency bands: 'Low' (octave-bands with centre frequencies of 63 – 250 Hz), 'Middle' (centre frequencies of 500 Hz – 2 kHz) and 'High' (centre frequencies of 4 kHz – 16 kHz). These three frequency bands were each decorrelated with varying degrees of ICC and presented in multiple comparison trials (both horizontally and vertically), to observe whether an ICC effect is apparent for the different groups of frequencies.

3.1 Experimental Hypotheses

Firstly, considering the horizontal part of the experiment, the accepted relationship between interchannel cross correlation (ICC), interaural cross-correlation (IAC) and the perceived horizontal image spread (HIS) has already been researched extensively (Zotter & Frank, 2013). Given this, it is hypothesised that decreasing the horizontal ICC between a pair of left and right loudspeakers will increase the perceived HIS (ASW), supporting the existing literature. In terms of frequency-dependency, previous research has demonstrated that the effect of IACC on the perception of ASW varies between different groups of frequencies (Okano et al., 1998; Mason et al., 2005). In concert hall acoustics, an established measure of ASW involves calculating the average IACC of the 500 Hz, 1 kHz, and 2 kHz octave-bands only ($IACC_3$) (Hidaka et al., 1995), which were chosen as the IACC at these frequencies appeared to align with changes to the absolute ASW angle. The 4 kHz octave-band also demonstrated a similar relationship; however, it was not included in the measure as musical signals have less relative energy at higher frequencies, and there is little contribution to ASW above 3 kHz. Furthermore, lower frequencies are considered inherently broad, and it is thought that decorrelation results in a relatively small change to the overall HIS, thus the low frequency exclusion from the $IACC_3$ measure. From this, it is hypothesised that the greatest degree of HIS change by horizontal decorrelation will be observed for the 'Middle' frequency band (500 Hz – 2 kHz octave-bands). The assessment of three separate frequency bands should provide novel insights of decorrelation perception, both horizontally and vertically.

Since previous studies have not formally assessed the effect of interchannel decorrelation in the vertical plane, a focus must be placed on existing vertical localisation and vertical panning literature. As mentioned above, research of vertical panning in the median plane indicates that two discrete coherent signals can be interpreted by the hearing system simultaneously, where an elevated phantom image is perceived between the two spaced positions (Pulkki, 2001; Barbour, 2003; Mironovs & Lee, 2017). As this level of cognition is well established, it is

hypothesised that the hearing system is also able to perceive two partially correlated signals as a phantom image. Furthermore, partial correlation of two discrete signals from independent directions would imply that they emanate from a single source of great spatial extent, given the subtle phase and amplitude differences at arrival. It is also hypothesised that as correlation between the signals decreases, the two signal locations become more independent (yet remain sonically fused due to partial correlation), which in turn causes a relative increase to the perceived vertical image spread (VIS). Moreover, since vertical localisation and vertical panning are most effective for signals with higher frequency content (> ~3 kHz) (Roffler & Butler, 1968a; Hebrank & Wright, 1974; Mironovs & Lee, 2017), it leads to the hypothesis that any effect of vertical decorrelation is likely to be strongest within this frequency region.

3.2 Experimental Design

3.2.1 Stimuli Creation

As discussed in Section 2.3, there are many approaches to achieve a decorrelation of signals – this is mostly through slight phase and/or spectral-amplitude alterations of an input signal to create two partially correlated output signals. For these two experiments, the amplitude-based complementary comb-filtering method has been implemented, due to its simplicity and easy control over the degree of correlation between the output signals. First discovered by Lauridsen (1954) and investigated further by Schroeder (1958), the technique works on a basis of alternating frequency panning across the spectrum. It was found that summing and subtracting an input signal with a delayed version of itself creates a pair of comb-filtered signals with opposing amplitude differences (as demonstrated in Figure 3.1).

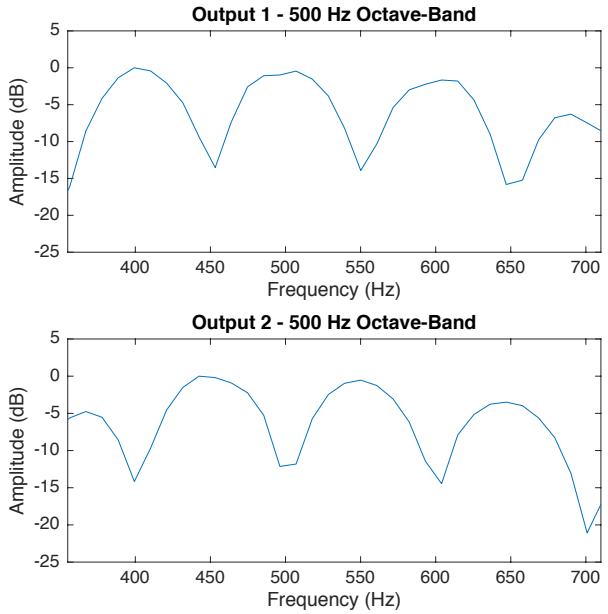

Figure 3.1 FFT plots of the Complementary Comb-Filtered output signals.
(500 Hz octave-band with a 10 ms time-delay and gain factor of 1.0)

The regularity of these amplitude differences (i.e. the 'tooth' bandwidth) is dictated by the time-delay (T) of the secondary signal, and a gain factor (G) applied to the delayed signal controls the notch depth and degree of decorrelation (between 0 and 1, where 1 is maximum

decorrelation) – a block diagram of this process can be seen in Figure 3.2 below. Irwan and Aarts (2002) used a time-delay of 10 ms in their proposed upmixing algorithm, which was determined experimentally as a compromise between adequate widening and avoiding confusion that can be experienced with longer time-delays. As far as the author is aware, there has been no formal assessment of the complementary comb-filtering parameters. It was thought that performing such an assessment within the context of the present experiment would provide a useful insight into the general perception of decorrelation, both horizontally and vertically. To this end, test stimuli were created with 1 ms, 5 ms, 10 ms and 20 ms time-delays for each frequency band; and to observe the effect of interchannel cross-correlation (ICC), the gain factor was set between 0.0 and 1.0 with increments of 0.2. This resulted in six stimuli being compared for each of the four time-delays within a particular frequency band – the six stimuli were judged in a multiple comparison format based on MUSHRA (ITU-R, 2015b) for each time-delay independently (as described further in Section 3.2.4 below).

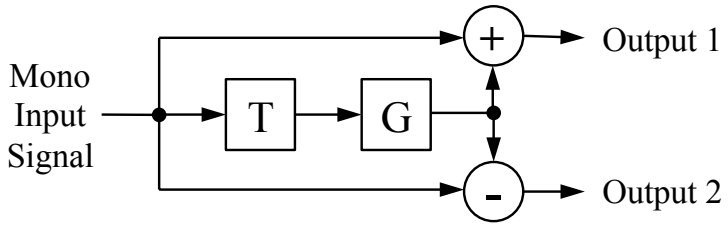

Figure 3.2 Structure of the Complementary Comb-Filter decorrelator (after Breebaart & Faller, 2007).

To assess the frequency-dependency of interchannel decorrelation, a continuous monophonic pink noise sample was band-pass filtered into three frequency bands using the FFT filter function in Adobe Audition. Each frequency band spanned three octave-bands: 'Low' (octave-bands with centre frequencies of 64 Hz, 125 Hz and 250 Hz), 'Middle' (500 Hz, 1 kHz and 2 kHz) and 'High' (4 kHz, 8 kHz and 16 kHz). The Lauridsen decorrelation algorithm was implemented in MATLAB to process the three frequency bands, using the time-delay and gain factor settings described above. This resulted in 12 multiple comparison trials for both the horizontal and vertical domains, made up of each frequency band and time-delay combination

(3×4). During testing, all stimuli were level-matched and presented at an un-weighted sound pressure level (SPL) of ~72 dB(Z), producing a comfortable listening level for the subjects.

To confirm that a variation of ICC is present amongst the stimuli, calculations of the ICC coefficients (ICCC) for all stimuli (taken as the average of 50 ms windows over time ($ICCC_{avg}$), where '1.0' is full correlation) are presented in Table 3.1 below. It can be seen that the resultant $ICCC_{avg}$ values within each gain factor are very similar across all frequency bands, demonstrating the consistency of the method when varying the time-delay and frequency content. A 1 ms time-delay applied to the 'Low' frequency band sees a slight increase of $ICCC_{avg}$ for the middle gain factors ('0.4' to '0.8'); however, the maximum $ICCC_{avg}$ achieved with a gain factor of '1.0' remains suitably low (0.02). Generally, there appears to be an almost linear relationship between $ICCC_{avg}$ and gain factor across all conditions, which provides the broad range of $ICCC_{avg}$ values required for the current experiment.

Table 3.1 Interchannel cross-correlation coefficients (ICCCs) of the complementary comb-filter decorrelated stimuli (calculated as the average of 50ms windows over time).

	Time-Delay	**Gain Factor (G)**					
	(TD)	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
Low	**1 ms**	1.00	0.96	0.85	0.64	0.33	0.02
	5 ms	1.00	0.93	0.74	0.48	0.23	0.04
	10 ms	1.00	0.93	0.73	0.48	0.22	0.07
	20 ms	1.00	0.92	0.73	0.47	0.23	0.11
Middle	**1 ms**	1.00	0.93	0.73	0.48	0.22	0.01
	5 ms	1.00	0.92	0.72	0.47	0.22	0.02
	10 ms	1.00	0.92	0.72	0.47	0.22	0.03
	20 ms	1.00	0.92	0.72	0.47	0.22	0.04
High	**1 ms**	1.00	0.92	0.72	0.47	0.22	0.00
	5 ms	1.00	0.92	0.72	0.47	0.22	0.01
	10 ms	1.00	0.92	0.72	0.47	0.22	0.01
	20 ms	1.00	0.92	0.72	0.47	0.22	0.02

3.2.2 Physical Setup

The listening tests were carried out in a semi-anechoic chamber at the University of Huddersfield, featuring a rubber floor and sound absorption on the walls and ceiling. Further absorption was placed on the floor between the loudspeaker and listener to reduce the effect of floor reflections. All loudspeakers in both experimental parts were hidden from view by an acoustically

transparent curtain, in order to conceal the test setup and avoid any visual bias that may occur. For the horizontal part, two Genelec 8040A loudspeakers (Frequency response: 48 Hz – 20 kHz (±2 dB)) were positioned in a left / right two-channel stereophonic loudspeaker setup with a base angle of 60° (±30° azimuth), positioned at a distance of 1.5 m from the listener and 1.5 m from each other (see Figure 3.3). The listener was positioned at a height so that their ears were in line with the acoustic centre of both loudspeakers.

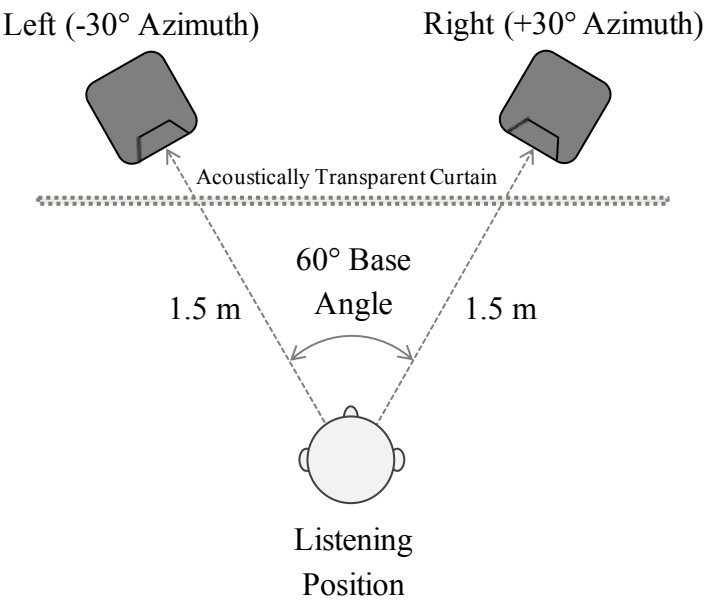

Figure 3.3 Horizontal loudspeaker setup with a 60° base angle.

In the assessment of vertical decorrelation, two Genelec 8040A loudspeakers were vertically-arranged in the median plane, with the lower main-layer loudspeaker positioned 1.5 m in front of the listener, and the upper height-layer loudspeaker elevated by +30° at a distance of 0.9 m directly above the lower loudspeaker (Figure 3.4). The two loudspeaker signals of the vertical pair were time- and level-aligned at the listening position, to accommodate for a difference in distance from source to receiver. As with the horizontal test, loudspeakers were hidden by an acoustically transparent curtain, and the listener was positioned so that their ears were in line with the acoustic centre of the lower (main-layer) loudspeaker.

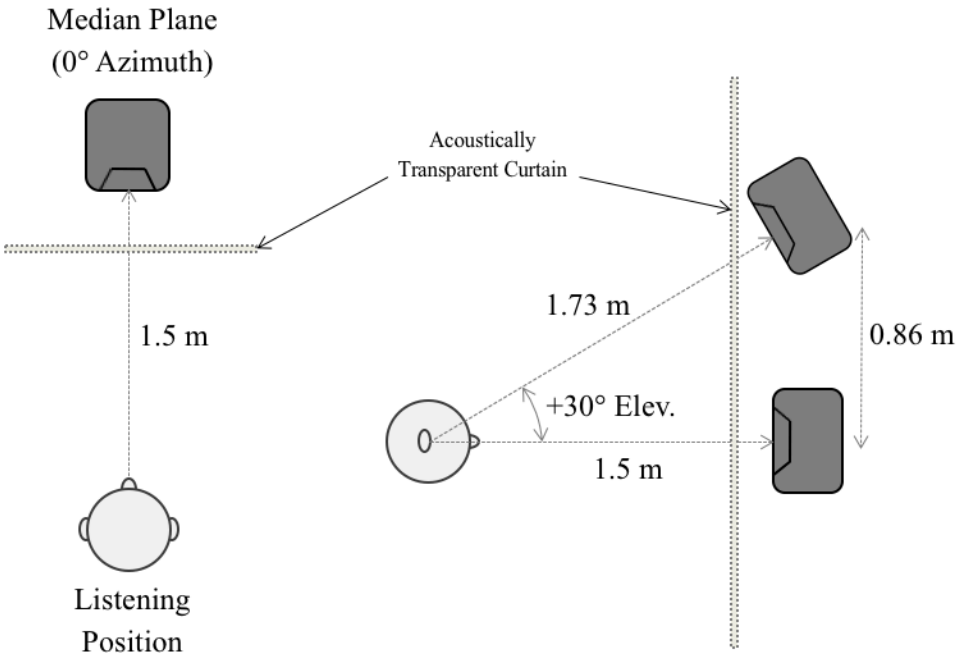

Figure 3.4 Vertical loudspeaker setup at 0° azimuth with a +30° elevation.

3.2.3 Subjects

The horizontal and vertical tests were carried out at separate times to ensure the testing conditions remained constant for each listener. As a result, not all subjects were available to sit both parts of the test. In total, 14 subjects took park in the horizontal test and 13 in the vertical test, with 10 subjects contributing to both. The subjects were trained listeners affiliated with the University of Huddersfield's music technology courses – comprising staff members, final year undergraduate students and post-graduate research students – all of who were experienced with critical listening and analysis of spatial content in a listening test environment.

The use of 13 subjects achieves a minimum statistical power of 0.51 for the current experiment, based on a two-tailed t-test with an effect size of 0.6 and α error probability of 0.05 (type I error), as calculated using G*Power 3.1 (Faul, Erdfelder, Lang & Buchner, 2007). This indicates that the probability of a type II statistical error (false-negative) is 0.49 (β), meaning there is a fairly high chance that some significant (perceivable) differences may be reported as insignificant. In the context of the present study, this is not of real concern as it is likely that the

perceptual differences between these conditions are borderline perceivable, which may reduce even further under real listening conditions (i.e. non-anechoic). In contrast, it is more important that a type I error does not occur (false-positive indication of significance), which is accounted for with the Bonferroni correction in the analysis below. Bech and Zacharov (2007) also state that 5-15 subjects are adequate for subjective critical listening tests when listeners are experienced, as is the case for the current experiments. It is therefore considered that the use of 13 and 14 subjects for each part, respectively, is sufficient for the present set of tests.

3.2.4 Test Method

As previously mentioned, twelve multiple comparison trials were presented for both the horizontal and vertical conditions, made up of each time-delay and frequency band combination. Each multiple comparison trial featured 6 buttons and sliders to control and grade the 6 gain factor stimuli for a particular condition (with gain factors ranging from 0.0 to 1.0 at 0.2 increments). The multiple comparison test format was based on the MUSHRA standard in ITU-R BS.1534-3 (2015b). However, rather than a scale of 0 to 100, a bipolar scale was utilised ranging from -50 to 50, with a button for a reference stimulus positioned at 0 on the scale. The reference chosen was the '0.0' gain factor condition of that particular trial, creating a hidden reference amongst the stimuli. In the case of the horizontal experiment, listeners were asked to grade the relative horizontal image spread (HIS) of each stimulus against each other and the reference; and with the vertical test, listeners were instructed to relatively grade vertical image spread (VIS). The testing interface was constructed in Cycling '74's Max 7 and can be seen in Figure 3.5 below. Subjects could freely switch between stimuli and the reference throughout the test. The order of the presented stimuli and trials were randomised for each listener to reduce any psychological bias. Listeners were trained beforehand by being presented with an example of the grading interface and extreme stimuli (gain factors = 0.0 and 1.0), to familiarise them with the attribute(s) (HIS / VIS) and the format of testing.

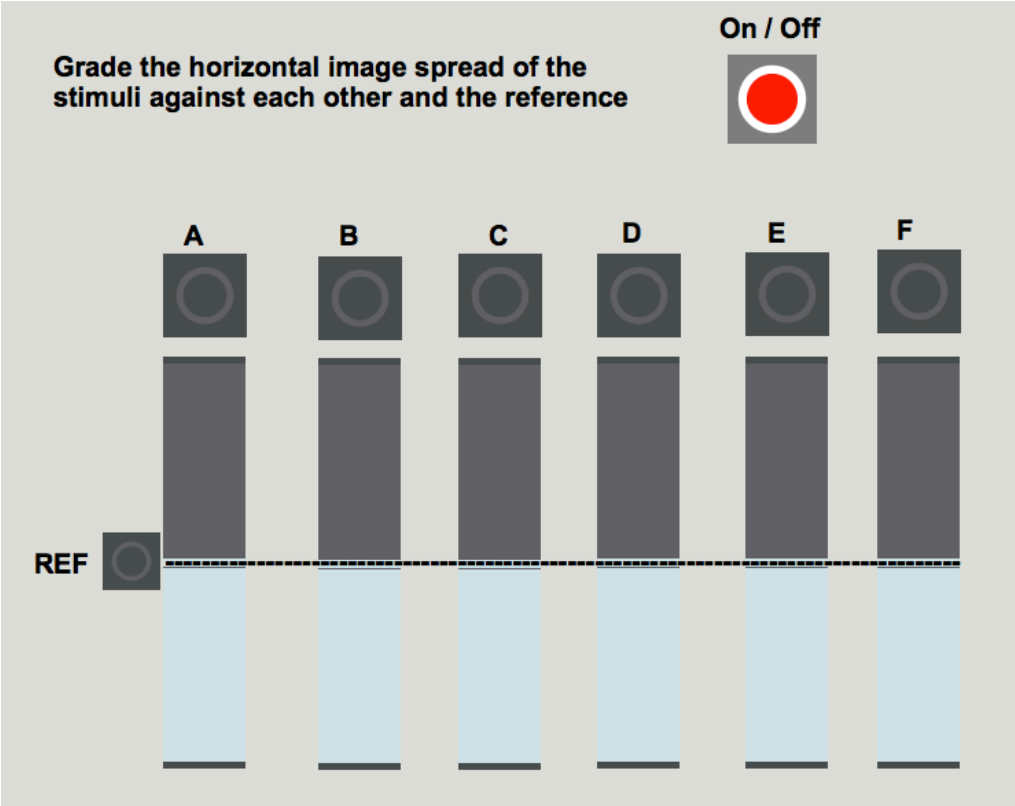

Figure 3.5 Multiple comparison interface used during testing, constructed in Max 7.

A MUSHRA-type multiple comparison test was the preferred listening test format as it is commonly used to detect medium to large differences between auditory stimuli (ITU-R, 2015b). A double-blind triple-stimulus test was also considered (ITU-R, 2015a), however, this approach is more suited to small impairments of quality between stimuli. The double-blind triple-stimulus format means that only one stimulus is assessed against a reference in each trial, leading to a greater testing load for the subject. Moreover, the subject is required to be particularly consistent with their grading between trials, in order for the relative differences between stimuli to be revealed. With this in mind, a multiple comparison is considered to be easier and more efficient for the listener, while also providing useful information on the relative differences between stimuli. Similarly, an ABX (forced choice) test was not utilised as it is often used to

determine whether a difference can be perceived between stimuli, however, this does not reliably determine the extent of that difference and can also be time-consuming when comparing a large number of stimuli.

The MUSHRA test in the current experiment featured bipolar scale to avoid any grading bias, as it is not yet known whether decorrelation can cause the perception of VIS to decrease. The MUSHRA format was initially designed to assess the audio quality of processed signals against an unprocessed 'high quality' reference, which was positioned at 100 on a 0 to 100 scale. However, unlike quality testing, it is impossible to designate a stimulus that has objectively the greatest or least amount of a spatial attribute. Positioning an audible reference in the middle of the scale allows the listener to grade both above and below the reference for a given spatial attribute, as has been employed in previous spatial audio studies (George et al., 2010). Additional detail and discussion on subjective listening tests can also be found in Appendix A, where a listening test interface tool named HULTI-GEN is presented.

3.3 Horizontal Decorrelation: Results and Analysis

Results for the horizontal decorrelation test are presented in Figure 3.6 below – all data has been normalised in accordance with ITU-R BS.1116-3 (2015a) and analysed in SPSS. Given that the scale used during testing was continuous, normalisation is required to compensate for differences of scale use between subjects. For instance, one subject may have only used a small section of the scale to reflect slight differences, while another may have used the full range despite only perceiving slight differences. Data normalisation averages these variations out and involves calculating the mean and standard deviation of data, both for the entire dataset and for the data of each subject independently. Each individual score (x_i) is then altered based on these mean and standard deviation results, as seen in Equation 3.1 below.

$$Z_i = \frac{(x_i - x_{si})}{s_{si}} \cdot s_s + x_s \qquad (3.1)$$

where Z_i is the normalised result, x_i is the score of subject i, x_{si} is the mean score for subject i, s_{si} is the standard deviation for subject i, x_s is the mean score of all subjects and s_s is the standard deviation for all subjects.

The graphs below display the normalised median scores of relative horizontal image spread (HIS) with bars to signify notch edges, representing non-parametric 95% confidence intervals (McGill, Tukey & Larsen, 1978). Shapiro-Wilk tests for normality indicated that the data of each condition was not always normally distributed; therefore, non-parametric statistical tests were performed across all conditions for consistency and comparison. Friedman repeated measure tests have been conducted on each frequency band and time-delay combination, in order to observe the gain factor (ICC) effect on HIS. Where a significant effect is apparent, Wilcoxon pairwise comparison tests have been carried out between each condition of that combination.

Statistical correlation results are presented in Table 3.2, where Spearman's rank-order and Pearson's product-moment coefficients have been calculated. Spearman's rank-order is non-parametric and observes a monotonic relationship, whereas Pearson's product-moment determines

the linearity of a relationship between two variables. Given that the data under analysis is not normally distributed, Spearman's coefficients (r_s) shall be referred to primarily; however, if agreement is seen with the respective Pearson coefficient (r), a linear relationship may also be suggested. Interpretation for both sets of correlation coefficients is as follows: 0.3-0.49 = weak, 0.5-0.69 = moderate, 0.7-0.89 = strong, and 0.9-1.0 = very strong.

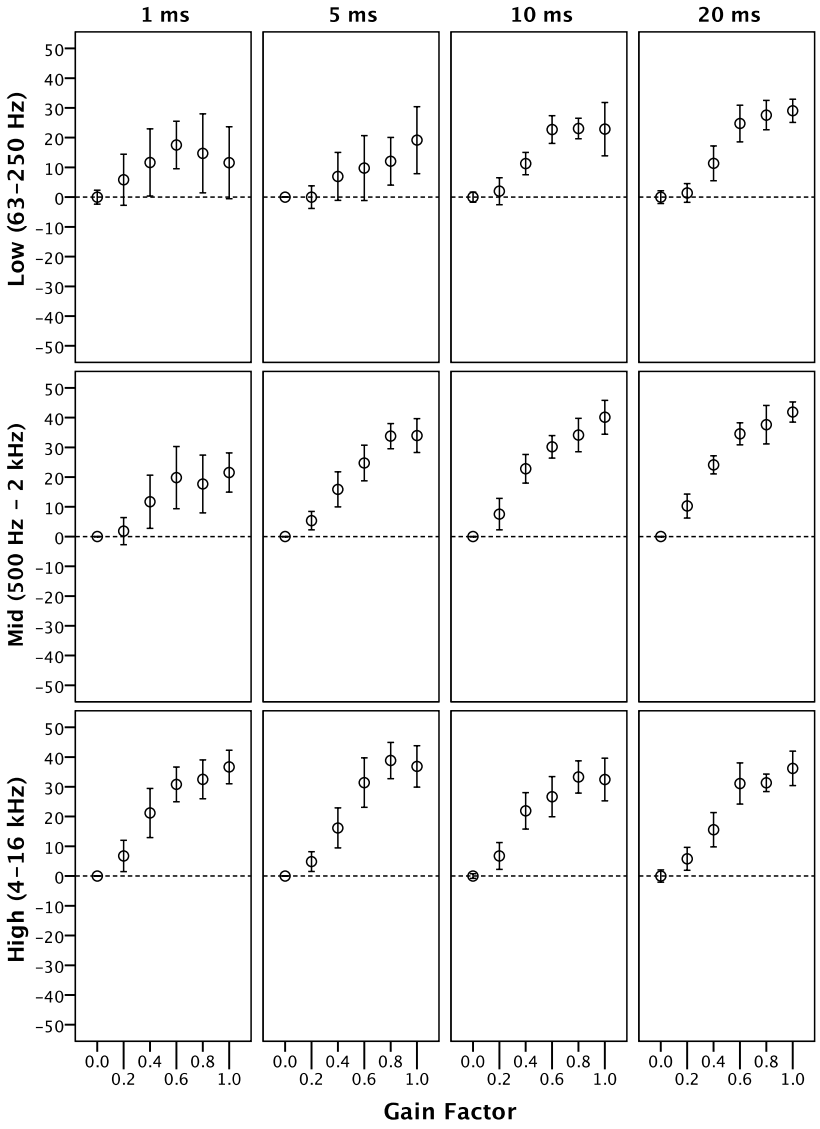

Figure 3.6 Results of the relative horizontal image spread (HIS) by interchannel decorrelation. Median values and notch edges (95% confidence).

Table 3.2 Statistical Correlation between the gain factor of the complementary comb-filtering method and the relative Horizontal Image Spread (HIS) Scores (** $p < 0.01$; * $p < 0.05$)

Frequency Band	Spearman's Rank-Order (r_s)				Pearson's Product-Moment (r)			
	1 ms	5 ms	10 ms	20 ms	1 ms	5 ms	10 ms	20 ms
Low (63-250 Hz)	0.27*	0.46**	0.65**	0.82**	0.25*	0.40**	0.65**	0.83**
Middle (0.5-2 kHz)	0.59**	0.88**	0.80**	0.88**	0.55**	0.87**	0.81**	0.88**
High (4-16 kHz)	0.80**	0.81**	0.78**	0.85**	0.80**	0.81**	0.78**	0.86**

3.3.1 Horizontal Results: Low Frequency Band

The Friedman test results for the 'Low' frequency band (centre frequencies from 63 Hz to 250 Hz) show a significant gain factor effect on HIS for the 5 ms, 10 ms and 20 ms time-delays ($p < 0.01$), but not the 1 ms time-delay ($p > 0.05$). With the 5 ms time-delay, post-hoc Wilcoxon tests indicate there are no significant differences between conditions following Bonferroni correction ($p > 0.05$). Wilcoxon tests with Bonferroni correction on the 10 ms data show significant differences between some conditions – gain factors of '0.6', '0.8' and '1.0' had significantly greater HIS than the '0.0' and '0.2' gain factors ($p < 0.05$); and a gain factor of '0.8' was also significantly greater than '0.4' ($p = 0.03$). For the 20 ms time-delay, the Bonferroni-corrected Wilcoxon results indicate that gain factors of '0.6', '0.8' and '1.0' all had significantly greater HIS than the '0.0', '0.2' and '0.4' gain factors ($p < 0.02$). These results suggest that shorter time-delays are less effective at increasing HIS for lower frequencies, which is further reflected in the statistical correlation results of Table 3.2. It is seen that the statistical relationship between gain factor and HIS also increases as time-delay increases – 10 ms has a moderate relationship between the two variables ($r_s = 0.65$), while the correlation for a 20 ms time-delay is considered strong ($r_s = 0.82$) ($p < 0.01$).

3.3.2 Horizontal Results: Middle Frequency Band

Friedman tests on the 'Middle' frequency band (centre frequencies from 500 Hz to 2 kHz) reveal a significant gain factor effect for all time-delays ($p < 0.01$). The 1 ms Wilcoxon results with Bonferroni correlation show that the gain factor of '1.0' had a significantly greater HIS than the '0.0' and '0.2' gain factors ($p < 0.02$). With the 5 ms time-delay, the Wilcoxon results

demonstrate that the gain factors of '0.6', '0.8' and '1.0' were all significantly greater than the '0.0', '0.2' and '0.4' gain factors ($p < 0.02$), but not each other ($p > 0.05$). For the 10 ms time-delay, all gain factor conditions were significantly greater than '0.0' ($p < 0.04$); and '0.6' and '0.8' were also significantly greater than the '0.2' and '0.4' gain factors ($p < 0.04$). Lastly, the Wilcoxon results for a 20 ms time-delay show that gain factors of '0.4', '0.6', '0.8' and '1.0' all have a significant HIS increase over '0.0' and '0.2' gain factors; and gain factor '1.0' is also significantly greater than '0.4' and '0.6' ($p < 0.04$). To support these results, the statistical correlation coefficients in Table 3.2 demonstrate that time-delays of 5 ms, 10 ms and 20 ms have a strong relationship between gain factor and HIS ($r_s > 0.7$). On the other hand, the correlation for a 1 ms delay is only moderate ($r_s = 0.59$) ($p < 0.01$), suggesting that longer time-delays are also more effective at middle frequencies, similar to that seen with the 'Low' frequency band. In general, the relationship between ICC and HIS appears to be particularly strong for the 'Middle' band, as was hypothesised at the beginning of the chapter.

3.3.3 Horizontal Results: High Frequency Band

The Freidman tests on the 'High' frequency band data (centre frequencies from 4 kHz to 16 kHz) show a significant gain factor effect for each of the time-delays ($p < 0.01$). Statistical correlation coefficients in Table 3.2 also indicate a strong relationship between gain factor and HIS for all time-delays ($r_s > 0.7$) ($p < 0.01$). With the 1 ms delay, Bonferroni-corrected Wilcoxon results indicate that a gain factor of '1.0' had significantly greater HIS than all other conditions (0.0-0.8) ($p < 0.05$); and '0.6' and '0.8' were also significantly greater than the '0.0' and '0.2' ($p < 0.04$). For 5 ms, '1.0' and '0.8' were significantly greater than '0.0', '0.2' and '0.4' ($p < 0.05$). The 10 ms results show that gain factors of '0.4', '0.6', '0.8' and '1.0' were all significantly greater than '0.0' and '0.2' gain factors ($p < 0.02$); and '0.8' is also significantly greater than a gain factor of '0.4' ($p = 0.03$). Finally, for the 20 ms time-delay, a gain factor of '1.0' was significantly greater than the '0.0', '0.2', '0.4' and '0.6' gain factors ($p < 0.02$); and '0.6' and '0.8' were significantly greater than '0.0', '0.2' and '0.4' ($p < 0.02$).

3.4 Vertical Decorrelation: Results and Analysis

The results for the vertical decorrelation part of the experiment are presented in Figure 3.7 below, displaying the median and notch edge values (non-parametric 95% confidence equivalent). As with the HIS results, the relative vertical image spread (VIS) data was normalised in accordance with ITU-R BS.1116-3 (2015a) (see Section 3.3) and analysed in SPSS. Shapiro-Wilk tests of normality revealed that not all conditions had normally distributed data; as a result, non-parametric statistical tests were used to assess for significance within the data. The same statistical testing process was used as with the horizontal results, where Friedman tests were initially conducted to observe the gain factor effect within each time-delay and frequency band combination. Then, if a significant effect was detected, pairwise Wilcoxon tests with Bonferroni correction were performed between conditions to identify any significant difference. Statistical correlation results are presented in Table 3.3 below, where both Spearman's rank-order and Pearson's product-moment coefficients have been calculated. As described above, Spearman's test is non-parametric and looks at a monotonic relationship, whereas Pearson's observes the linearity of correlation – both approaches can be interpreted as follows: 0.3-0.49 = weak, 0.5-0.69 = moderate, 0.7-0.89 = strong, and 0.9-1.0 = very strong.

Table 3.3 Statistical Correlation between the gain factor of the complementary comb-filtering method and the relative Vertical Image Spread (VIS) Scores (** $p < 0.01$; * $p < 0.05$)

Frequency Band	Spearman's Rank-Order (r_s)				Pearson's Product-Moment (r)			
	1 ms	5 ms	10 ms	20 ms	1 ms	5 ms	10 ms	20 ms
Low (63-250 Hz)	0.42**	0.49**	0.63**	0.58**	0.45**	0.45**	0.66**	0.57**
Middle (0.5-2 kHz)	0.41**	0.13	0.34**	0.64**	0.39**	0.16	0.31**	0.63**
High (4-16 kHz)	0.59**	0.06	0.29**	0.62*	0.54**	0.02	0.22*	0.63*

3.4.1 Vertical Results: Low Frequency Band

Friedman tests of the 'Low' frequency band data (octave-band centre frequencies from 63 Hz to 250 Hz) reveal that all time-delays have a significant gain factor effect ($p < 0.01$). Post-hoc Wilcoxon tests with Bonferroni correction on the 1 ms data indicate that gain factors of '0.6'

and '1.0' had significantly greater VIS than '0.0' and '0.2' ($p < 0.05$); furthermore, the '0.8' gain factor was also significantly greater than '0.0' ($p = 0.03$). For the 5 ms and 10 ms time-delays, the Wilcoxon tests show no significant difference between any conditions, following Bonferroni correction ($p > 0.05$). With the 20 ms delay, a gain factor of '0.8' had significantly greater VIS than '0.2' ($p < 0.05$). Furthermore, the statistical correlation results in Table 3.3 suggest that the relationship between gain factor and VIS is moderate for the 10 ms and 20 ms time-delays ($r_s = 0.58$-0.63), but weaker for the 1ms and 5 ms delays ($r_s < 0.5$) ($p < 0.01$).

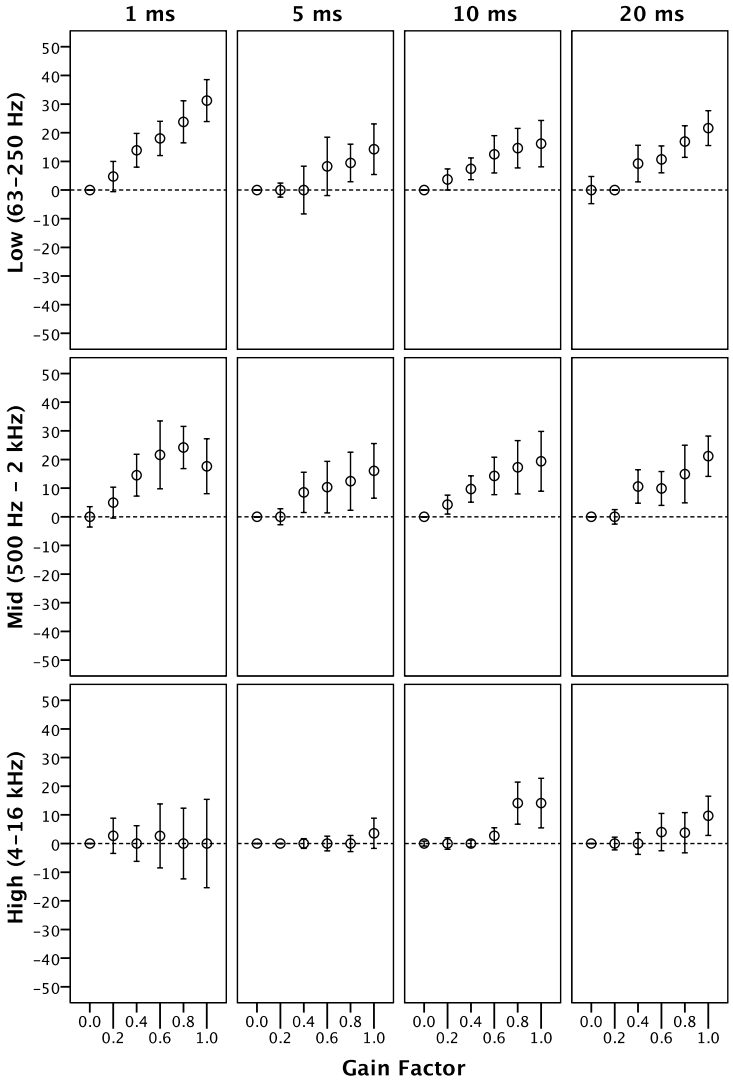

Figure 3.7 Results of the relative vertical image spread (VIS) by interchannel decorrelation. Median values and notch edges (95% confidence).

3.4.2 Vertical Results: Middle Frequency Band

The Friedman results for the 'Middle' frequency band (octave-band centre frequencies from 500 Hz to 2 kHz) show a significant gain factor effect for all time-delays ($p < 0.05$). The post-hoc Wilcoxon tests indicated no significant difference between gain factors for both the 1 ms and 5 ms delays, following Bonferroni correction ($p > 0.05$). On the other hand, with a 10 ms time-delay, gain factors of '0.4', '0.6' and '1.0' had significantly greater VIS than '0.0' and '0.2' ($p < 0.04$); and for the 20 ms delay, a '1.0' gain factor was perceived as significantly greater than '0.2' ($p = 0.03$). The statistical correlation results in Table 3.3 indicate a moderate correlation between gain factor and VIS for the 20 ms time-delay ($r_s = 0.64$), however, the relationship for the other time-delays is considered weak ($r_s < 0.5$) ($p < 0.01$).

3.4.3 Vertical Results: High Frequency Band

Results from the Friedman tests on the 'High' frequency band data (octave-band centre frequencies from 4 kHz to 16 kHz) reveal a significant gain factor effect for the 5 ms, 10 ms and 20 ms time-delays ($p < 0.04$), but not the 1 ms delay ($p > 0.05$). The Bonferroni-corrected Wilcoxon tests on the 5 ms and 20 ms data show no significant difference between any of the conditions ($p > 0.05$). However, the 10 ms Wilcoxon tests reveal that a gain factor of '1.0' produces a significantly greater VIS than a gain factor of '0.0' ($p < 0.05$). Furthermore, Table 3.3 indicates that the statistical correlation between gain factor and VIS is moderate for the 1 ms and 20 ms time-delays ($r_s = 0.59$-0.62), but very weak for the 5 ms and 10 ms delays ($r_s < 0.3$) ($p < 0.01$).

3.5 Discussion of Results

Comparing both sets of results, there is a noticeable difference of relative spread perception between horizontal and vertical interchannel decorrelation. The horizontal results show that decorrelation is similarly effective at increasing horizontal image spread (HIS) for all frequency bands – however, a longer time-delay is required for the 'Low' and 'Middle' frequency bands to generate significantly greater levels of HIS. In contrast, vertical decorrelation in the median plane appears to be most effective for the 'Low' frequency band, though with little significant difference between conditions (unlike the results for horizontal decorrelation). The statistical correlation between gain factor and vertical image spread (VIS) is also noticeably lower for all vertical decorrelation conditions, in comparison to those for horizontal decorrelation. This suggests that, although changes to VIS by vertical decorrelation are observed in the median plane, the effect is weaker than that of horizontal decorrelation between a pair of left and right loudspeakers.

It is assumed that the significant effect of decorrelation on HIS is directly related to the interaural cross-correlation (IAC), which is well documented in the literature (Zotter & Frank, 2013). It is interesting to note that significant changes of horizontal decorrelation were perceivable for all three frequency bands, whereas the $IACC_{E3}$ measurement considers that the greatest contributors of ASW lie within the 'Middle' band (500 Hz – 2 kHz octave-bands) (Hidaka et al., 1995). For the 'Low' band, it may be that these frequencies are mostly correlated in concert halls when summing at the ear, thus providing no contribution to the measurement of IACC. However, when two low frequency signals are artificially decorrelated between a left and right loudspeaker pair, the differences of interaural correlation would likely be greater, seemingly causing an increase of HIS. High frequencies were not included in the $IACC_{E3}$ measure due to a lack of reflection energy at higher frequencies in concert halls – with that in mind, the results presented here strongly suggest that measuring the IACC of high frequencies can also contribute to the measurement of HIS. A greater consideration of higher frequencies (4 kHz to 16 kHz

octave-bands) could be the basis of accurate HIS measurement in surround sound reproduction, where high frequency energy may be greater.

3.5.1 Low Frequency Band Discussion

It is apparent that the perception of the 'Low' frequency band differs between horizontal and vertical decorrelation. In order to observe the effect of time-delay on the source signals for the 'Low' band, Figure 3.8 displays the difference of spectrum between the two output channels, with gain factors of '0.2', '0.6' and '1.0' for each time-delay. Spectra were calculated as the long-term average FFT using 4096 FFT points and a frame size of 4096 samples (with 50% overlapping windows and no spectral smoothing). In the plots, a positive amplitude indicates a bias towards the right / height loudspeaker channel (for horizontal / vertical decorrelation, respectively), whereas a negative amplitude is a bias to the left / main loudspeaker channel.

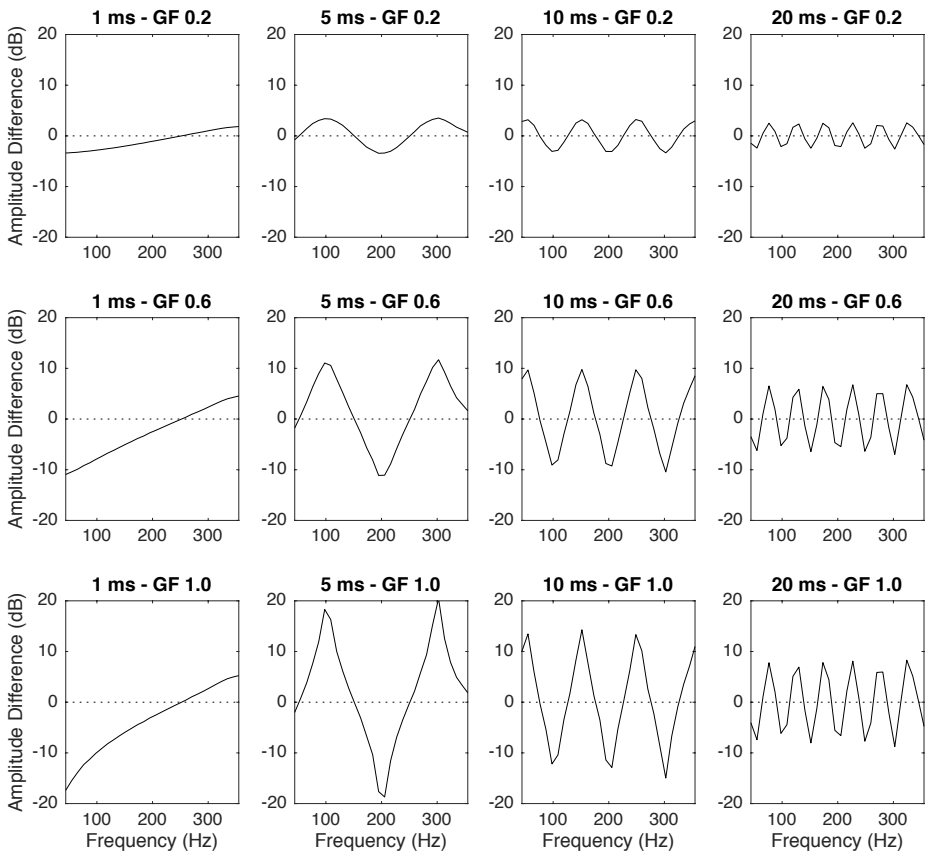

Figure 3.8 Delta spectra between the output signals for the 'Low' frequency band: 'S2 – S1', where S1 is the left/main source signal and S2 is the right/height source signal.

From the plots in Figure 3.8, it can be seen that as time-delay decreases, the distribution of frequencies between the two channels becomes unbalanced. This is further reflected in Table 3.4, where the RMS of the two output signals has been calculated for each time-delay, using a gain factor of '1.0'. With a 1 ms time-delay and gain factor of '1.0', all frequencies below around 250 Hz are boosted in the left / main loudspeaker channel, resulting in an RMS difference between the two channels of 4.4 dB. In the case of horizontal decorrelation, this bias to the left loudspeaker may have caused the greater deviation of responses seen in the 'Low' frequency band subjective results (as suggested by the larger error bars for the 1 ms time-delay in Figure 3.6). Whereas for vertical decorrelation, the uneven frequency distribution with a 1 ms time-delay would have resulted in more energy in the lower main-layer loudspeaker below 250 Hz, potentially causing an increase of perceived loudness (despite SPL level-matching the conditions). Such a change in perceived loudness could have caused the greater perception of VIS seen for the 1 ms time-delay in Figure 3.7, potentially from an enhanced floor reflection; having said that, the statistical correlation between gain factor and VIS for a 1 ms delay remains weak ($r_s = 0.42$). These results suggest a 1 ms time-delay is unsuitable for decorrelating low frequency content. In future experimentation, it may also be useful to RMS level-match the two decorrelated outputs of the complementary comb-filtering method, in order to reduce a bias of energy towards one loudspeaker channel.

Table 3.4 RMS of the two output channels for the 'Low' frequency band with a gain factor of 1.0.

	Left / Main Channel	**Right / Height Channel**
1 ms	-5.8 dB	-10.3 dB
5 ms	-7.2 dB	-7.9 dB
10 ms	-7.4 dB	-7.6 dB
20 ms	-7.5 dB	-7.5 dB

While uneven frequency distribution may account for the 1 ms vertical decorrelation results, VIS change was also seen for longer time-delays at low frequencies (though the differences were largely insignificant). As suggested above, a floor reflection may have influenced the

perception of VIS, particularly when more energy is present in the lower main-layer loudspeaker. To look for a potential effect of the listening room on VIS perception, binaural room impulse responses (BRIRs) of the semi-anechoic chamber were captured using the HAART impulse response toolbox (Johnson, Harker & Lee, 2015), which utilises the exponential sine sweep approach (Farina, 2000). Sine sweeps were reproduced from both the main- and height-layer loudspeakers independently, with the signals captured by a Neumann KU 100 dummy head located in the listening position. The main- and height-layer BRIRs were then time- and level-aligned, before being summed together to replicate the vertical test condition.

Figure 3.9 displays the FFT of the summed BRIRs, calculated using 4096 FFT-points and a frame length of 4096 samples (with 50% overlapping Hann windows). On inspection of the spectrum, a large notch can be seen in the low frequency region around 140 Hz, as well as smaller notches in the 'Middle' frequency band (up to around 2 kHz). Given the regularity of the notches, it suggests a comb-filter effect due to a first reflection interacting with the direct sound – presumably from the rubber flooring of the semi-anechoic chamber (despite placing absorption on the floor between the listener and the loudspeakers). The first frequency notch of comb-filtering when two similar signals interact can be determined by Equation 3.2 below. The main-layer loudspeaker was located at 1.15 m above the ground and 1.5 m from the listening position – this results in a floor reflection path of around 1.25 m greater than the direct signal, with a delay of ~3.6 ms between their arrival at the ear. From this, it is calculated that the first comb-filter notch from a floor reflection should theoretically occur at 139 Hz – the similarity between this and the large notch observed in Figure 3.9 suggests that a floor reflection is indeed present. Previous research has shown that a single ceiling reflection can increase the perception of VIS (Furuya et al., 1995; Robotham et al., 2016). It is possible that a single floor reflection may also have a similar effect on the vertical image.

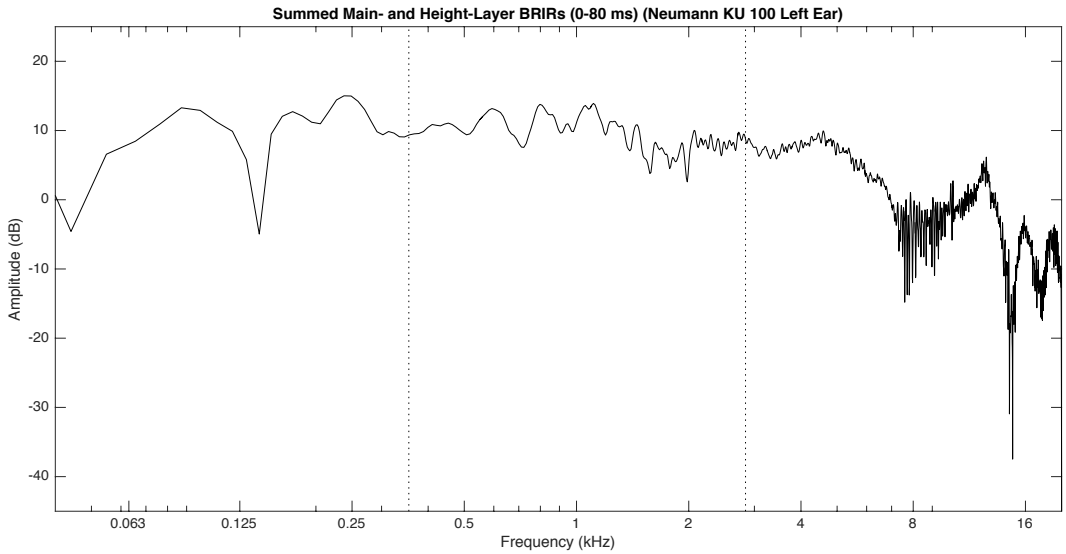

Figure 3.9 FFT of the summed main- and height-layer BRIRs (0-80 ms) from the semi-anechoic chamber. The vertical dotted lines signify the cut-off frequency between the frequency bands.

$$f = 1/(2t) \tag{3.2}$$

where t is the time-delay between signals and f is the first notch frequency of the comb-filter.

Further analysis of the summed main and height BRIRs show that the ratio of early reflection energy to direct sound energy (ER/D) (Equation 3.3) is noticeably greatest in the 'Low' frequency band (-1.9 dB) (Table 3.5) – in other words, the floor reflection observed in Figure 3.9 is likely to have been heavily weighted with low frequency energy. Hypothetically, a decorrelation of enhanced reflections might have led to a further increase of VIS with the 'Low' frequency band. If this were the case, it is possible that the subjective results presented here are specific to the listening environment in which the testing was conducted. However, despite this, the results still indicate that some change of VIS is perceivable at low frequencies, with further investigation required to ascertain the exact cause of the perception.

$$ER/D\ Energy\ Ratio = 10\log_{10}\left(\frac{\int_{t2}^{t3} x^2 dt}{\int_{t1}^{t2} x^2 dt}\right) \tag{3.3}$$

where x is the impulse signal, $t1$ is 0 ms, $t2$ is 2.5 ms and $t3$ is 80 ms.

Table 3.5 Early Reflection Energy (2.5-80 ms) to Direct Sound Energy (0-2.5 ms) Ratio (ER/D) for the summed main- and height-layer BRIRs

Low Frequency Band	Middle Frequency Band	High Frequency Band
-1.9 dB	-14.2 dB	-17.3 dB

3.5.2 High Frequency Band Discussion

On further inspection of the vertically summed BRIR spectra in Figure 3.9, large notches can also be seen within the 'High' frequency band – these are presumably due to HRTF filtering at the pinna (Hebrank & Wright, 1974). To investigate the effect of the HRTF further, the vertical stimuli have been convolved with the sum of two anechoic head-related impulse responses (HRIRs) from MIT's KEMAR dummy head database (Gardner & Martin, 1994) – where one HRIR represents the main-layer loudspeaker angle (0° azimuth, 0° elevation), and the other the height-layer loudspeaker (0° azimuth, +30° elevation). The choice of convolution with KEMAR HRIRs rather than the BRIRs captured in the semi-anechoic chamber was to maintain consistency with the spectral analysis of octave-band pink noise in Chapter 5.

The HRIR-convolved stimuli spectra have been plotted in Figure 3.10 below for each time-delay, with gain factors of only '0.0', '0.4' and '1.0' to improve the clarity of the plots – these are long-term averaged FFTs calculated using 4096 FFT-points and a frame length of 4096, with 1/96 octave spectral smoothing and 50% overlapping Hann windows. The spectra in Figure 3.10 display similar high frequency spectral notches to those observed in Figure 3.9 – these notches are around 11.5 kHz and 17 kHz, and appear greatest when the signals are correlated (gain factor = 0.0). As the gain factor increases (i.e. as correlation between the main- and height-channels decreases), a spectral boost occurs in these regions and the notches become 'filled in'. This is most apparent for the 10 ms and 20 ms time-delays, whereas with shorter delays, a comb-filtering effect occurs that noticeably distorts the definition of the notches. If the spectral notches (and subsequent filling from decorrelation) are an important cue for VIS perception, the increased depth of comb-filter distortion seen for shorter time-delays would inevitably have

an impact on the detection of such cues. This is reflected in the vertical decorrelation results (Figure 3.7), where larger error bars are seen for the 1 ms time-delay, and a significant gain factor effect is only apparent for the 5 ms time-delay and above. Furthermore, the only significant difference between individual gain factor conditions for the 'High' frequency band was with a time-delay of 10 ms.

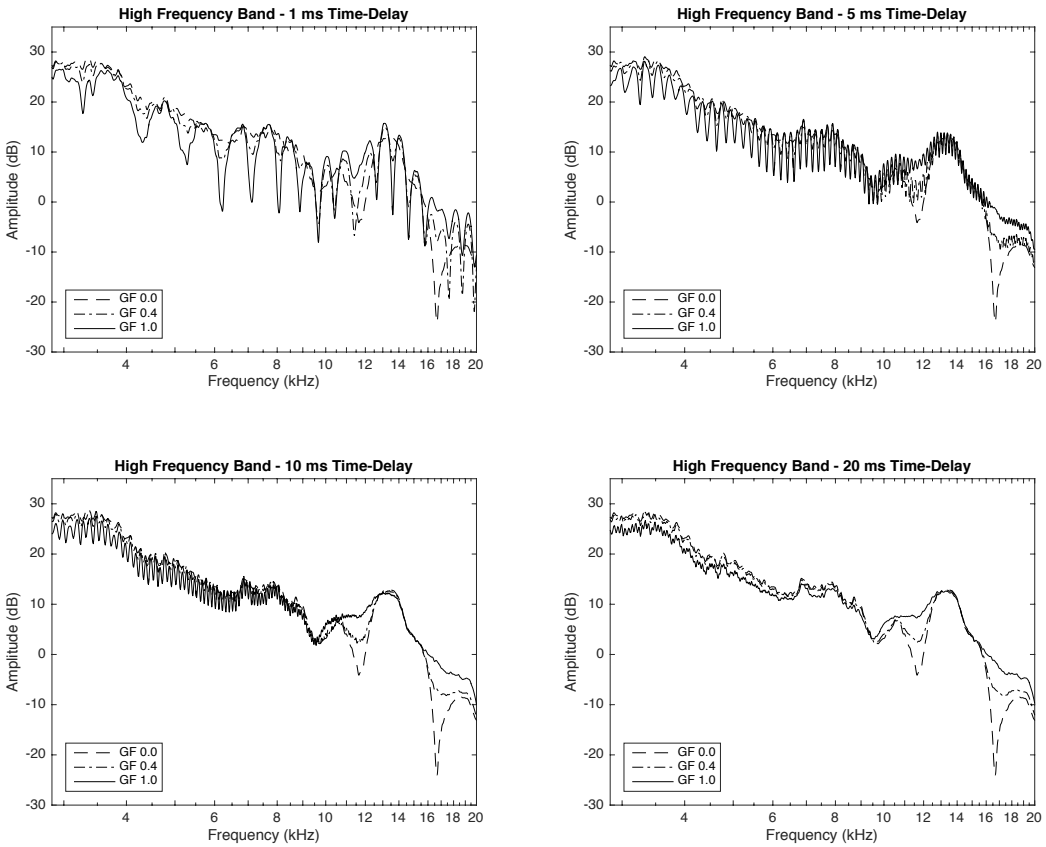

Figure 3.10 FFTs of HRIR-convolved stimuli for the vertical decorrelation conditions. Gain factors of 0.0, 0.4 and 1.0 are plotted for each time-delay.

Despite clear spectral changes in the 'High' frequency band as correlation decreases, little significant change of VIS is seen between the different gain factor conditions in the subjective results. It is possible that this may relate to the un-weighted sound pressure level (SPL) used during testing. It is known from the literature that presentation level has an impact on the perceived extent of a source, where a greater level increases the size of an auditory event (Cabrera

& Tilley, 2003). To quantify the differences in loudness between the frequency bands, Table 3.6 displays the LUFS (LKFS) values (ITU-R, 2015c) for the correlated stimuli source signals used during testing, as calculated by Adobe Audition. The results show a +4 dB increase of loudness for the 'High' frequency band compared to the 'Low' frequency band (when both have been SPL level-matched). Table 3.6 also displays the LUFS values calculated for pink noise that has been filtered into the frequency bands used during testing. Pink noise has equal energy for each octave-band, and is thought to roughly represent the typical octave-band relationship within a complex signal. The pink noise LUFS results demonstrate that the relative loudness of the 'High' frequency band is considerably lower than that of the 'Low' frequency band. When comparing this against the LUFS of the test stimuli, it is clear that the levels used during testing are not representative of a typical frequency relationship found in a complex source. Given that the loudness of the high frequency stimuli was comparatively high during testing, it may have resulted in an increased perception of VIS for all conditions, resulting in more subtle VIS changes from decorrelation.

Table 3.6 A comparison of loudness level (LUFS) between the correlated stimuli used during testing (gain factor = 0.0) and pink noise filtered into the three frequency bands.

	Low Frequency Band	**Middle Frequency Band**	**High Frequency Band**
Test Stimuli Signals	-16 dB LUFS	-14 dB LUFS	-12 dB LUFS
Filtered Pink Noise	-16 dB LUFS	-22 dB LUFS	-27 dB LUFS

The loudness level of the 'High' frequency band signals could also be related to a lack of reflective energy at high frequencies. As seen in Table 3.5, the 'Low' frequency band has the greatest amount of early reflective energy, suggesting that less amplification would be required to meet the target SPL. In contrast, given the greater absorption at high frequencies in the semi-anechoic chamber, it is thought that the 'High' frequency band would require more amplification at line level to match the same SPL. To better assess the frequency-dependency of VIS, it would be useful to observe changes for octave-band noise stimuli, which are reproduced at the inherent SPL of each octave-band within a broadband pink noise signal (i.e. maintaining the

energy relationship apparent within pink noise) – this experiment has been conducted in Chapter 4 of the present thesis.

Another reason for a lack of significant VIS difference between the 'High' frequency band conditions could be related to the "pitch-height effect" (Cabrera & Tilley, 2003; Wallis & Lee, 2016; Lee, 2016b) and "directional bands effect" (Blauert, 1969/70; Wallis & Lee, 2015a). From the directional band research, it is known that 4 kHz and 16 kHz bands tend to be perceived in front, and an 8 kHz band is often perceived above, under anechoic conditions. Similarly, when octave-band noise signals are presented at ear height from in front of the listener, a "pitch-height effect" occurs which sees the 8 kHz octave-band elevated upwards (towards the position of the height-channel loudspeaker); whereas the 16 kHz band is localised towards the main-channel loudspeaker, and 4 kHz is perceived somewhere between the two (Cabrera & Tilley, 2003; Lee, 2016b).

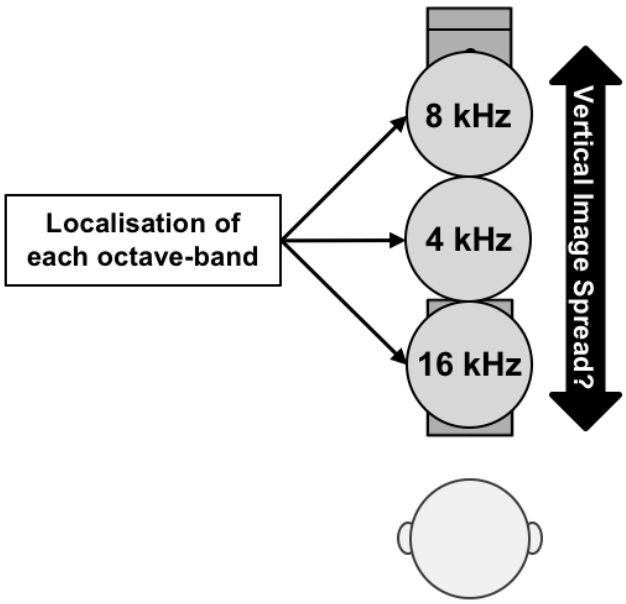

Figure 3.11 Possible perception of vertical image spread (VIS) for the 'High' frequency band, based on the "pitch-height effect" phenomenon (Cabrera & Tilley, 2003; Lee, 2016b; Wallis & Lee, 2016).

Wallis and Lee (2016) have demonstrated that the pitch-height effect also occurs when coherent octave-bands are presented in vertical stereophony i.e. the same signal reproduced in a main-layer loudspeaker and a height-layer loudspeaker simultaneously (both of which are located in the median plane, similar to the vertical test condition in the current experiment). It is thought that this natural vertical spread of frequencies may also be apparent in the 'High' frequency band signals, resulting in an initial broad VIS when both signals are correlated (gain factor = 0.0), with relatively small changes to VIS from decorrelation. The potential perception of this has been illustrated in Figure 3.11 above, showing the possible distribution of octave-bands across the frontal vertical image. To investigate the hypothesis further, both the relative VIS and absolute VIS of octave-band stimuli have been assessed in Chapter 4, to observe how narrow frequency bands are affected by vertical decorrelation.

3.6 Conclusion

Two listening tests have been conducted to observe the effects of interchannel decorrelation both horizontally and vertically. Decorrelated stimuli were generated using the complementary comb-filtering decorrelation method, where frequencies are alternately distributed between two channels throughout the spectrum. The decorrelation method is controlled by two variables: time-delay and gain factor. Time-delay determines the bandwidth between the comb-filter notches, and the gain factor defines the notch depth, which controls the degree of decorrelation (between 0.0 and 1.0, where 1.0 is maximum decorrelation). Variations of these variables were assessed during testing, with time-delays of 1 ms, 5 ms, 10 ms and 20 ms, and six gain factors from 0.0 to 1.0 at increments of 0.2. These conditions were applied to three frequency bands: 'Low' (octave-bands with centre frequencies of 63 Hz, 125 Hz and 250 Hz), 'Middle' (centre frequencies of 500 Hz, 1 kHz and 2 kHz), and 'High' (centre frequencies of 4 kHz, 8 kHz and 16 kHz).

For the horizontal decorrelation, the two decorrelated signals were routed to left and right loudspeakers, respectively, with a base angle of 60° (±30° azimuth). During the horizontal test, subjects were asked to grade the relative horizontal image spread (HIS) between the different gain factor stimuli. With the vertical decorrelation test, signals were decorrelated in the median plane (0° azimuth), between a main-layer loudspeaker positioned at ear height (0° elevation) and another positioned directly above at an elevation of +30°. Subjects were asked to grade the vertical image spread (VIS) between stimuli.

The key findings from the listening tests are as follows:

- A significant effect of interchannel decorrelation on auditory image spread is apparent both horizontally and vertically, where spread increases as correlation decreases.
- The decorrelation effect appears to be stronger in the horizontal domain, with moderate-strong statistical correlation between gain factor and HIS for all frequency bands.

- Vertical decorrelation also leads to significant increases of VIS, however, the relationship between gain factor and VIS appears to be weaker.
- The results also suggest that the perception of vertical decorrelation is frequency-dependent, with VIS change most apparent in the 'Low' frequency band.
- Perception of vertical decorrelation for the 'Low' and 'Middle' frequency bands could potentially be related to floor reflections within the listening room.
- Vertical decorrelation of the 'High' frequency band appears to be associated with spectral notches, which may act as cues for the perception of VIS.
- The 'High' frequency band may have been influenced by the 'pitch-height' phenomena, where different high frequencies are perceived at difference elevations.
- A time-delay of 1 ms causes an uneven distribution of frequencies in the 'Low' frequency band, making it unsuitable for low frequency band-limited decorrelation.

The above findings suggest that vertical interchannel decorrelation has some influence on the perception of VIS in the median plane. However, the frequency bands used during this investigation were relatively broad and, given some indication of frequency-dependency, assessment of narrower pink noise bands would be beneficial. Moreover, the results here indicate relative changes of VIS are perceivable for all frequency bands, but they do not reveal the extent of those changes. Considering this, the following experiments in Chapter 4 investigate the perception of decorrelated octave-band pink noise, both in a relative and absolute sense. The vertical decorrelation testing in the current chapter was also limited to the median plane i.e. with minimal interaural differences; therefore, the following investigations also consider vertical decorrelation for different azimuth positions around the listener, to observe the interaural effect on VIS perception.

4 RELATIVE AND ABSOLUTE GRADING OF VERTICAL IMAGE SPREAD FOR OCTAVE-BAND PINK NOISE STIMULI[3,4]

This chapter describes two investigations that have been conducted to observe the effect of vertical interchannel decorrelation on octave-band and broadband pink noise stimuli. The first experiment compares vertically decorrelated conditions against each other and a correlated reference, in terms of relative vertical image spread (VIS); whereas the second experiment looks at the absolute extent of VIS, assessing extreme stimuli from the first experiment. Chapter 3 demonstrated that changes to VIS from vertical decorrelation in the median plane (0° azimuth) were perceivable for all three frequency bands tested ('Low', 'Middle' and 'High'). Differences of trend were apparent between the results of the three bands, where the greatest degree of change was observed for the 'Low' and 'Middle' bands (Section 3.4). Given the suggested frequency-dependency of vertical decorrelation, the current experiments aim to investigate how narrower frequency bands (octave-bands) and broadband signals respond to vertical interchannel decorrelation.

Furthermore, since only the median plane (0° azimuth) was assessed in Chapter 3, the vertically decorrelated stimuli in the current chapter have been presented independently from three different azimuth angles (0°, ±30° and ±110°), to observe whether interaural differences can also contribute to the perception of VIS. It is still useful to assess the effects of vertical decorrelation in the median plane, however, the inclusion of the wider azimuth angles makes the assessment more representative of auditory presentation in commercial 3D surround sound systems, such as, Auro-3D 9.1 (Auro Technologies, 2015a). This provides a more detailed understanding of

[3] Gribben, C., & Lee, H. (2017). The Perceptual Effect of Vertical Interchannel Decorrelation on Vertical Image Spread at Different Azimuth Positions. Presented at the *142th Convention of the Audio Engineering Society*, 9747.
[4] Gribben, C., & Lee, H. (2018). The Frequency and Loudspeaker-Azimuth Dependencies of Vertical Interchannel Decorrelation on Vertical Image Spread. *Journal of the Audio Engineering Society,* (accepted May 2018).

vertical interchannel decorrelation for potential upmixing applications, as well as the control of auditory phantom sources around the head.

In order to observe the ICC effect at the different azimuth angles, three degrees of interchannel cross-correlation (ICC) have been precisely generated using two separate decorrelation approaches. Conventional decorrelation methods can broadly be split into phase-based and spectral-amplitude-based techniques, as discussed in Section 2.3 of Chapter 2. The first experiment in the current chapter uses one phased-based technique proposed by Kendall (1995) (all-pass filtering) and one amplitude-based technique by Lauridsen (1954) (complementary comb-filtering, which was also the method utilised in Chapter 3), both of which are described in Section 4.2.1 below. Comparing these two approaches side-by-side (with the same controlled levels of ICC) should better indicate whether a general relationship between vertical ICC and VIS exists, rather than simply assessing the effects for a single decorrelation method.

From this background, the following research questions are proposed:

- At which frequencies is the relationship between vertical ICC and VIS strongest?
- What effect does the azimuth angle of presentation have on VIS perception?
- Are phase- and amplitude-based decorrelation methods perceived similarly?
- What is the extent of change to VIS by vertical decorrelation?
- Is the general extent of VIS dependent on frequency and/or azimuth angle?

In order to answer the above questions, broadband pink noise was band-pass filtered into nine octave-bands (with centre frequencies of 63 Hz to 16 kHz). These octave-bands and the broadband pink noise were then decorrelated with the two decorrelation methods mentioned above, generating three interchannel cross-correlation coefficients (ICCCs) of '0.7', '0.4' and '0.1' – monophonic (lower main-layer loudspeaker only) and correlated (ICCC = 1.0) conditions were also included in a multiple comparison alongside the decorrelated stimuli. In the second experiment, subjects graded the absolute extent of VIS for extreme stimuli from the first experiment (ICCC = 0.1, ICCC = 1.0 and the monophonic condition) at azimuth angles of 0° and ±30°.

Responses were captured with the aid of a light emitting diode (LED) strip connected to Max and a rotary controller, where the listeners controlled the LEDs to visually define the upper and lower VIS boundaries of the stimuli for each frequency band.

4.1 Experimental Hypotheses

The results in Chapter 3 demonstrate that some significant change of VIS from vertical inter-channel decorrelation could be perceived for all three of the frequency bands ('Low', 'Middle' and 'High'). Consequently, it is hypothesised that spatial changes from vertical decorrelation will be perceivable for all nine octave-bands to some degree. Given that vertical localisation in the median plane is mostly determined by pinna spectral cues around 8 kHz (Roffler & Butler, 1968a; Hebrank & Wright, 1974), and since potential spectral cues of vertical decorrelation were observed in a similar region in Chapter 3 (Section 3.5.2), it is anticipated that the 8 kHz octave-band may display the greatest degree of change by vertical decorrelation. In terms of azimuth angle, it is seen in the literature that interaural level differences (ILD) have more of an impact on higher frequencies at wider azimuth angles (Section 1.1.1.1); therefore, it is further hypothesised that the vertical decorrelation from ±110° will be aided by the additional head-shadowing of high frequencies (> 1.5 kHz).

The discussion in Section 3.5.2 suggests that the perception of the 'High' frequency band (octave-bands with centre frequencies of 4 kHz, 8 kHz and 16 kHz) in the median plane may have been influenced by the 'pitch-height' effect (Cabrera & Tilley, 2003; Lee, 2016b). That is, higher frequencies are inherently spread along the vertical plane, with the 8 kHz octave-band perceived as most elevated towards the height-channel loudspeaker. As a result, an additional hypothesis is that the absolute testing results will demonstrate different frequency bands at different heights in the median plane, specifically an elevation of the 8 kHz octave-band. It is further hypothesised that the lower octave-bands will generally display a greater VIS than higher octave-bands, in line with the literature on the inherent frequency-dependent extent of sound in Section 1.2.1 (Cabrera & Tilley, 2003).

4.2 Experiment 1: Relative Grading of Vertical Image Spread (VIS)

The first experiment of the present chapter looks to establish the relative perception of vertical image spread (VIS) between vertically decorrelated and correlated pink noise stimuli. Two decorrelation methods (Sections 4.2.1.3 and 4.2.1.4 below) were employed to control the degree of signal correlation (ICC) between pairs of vertically-arranged loudspeakers. Stimuli were created using octave-band and broadband pink noise, in order to observe the frequency dependency of ICC on VIS; and were presented at three different azimuth angles to the listening position (0°, ±30° and ±110°). The results from this first experiment then informed the second experiment of the current chapter, where absolute VIS was measured to examine changes of the upper and lower image boundaries.

4.2.1 Experimental Design

4.2.1.1 Physical Testing Setup

Ten Genelec 8040A loudspeakers (Frequency response: 48 Hz – 20 kHz (±2 dB)) were used during testing, divided into two separate layers (main and height) with five loudspeakers in each. The five main-layer loudspeakers were spaced around the listener at azimuth angles of 0°, ±30° and ±110°, in accordance with ITU-R BS.775-3 (ITU-R, 2012) – each loudspeaker was positioned 2 m from the listening position, with the acoustic centre at a height of 1.27 m and in line with the ear position. The height-layer loudspeakers were positioned directly above the main-layer azimuth positions at an elevation angle of +30° to the listener, as per Auro-3D 9.1 (with an additional centre height) (Auro Technologies, 2015a). This resulted in five vertically-arranged loudspeaker pairs around the listener, with a vertical spacing of 1.15 m between layers (see Figure 4.1). Auro-3D 9.1 is a well-established 3D multichannel surround sound format, and was chosen for this experiment due to its relationship with the standard '2D' 5.1 format (ITU-R, 2012). That is, it features four vertically-arranged pairs of loudspeakers at azimuth angles of ±30° and ±110°, where the four main- and height-layer loudspeakers align at the exact same position (i.e. left, right, left surround and right surround in 5.1 surround sound).

A median vertical pair ('Centre' at 0° azimuth, directly in front of the listener) was also included to observe VIS perception where interaural differences are limited. The objective analysis in Chapter 3 (Section 3.5.2) suggests that the perception of high frequency VIS from vertical decorrelation in the median plane may be related to spectral cues – therefore, the inclusion of the median plane condition also provides continuity from that experiment, in order to explore these potential spectral cues further.

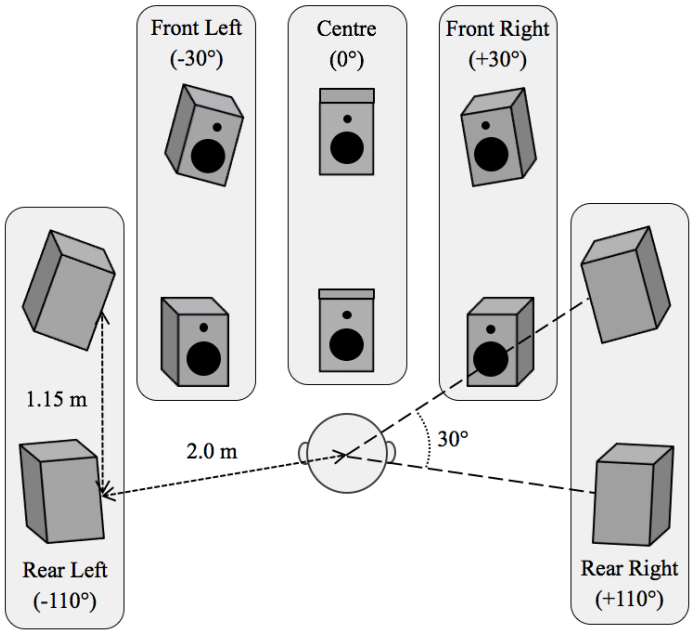

Figure 4.1 Physical loudspeaker setup used during testing (based on Auro-3D 9.1 (Auro Technologies, 2015a) with an additional Centre height-channel). Five main-layer loudspeakers positioned 2 m from the listener at ear height with azimuth angles of 0°, ±30° and ±110°. Five upper height-layer loudspeakers elevated directly above its main-layer pair by +30° to the listener.

Listening tests were conducted at the University of Huddersfield in a critical listening room that fulfils the specification of ITU-R BS.1116-3 (ITU-R, 2015a) (6.2m x 5.6m x 3.8m; RT = 0.25 s; NR 12). Time and level alignment were applied between the two loudspeaker layers, to compensate for interlayer difference of signal arrival at the listening position. An acoustically transparent curtain was also used to obscure the loudspeakers from view, so as to avoid visual bias during testing.

4.2.1.2 Decorrelation Methods

For stimuli creation, continuous broadband pink noise was band-pass filtered into nine octave-bands, with centre frequencies of 63 Hz to 16 kHz, using a 16th-order linear phase Butterworth filter (96 dB/octave). To generate the stimuli, each pink noise octave-band and the original broadband sample were processed using the two decorrelation techniques: Complementary Comb-Filtering (CF) (the amplitude-based method implemented in Chapter 3) and Phase Randomisation (PR) by use of all-pass filters (a phase-based approach).

With regard to phase-based decorrelation, various suggestions have been made to improve these approaches in terms of colouration and transient handling – these include the use of exponentially decaying white noise bursts (Faller, 2006; Pulkki, 2007), the random time-shifting of whole critical frequency bands (Bouéri & Kyirakakis, 2004; Pihlajamäki, Santala & Pulkki, 2014) and extraction of transient information (Laitinen et al., 2011) – however, the added complexity can make achieving low levels of ICC more difficult. The first experiment in the present chapter has a focus on controlling a broad range of $ICCC_{avg}$ values, using continuous pink noise as the stimulus (i.e., limited transients and little consideration towards colouration). Given this, it was deemed that the PR all-pass filter approach (Kendall, 1995) would be best suited for that purpose, as it is relatively straightforward to implement and able to achieve low levels of ICC across all frequencies. Some existing surround sound upmixing models also indicate the generic use of all-pass filters to decorrelate signals by phase randomisation (Jot & Avendano, 2003; Li & Driessen, 2005).

Considering amplitude-based methods, these mostly feature frequency panning where groups of frequencies are alternately panned between the two output channels across the frequency spectrum (Section 2.3.3). The CF method in the present experiment generates fixed and regular frequency panning between two channels by creating two opposing comb-filter responses. This is also the method utilised in the experiments of Chapter 3 in the current thesis, and full details of its implementation can be seen in Section 3.2.1 – an FFT example of the resulting

complementary amplitude differences is also presented in the same section (Figure 3.1), displaying the groups of panned frequencies for a 500 Hz octave-band. The CF method is very simple and cost-effective to implement, and the degree of ICC is easily controlled by the gain factor applied to the delayed signal (where a gain factor of 1.0 is maximum decorrelation). Many other amplitude-based methods also work on a similar principle of frequency distribution between two channels (Zotter & Frank, 2013; Fink et al., 2015); however, suitably low levels of ICC can easily be realised with CF, making it an appropriate technique to assess in the present experiment. Furthermore, the method is also featured in some proposed surround sound upmixing algorithms (Irwan & Aarts, 2002; Adami, Brand & Herre, 2017).

Although more sophisticated methods of decorrelation have been developed in recent years to improve tonal quality (as discussed in Section 2.3), the two approaches selected for the experiments in the current chapter (PR and CF) allow for a simple controlled assessment of the ICC effect, due to fewer parametric variables. Moreover, informal calculations during the stimuli creation process demonstrated that both were easily able to achieve an $ICCC_{avg}$ of at least '0.1', for each of the octave-bands under testing (63 Hz – 16 kHz) – this is considered a minimum requirement of the decorrelation methods used during these experiments.

4.2.1.3 Phase Randomisation Stimuli (All-Pass Filtering)

To implement the PR method, an original monophonic signal can simply be convolved with two impulses (short white noise bursts) of random phase and unit magnitude (Equation 4.1). These impulses represent an FIR all-pass filter, with the intention that the amplitude frequency responses between the input and output signals are identical.

$$\begin{aligned} s_1(n) &= x(n) * h_1(n) \\ s_2(n) &= x(n) * h_2(n) \end{aligned} \tag{4.1}$$

where s_1 and s_2 are the two output signals, x is the monophonic input signal, and h_1 and h_2 are two FIR filter impulses.

To create the random phase coefficients of the FIR all-pass filters, a random number sequence is generated for each filter (featuring random values between $-\pi$ and π). This produces an inherent phase decorrelation between the two filter responses, where the degree of correlation can be controlled by a mixing matrix of the two sequences, as seen in Equation 4.2 below. The random numbers represent the phase component of each FFT bin in the frequency-domain, where the magnitude of the frequency is set to unity (1). To obtain the impulse response of each filter for time-based convolution with the input signal, an inverse FFT (IFFT) is performed to convert the FFT frequency components into the time-domain. Alternatively, convolution could be carried out in the frequency domain by performing an FFT of the input signal, before converting back to the time-domain with an IFFT, which may be more computationally efficient.

$$h_1[n] = a[n]$$
$$h_2[n] = \frac{1}{1+k}(a[n] + (b[n] \cdot k)) \qquad (4.2)$$

Where: h_1 and h_2 are the two random filter coefficients after mixing
a is the first random number sequence
b is the second random number sequence
k is the mixing factor between 0 and 1, where 1 is maximum decorrelation

The length of the random number sequences defines the resolution of the frequencies that are phase-randomised, as well as the temporal length of the filter's impulse response. To create the PR stimuli in the current experiment the filters were generated with a length of 30 ms, resulting in 1323 FFT bins at a sampling rate of 44.1 kHz. A length of 30 ms for the sequences was found to easily decorrelate the 63 Hz octave-band to $ICCC_{avg}$ 0.1, where the lowest frequency (44 Hz) has a cycle length of around 23 ms. It is thought that the correlation between the two generated impulses directly relates to the correlation between the two outputs; however, given the random generation of number sequences, the actual degree of maximum ICC can vary drastically between each generation, often requiring repetition until the desired ICC level is achieved.

It is assumed that the use of all-pass decorrelation should have little effect on the frequency response of the signals, though in practice, the length of the filter (white noise burst) can cause smearing of transient information (Laitinen et al., 2011). In the current experiment, this issue is not of particular concern as the subjective assessment has been conducted with continuous pink noise sources. In general, the longer the filter length, the greater or easier the decorrelation, yet at the expense of increased signal colouration. The waveform of the output can also be distorted by significant opposing phase-shifts of neighbouring frequencies (Bouéri & Kyirakakis, 2004) – it is this random interaction of phase-shifts that leads one implementation of the all-pass filter to sound noticeably different from another. Although it is relatively simple to achieve a low level of ICC with all-pass filters, the random element requires a trial-and-error approach in terms of both colouration and ICC, which may not be suitable for practical applications. In this instance, MATLAB was used for the PR decorrelation processing – to achieve the $ICCC_{avg}$ values required for the present experiment ('0.1', '0.4' and '0.7'), a MATLAB script was used to loop the decorrelation code until the correct $ICCC_{avg}$ value was calculated.

4.2.1.4 Complementary Comb-Filtering Stimuli

The two parameters of the CF method are the time-delay and the gain factor, both of which are applied to the summed / subtracted secondary signal (see Section 3.2.1). In Chapter 3, the CF method was assessed with four different time-delays (1 ms, 5 ms, 10 ms and 20 ms) and varying degrees of gain factor ('0.0' to '1.0', with increments of 0.2, where '1.0' is maximum decorrelation). It was generally found that longer time-delays (10-20 ms) were required to significantly change VIS for all three frequency bands. With the 'Middle' band, both 10 ms and 20 ms time-delays generated a significant increase of VIS between conditions; for the 'High' band, only the 10 ms time-delay significantly increased VIS; and with the 'Low' frequency band, both the 1 ms and 20 ms time-delays demonstrated a significant increase of VIS. In Chapter 3 (Section 3.5.1), it was discussed that a 1 ms time-delay would be unsuitable for decorrelation of low frequencies in practice, due to the uneven distribution of frequencies within that region (Figure

3.8) (which may have been perceived as an increase of VIS) – as a result, a time-delay of 1 ms has not be been considered for the current experimental stimuli.

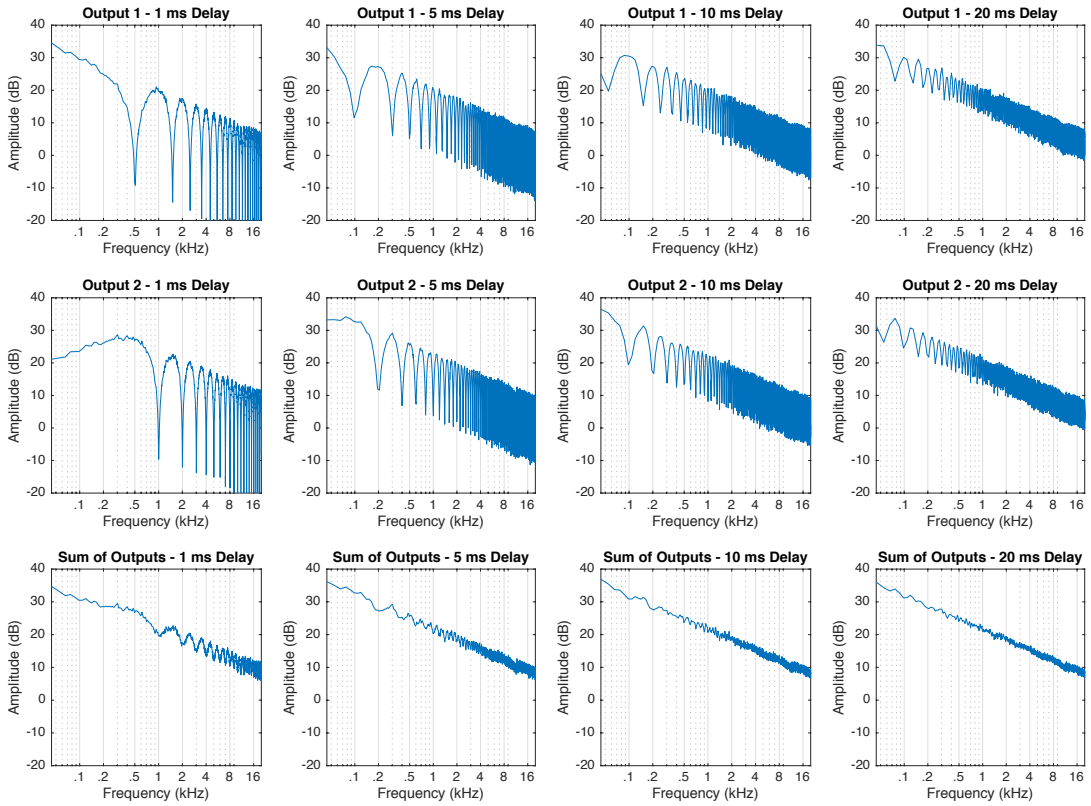

Figure 4.2 The effect of time-delay (T) on the comb-filter notch depth and bandwidth between notches with the CF method, calculated with 4096 FFT-points and a frame length of 4096 samples (50% overlapping windows). (Broadband pink noise decorrelated with 1-20 ms delays and a gain factor of '1.0')

In order to determine a suitable time-delay, the effects of 1 ms, 5 ms, 10 ms and 20 ms time-delays have been analysed further, with the comb-filtered results displayed in Figure 4.2 above (where the time-delays are split by column). The first two rows of plots represent the two output signals with a gain factor of '1.0' (maximum decorrelation) – in these, the notch depth becomes more apparent for a time-delay of 10 ms or shorter (> ~12 dB), with the notch increasing considerably as time-delay decreases. Calculating the $ICCC_{avg}$ for the 63 Hz octave-band, the minimum achievable $ICCC_{avg}$ with a gain factor of '1.0' (maximum decorrelation) is 0.0 for 1 ms, 0.1 for 5 ms, 0.1 for 10 ms and 0.2 for 20 ms. Given that an aim of this study is to test as close to full decorrelation as possible without separation of image ($ICCC_{avg}$ = 0.1), a time-delay of

20 ms does not give a suitable range of $ICCC_{avg}$ values at lower frequencies, presumably due to the shallower notch depth.

Further seen in Figure 4.2, the frequency bandwidth of the individual comb-filter peaks (taken as the distance between notch positions) are consistent across the frequency spectrum; from observing the plots, this bandwidth is 1 kHz for a 1 ms time-delay, 200 Hz for 5 ms, 100 Hz for 10 ms and 50 Hz for 20 ms. Due to these variations in bandwidth, an uneven distribution of low frequencies is particularly evident in the 1 ms plots – this is the effect that was observed in Chapter 3 of the present thesis (Section 3.5.1). Figure 4.2 also indicates a comb-filter distortion when summing the outputs of shorter time-delays (as seen in the third row of plots); since both outputs have been RMS-matched, this spectral distortion may be caused by a combination of the spectral energy imbalance and also the greater notch depths seen with shorter time-delays. On the other hand, looking at the 10 ms and 20 ms plots, the comb-filter distortion of the summed outputs appears to reduce and the frequency distribution is more evenly spread throughout the spectrum.

Given the above assessment, the present study chooses to utilise a time-delay of 10 ms, due to a suitable maximum notch depth (~12 dB with an $ICCC_{avg}$ of 0.1 for 63 Hz octave-band) and a relatively small comb-filter peak bandwidth between notches (100 Hz). Furthermore, the 10 ms condition in the vertical decorrelation experiment of Chapter 3 (Section 3.4) demonstrated a significant change of VIS for both the 'Middle' and 'High' frequency bands. Other researchers have also suggested that a 10 ms time-delay is a compromise to give the desired perception of widening, while avoiding any confusion that may be experienced with longer time-delays and greater diffusion (Irwan & Aarts, 2002).

4.2.1.5 Stimuli Conditions

For each technique described above, three $ICCC_{avg}$ levels of decorrelation were generated: '0.1', '0.4' and '0.7'. Additionally, $ICCC_{avg}$ '1.0' (fully correlated) and a monophonic signal routed to the main-layer loudspeaker only were also included as stimuli conditions. This

resulted in eight stimuli for each of the ten frequency conditions (nine octave-band and one broadband). Both methods were implemented using MATLAB looping scripts that repeated the decorrelation process until the desired $ICCC_{avg}$ was achieved. In the case of CF, the gain factor was incremented with each loop, and for PR, new random number sequences of phase were generated each time (with the mixing matrix set to a reasonable level). Each of the two decorrelated output signals were RMS level-matched with the input signal to maintain an equal balance of energy between the channels – this was due to the uneven frequency distribution observed for low frequencies in Section 3.5.1. Of the two-channel decorrelated outputs, 'Output 1' was routed to the main-layer loudspeaker and 'Output 2' to the height-layer loudspeaker of a given vertical pair.

The level of each frequency band stimulus was determined by octave-band filtering a broadband pink noise signal at 75 dB LAeq, with each of the eight stimuli matched to the respective sound pressure level (LAeq) of the resultant bands – the SPLs for each frequency band are displayed in Table 4.1 below. Rather than matching all octave-bands to the same SPL, it was considered that maintaining the original inter-band energy relationship of pink noise (equal energy per octave-band) would be more appropriate. The motivation here was to examine the effectiveness of decorrelation for each octave-band, whilst maintaining the band's inherent loudness within a broadband signal. This would also be more representative of a potential practical application, where some octave-bands are selectively decorrelated for vertical upmixing, without changing the spectral energy weighting of the original signal.

Table 4.1 Sound pressure level (SPL) (LAeq) of the octave-band pink noise stimuli, as calculated from the octave-band filtering of a broadband pink noise signal with an SPL of 75 dB (LAeq).

63 Hz	**125 Hz**	**250 Hz**	**500 Hz**	**1 kHz**	**2 kHz**	**4 kHz**	**8 kHz**	**16 kHz**	**Broad-band**
49 dBA	51 dBA	63 dBA	66 dBA	68 dBA	68 dBA	67 dBA	64 dBA	53 dBA	75 dBA

4.2.1.6 Subjects

A total of 12 subjects took part in the first experiment of this study, comprising staff, postgraduate students and final year students from the University of Huddersfield's Music Technology courses. All participants reported to have normal hearing and were familiar with critical listening exercises in spatial audio. The use of 12 subjects achieves a statistical power of 0.47 for the current experiment, based on a two-tailed t-test with an effect size of 0.6 and α error probability of 0.05 (type I error i.e. false-positive), as calculated using G*Power 3.1 (Faul et al., 2007). This indicates that the probability of a type II statistical error (false-negative) is 0.53 (β), suggesting there is a fairly high chance that some significant (perceivable) differences may be reported as insignificant. If a type II error were to occur, it is thought that the differences between these conditions are likely to be particularly subtle, and it is better to be more restrictive than sensitive when testing for significance within the current context. As discussed in Section 3.4.2, a type I error is of more concern for the present subjective listening experiments, where a difference is reported as significant when it is insignificant – this is accounted for with Bonferroni correction of the results below, reducing the sensitivity of the statistical test.

4.2.1.7 Testing Procedure

During the test, subjects were presented with a total of 30 multiple comparison trials in a graphical user interface (GUI) that was developed in Cycling '74's Max (Figure 4.3 below) – this interface provided the basis for HULTI-GEN, a universal listening test interface generator that has since been developed by the author (Gribben & Lee, 2015) (Appendix A). For the reasons detailed in Section 3.2.4, trials were based on an adapted MUSHRA format (ITU-R, 2015b) with a bipolar scale (ranging from -30 to 30), featuring markers every 10 points to help maintain consistency between trials. Each trial consisted of eight buttons and sliders to trigger and grade the eight stimuli, with another button next to the centre of the scale (0) to trigger a reference stimulus. Six of the buttons/sliders corresponded to the decorrelated stimuli, with the other two

controlling the correlated and monophonic stimuli. The reference signal was the correlated sample ($ICCC_{avg}$ 1.0), which was also included as a hidden reference amongst the stimuli.

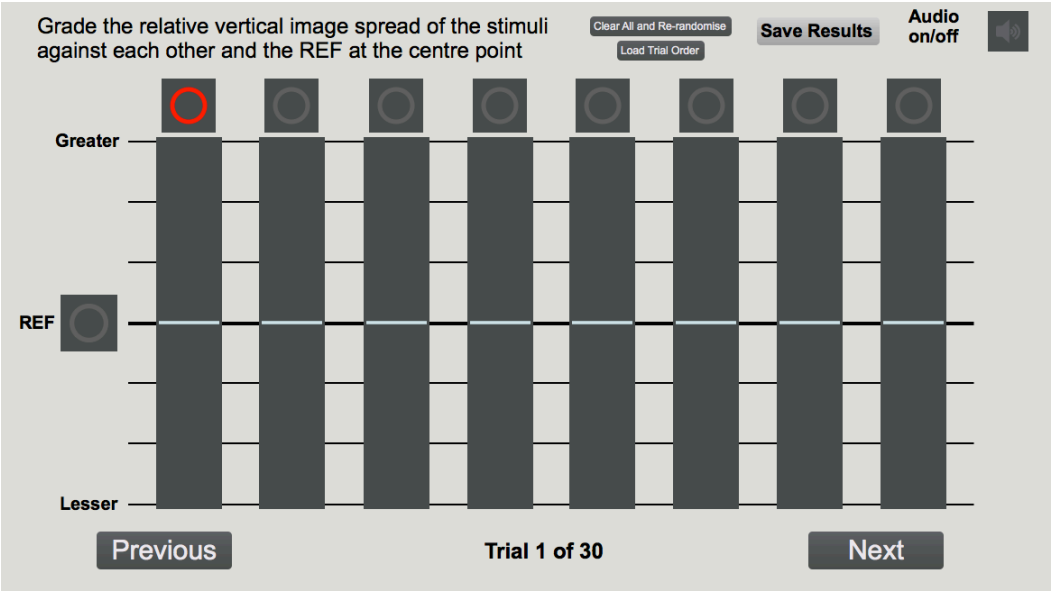

Figure 4.3 Multiple comparison interface used during testing, developed in Max 7.

Each stimulus was to be graded for vertical image spread (VIS) against each other and the reference – where above the reference (at 0 on the scale) was greater and below the reference was lesser. The 30 trials were made up of the ten frequency bands for three loudspeaker azimuth angle positions (0°, ±30° and ±110°, see Figure 4.1 above). Only one of the two 30° and 110° vertical loudspeaker pairs were tested for each band to limit the testing load – this was randomised between left and right for each 30°/110° trial. With every subject, the 30 trials were randomised in order and divided into three separate sessions, each of which took no more than 30 minutes for a listener to complete. Subjects were instructed to face forward and keep their head still throughout testing, which was ensured by the aid of a small headrest.

4.2.2 Results and Analysis

Results for relative vertical image spread (VIS) testing are presented in Figure 4.4 below – all data has been normalised in accordance with ITU-R BS.1116-3 (ITU-R, 2015a) (as described in Section 3.3) and analysed in SPSS. The graphs display the median scores of relative VIS with bars to signify notch edges, representing non-parametric 95% confidence intervals (McGill et al., 1978). Shapiro-Wilk tests for normality indicated that the data of each condition was not always normally distributed; therefore, non-parametric statistical tests were performed across all conditions for consistency and comparison.

Data was first analysed for significant difference between the two decorrelation methods, using Bonferroni-corrected Wilcoxon signed-rank tests to compare the two methods at each of the three interchannel cross-correlation coefficient ($ICCC_{avg}$) levels. Friedman repeated measure tests were then conducted to assess the effect of ICC on VIS for either the methods combined, if no difference was established, or each method independently. If a significant ICC effect was revealed, post-hoc Wilcoxon tests with Bonferroni correction indicated the significance of relationships between stimuli conditions. The monophonic stimulus was not included in the Friedman statistical testing, as it does not feature an $ICCC_{avg}$ value. Instead, additional Wilcoxon corrected tests were carried out between the monophonic stimulus and all other stimuli to observe significant differences, with the median and notch data of the monophonic stimuli also presented in the graphs for comparative purposes.

4.2.2.1 Comparing Decorrelation Methods

For the 63 Hz, 500 Hz, 4 kHz, 8 kHz and 16 kHz octave-bands at all azimuth angles, there was no significant difference of relative VIS between the phase randomisation (PR) and complementary comb-filtering (CF) methods, when performing a Wilcoxon test on each of the three $ICCC_{avg}$ pairs ($p > 0.05$). The 125 Hz, 1 kHz, 2 kHz and Broadband frequency bands also showed no significant difference between methods for 0° and ±30° azimuth ($p > 0.05$). However, with the 125 Hz, 2 kHz and Broadband frequency bands at ±110°, 'PR 0.1' was

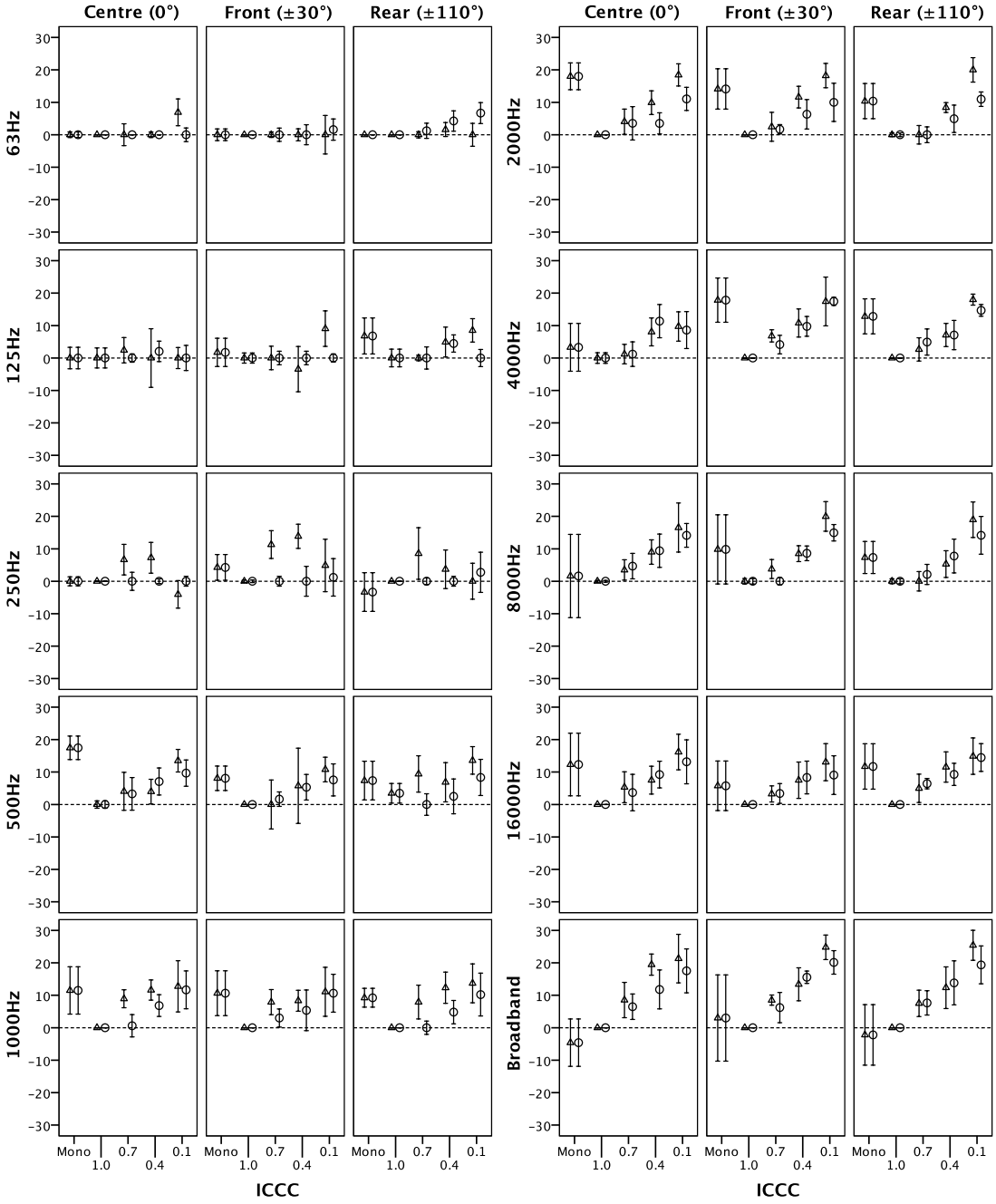

Figure 4.4 Median of the relative Vertical Image Spread (VIS) normalised scores with 95% confidence notch edge bars. Each panel features the results of all average running interchannel cross-correlation coefficient ($ICCC_{avg}$) and monophonic conditions for a given combination of loudspeaker position and frequency band. Results of the two decorrelation methods are clustered around the $ICCC_{avg}$/monophonic points on the X-axis:

Triangle / Left = Phase Randomisation (PR)
Circle / Right = Complementary Comb-Filtering (CF)

significantly greater than 'CF 0.l' ($p < 0.01$); and for the 1 kHz band at ±110°, 'PR 0.4' was significantly greater than 'CF 0.4' ($p = 0.02$). For the 250Hz band at 0°, there was no significant difference between methods ($p > 0.05$) – whereas PR was significantly greater than CF for '0.7' and '0.4' at ±30° ($p < 0.02$), and for '0.7' at ±110° ($p < 0.04$). In every case where there was a significant difference between methods, PR always had a greater perceived VIS than CF.

4.2.2.2 Interchannel Cross-Correlation Effect

For the 63 Hz octave-band, Friedman test results indicate a significant ICC effect for 0° and ±110° azimuth ($p < 0.05$), but not for ±30° ($p > 0.05$). Wilcoxon post-hoc tests with Bonferroni correction revealed no significant difference between conditions for 0°, however, '0.7', '0.4' and '0.1' were all significantly greater than '1.0' at ±110° azimuth ($p < 0.05$). With the 125 Hz band, only the PR method at ±110° azimuth showed a significant ICC effect, where '0.1' was significantly greater than '0.7' ($p < 0.02$). The 250 Hz band results indicate a significant ICC effect for PR at ±30° ($p < 0.01$) and both methods at ±110° ($p < 0.02$). For the PR method at ±30°, '0.7' and '0.4' were significantly greater than '1.0' ($p < 0.02$); and for PR at ±110°, '0.7' was significantly greater than '1.0' ($p < 0.05$). However, post-hoc tests of the CF method at ±110° reveal no significant differences following correction ($p > 0.05$).

Friedman tests on the 500 Hz, 1 kHz, 2 kHz, 4 kHz, 8 kHz, 16 kHz and Broadband frequency bands demonstrate a significant ICC effect at all azimuth angles (0°, ±30° and ±110°) ($p < 0.01$). For the 500 Hz at 0°, '0.1' was significantly greater than all other ICC conditions ($p < 0.04$); whereas at ±30° and ±110°, '0.1' was only significantly greater than '1.0' ($p = 0.01$). For the 1 kHz band at 0° and ±30°, '0.1' and '0.4' were both significantly greater than '1.0' ($p < 0.01$), but not each other ($p > 0.05$) – and at ±110°, 'PR 0.4' was significantly greater than 'PR 1.0' ($p = 0.02$), but there was no significant difference between the CF stimuli ($p > 0.05$). With the 2kHz band, '0.1' was significantly greater than all other ICC conditions for 0°, ±30° and PR at ±110° ($p < 0.01$); whereas for CF at ±110°, '0.1' was only significantly greater than '1.0' ($p < 0.04$). Results of the 4 kHz band at 0° indicate '0.1' and '0.4' were significantly

greater than '0.7' and '1.0' ($p < 0.02$), but not each other ($p > 0.05$) – whereas at ±30° and ±110°, '0.1' was significantly greater than all other ICC conditions ($p < 0.03$). For the 8 kHz band, '0.1' was significantly greater than all other ICC levels at each azimuth angle ($p < 0.04$). With the 16 kHz octave-band, '0.1' and '0.4' were significantly greater than '0.7' and '1.0' at all angles ($p < 0.04$); furthermore, '0.1' was also significantly greater than '0.4' at ±110° ($p < 0.01$). For the Broadband 0° results, '0.1' and '0.4' were both significantly greater than all other ICC levels ($p < 0.02$) – whereas at ±30° and ±110° azimuth, '0.1' was significantly greater than all other conditions ($p < 0.03$).

4.2.2.3 Statistical Correlation Between ICC and VIS

To further evaluate the apparent effect of ICC on VIS perception, the data sets of each decorrelation method were combined and the statistical correlation between ICC and VIS was calculated for each condition (Table 4.2). The correlation coefficients were determined using both Spearman's rank-order and Pearson's product-moment measurement techniques. Spearman's rank-order test is non-parametric and can be applied to ordinal data, with the results indicating the strength of a monotonic relationship between variables – that is, as one variable increases, so does the other – whereas Pearson's approach assesses the linearity of correlation between two variables. Interpretation for both sets of correlation coefficients is as follows: 0.3-0.49 = weak, 0.5-0.69 = moderate, 0.7-0.89 = strong, and 0.9-1.0 = very strong. A strong correlation between ICC and VIS for a particular octave-band would suggest that measurement of the $ICCC_{avg}$ for that band might contribute to the prediction of VIS (or potentially another spatial attribute), and also that it may be particularly important to controlling VIS.

Observing the Spearman rank-order coefficients in Table 4.2, it can be seen that the strongest correlation between ICC and VIS is for the 8 kHz band at ±30° azimuth ($r_s = 0.83$); this is further reflected in the Pearson results ($r = 0.80$), indicating that the relationship is relatively linear. The Spearman results also show strong correlation for the Broadband stimuli at ±30°, and 2 kHz, 4 kHz, 8 kHz, 16 kHz and Broadband at ±110°. There is general agreement between

the two correlation results, demonstrating that the strong correlation between variables is mostly linear. On the other hand, at 0° azimuth in the median plane, none of the frequency bands exhibit a strong correlation between ICC and VIS.

Table 4.2 Statistical Correlation Between the Interchannel Cross-Correlation Coefficient (ICCC) and the relative Vertical Image Spread (VIS) Scores (** $p < 0.01$; * $p < 0.05$)

	Spearman's Rank-Order (r_s)			**Pearson's Product-Moment (r)**		
	Centre (0°)	**Front (±30°)**	**Rear (±110°)**	**Centre (0°)**	**Front (±30°)**	**Rear (±110°)**
63 Hz	0.29**	0.11	0.35**	0.24*	0.17	0.30**
125 Hz	0.01	0.05	0.30**	0.03	0.07	0.29**
250 Hz	-0.09	0.01	0.14	-0.10	0.02	0.14
500 Hz	0.49**	0.41**	0.33**	0.47**	0.36**	0.33**
1 kHz	0.52**	0.42**	0.42**	0.54**	0.42**	0.41**
2 kHz	0.63**	0.68**	0.74**	0.63**	0.66**	0.68**
4 kHz	0.51**	0.67**	0.75**	0.50**	0.68**	0.74**
8 kHz	0.65**	0.83**	0.73**	0.58**	0.80**	0.70**
16 kHz	0.52**	0.56**	0.73**	0.50**	0.55**	0.70**
Broadband	0.62**	0.77**	0.71**	0.54**	0.73**	0.69**

Significant correlation ($p < 0.01$) between ICC and VIS occurs at all azimuth angles for the 500 Hz octave-band and above, and also for 63 Hz at 0°/±110° and 125 Hz at ±110°, broadly agreeing with the significant results in Figure 4.4. With the 63 Hz, 125 Hz and 500 Hz significant results, the correlation is considered to be weak ($r < 0.5$) – as is the 1 kHz band at ±30° and ±110° azimuth, whereas from 0°, the 1 kHz band shows a moderately linear correlation. For both the 500 Hz and 1 kHz bands, correlation between ICC and VIS is greatest at 0° azimuth, then appears to decrease as the azimuth angle increases – this suggests that the perception of VIS by decorrelation for these frequencies may be most detectable when changes occur in both ears equally (i.e. the median plane), potentially related to the interaural perception of decorrelated room reflections. In contrast, the 2 kHz, 4 kHz and 16 kHz bands show strongest correlation at ±110°, where interaural differences are greatest from head-shadowing – and the 8 kHz and Broadband frequency bands indicate strongest correlation at ±30°, where there is a combination of both interaural differences and spectral filtering by pinna reflections.

4.2.2.4 Monophonic Results

The only decorrelated stimuli to have a significantly greater VIS than the respective monophonic conditions were 63 Hz, 2 kHz and 8 kHz '0.1' at ±110° ($p < 0.04$), and Broadband '0.1' at both 0° and ±110° ($p < 0.03$). In some cases, the monophonic condition was perceived as having a significantly greater VIS than decorrelated stimuli. For the 500 Hz band, Wilcoxon tests revealed that the monophonic sample was significantly greater than '1.0' at both 0° and ±30° ($p < 0.03$), as well as greater than '0.7' and '0.4' for the PR method at 0° ($p < 0.04$). With the 1kHz band, the monophonic sample was significantly greater than '1.0' at ±110° ($p < 0.04$). The 2 kHz monophonic stimulus was significantly greater than '1.0' and '0.7' (PR method) at 0° ($p < 0.03$), as well as '1.0' and '0.4/0.7' (CF method) at ±30° ($p < 0.04$). With the 4 kHz band, the monophonic sample was significantly greater than '1.0' and '0.7' at ±30° ($p < 0.05$), and greater than '1.0' at ±110° ($p < 0.04$). It can be seen from these results that an $ICCC_{avg}$ of '0.1' was always perceived as having a similar or greater VIS than the monophonic condition for every frequency band and azimuth angle.

4.2.3 Discussion of Relative Testing Results

Generally speaking, there is little linear interchannel cross-correlation (ICC) effect with octave-bands below the 500 Hz band – that is, where an increase of spread is observed as correlation decreases. Looking at the median and notch edge plots in Figure 4.4 above, it can be seen that a direct relationship between ICC and vertical image spread (VIS) begins to develop around the 500 Hz octave-band point – this is confirmed by the significant statistical correlation results seen in Table 4.2. In every case for the 500 Hz octave-band and above, either the ICC conditions of '0.1' or '0.4' (decorrelated) were significantly greater than '1.0' (the correlated condition) across all loudspeaker angles. This perceptual effect is likely due to the introduction of vertical localisation cues generated by the pinna (Roffler & Butler, 1968a; Hebrank & Wright, 1974), torso (Algazi et al., 2001) and room reflections, as well as that of interaural cues by a head-shadowing effect when the source is presented off-centre. It also relates back to the potential

spectral cues of the median plane observed in Chapter 3 of the thesis. Consequently, the presence of these features may allow the brain to interpret two largely uncorrelated signals from different directions simultaneously. There also appears to be some direction dependency on the relationship between ICC and VIS, with the statistical correlation strongest for the 500 Hz and 1 kHz bands at 0° (the median plane), 2 kHz, 4 kHz and 16 kHz at ±110° (when head-shadowing is greatest), and 8 kHz and Broadband at ±30° (where frontal HRTF filtering and interaural differences are both present). These points have been explored further with the objective analysis in Chapter 5 of the present thesis (Section 5.3).

Comparing the results for the two decorrelation methods, phase randomisation (PR) and complementary comb-filtering (CF), there is generally little difference between them – 92% of all method comparisons exhibited no significant difference. In the cases where there was significant difference, the PR method was always perceived as having a greater VIS than CF. Given the general similarity of VIS between methods for each ICC level, and also the linear relationships seen at the 500 Hz octave-band and above, it can be suggested that the ICC relationship with VIS at middle to high frequency bands could be useful. Through further investigation and development, it may be found that vertical $ICCC_{avg}$ measurements at these frequencies can contribute to spatial attribute prediction (e.g. VIS or listener envelopment (LEV)).

It is worth noting that, for the majority of monophonic results, the monophonic stimuli were not perceived as significantly different from decorrelated conditions, when all conditions were SPL level-matched – the larger notch bars for the monophonic stimuli in Figure 4.4 suggest that a disagreement or confusion amongst listeners may have been the cause of this. In some cases between 500 Hz and 4 kHz, the monophonic sample was even perceived as having a significantly greater VIS than vertically decorrelated stimuli. One cause could be the impact of ICC on perceived loudness, where the interaction of two partially correlated signals at the ear may result in the cancelling of frequencies (e.g. by comb-filtering), leading to a perceived 'weakening' of the signal. There could have also been a room effect with strong early reflections influencing the perception of the monophonic stimuli, particularly for the middle

frequencies around 500 Hz to 1 kHz – this has been objectively investigated and discussed in Section 5.5 of Chapter 5. Furthermore, the 'directional bands' phenomenon may have influenced the grading of VIS for some of the monophonic octave-band stimuli (Blauert, 1969/70); in particular, the 1 kHz and 8 kHz bands tend to be perceived behind and above respectively when presented monophonically, which may have caused confusion of VIS perception and resulted in a greater spread of responses. Given the large spread of responses and increase of perceived VIS for some monophonic stimuli, the only frequency bands to feature stimuli with a significantly greater VIS than the monophonic condition were Broadband at 0° azimuth, and 63 Hz, 2 kHz, 8 kHz and Broadband at ±110°.

4.3 Experiment 2: Absolute Grading of Vertical Image Spread (VIS)

From the first experiment of the current chapter, it was established that the relative effect of interchannel cross-correlation (ICC) on vertical image spread (VIS) is both perceivable and significant, most notably for frequency bands of around 500 Hz and above. The aim of this second experiment is to determine the actual extent of change between vertically decorrelated, correlated and monophonic stimuli, through absolute measurement of the VIS upper and lower auditory boundaries. Results of the experiment will establish the accuracy and agreement of defining VIS boundaries, as well as in which direction any changes of VIS occur.

4.3.1 Experimental Design

The absolute grading experiment used the same physical loudspeaker setup as the relative grading experiment, based on Auro-3D 9.1 (Auro Technologies, 2015a) (see Section 4.2.1.1); however, only the Centre 0° and Front ±30° vertically-arranged loudspeaker pairs were tested (see Figure 4.5 below). This was due to the practical issue of localising and defining VIS in absolute terms from behind the listener. Listening tests of the second experiment were conducted in the same critical listening room as the first and under the exact same testing conditions, with time/level alignment and an acoustically transparent curtain. A vertical light-emitting diode (LED) strip was fixed beside the 0° loudspeakers, to aid the capture of the upper and lower auditory boundaries, whilst avoiding potential bias from visual markers on a physical scale (as seen in Figure 4.5). The LED strip was linked to a physical rotary controller (Griffin Powermate USB) through Cycling '74's Max, which the user could rotate to control the position of each boundary separately (depressing the controller knob to switch between boundaries). This response method was originally proposed and evaluated by Lee et al. (2016) – in their experiment on vertical phantom image localisation, it was found that the LED method improved the subject's response consistency and shortened the test duration, when compared against the conventional visual marker method.

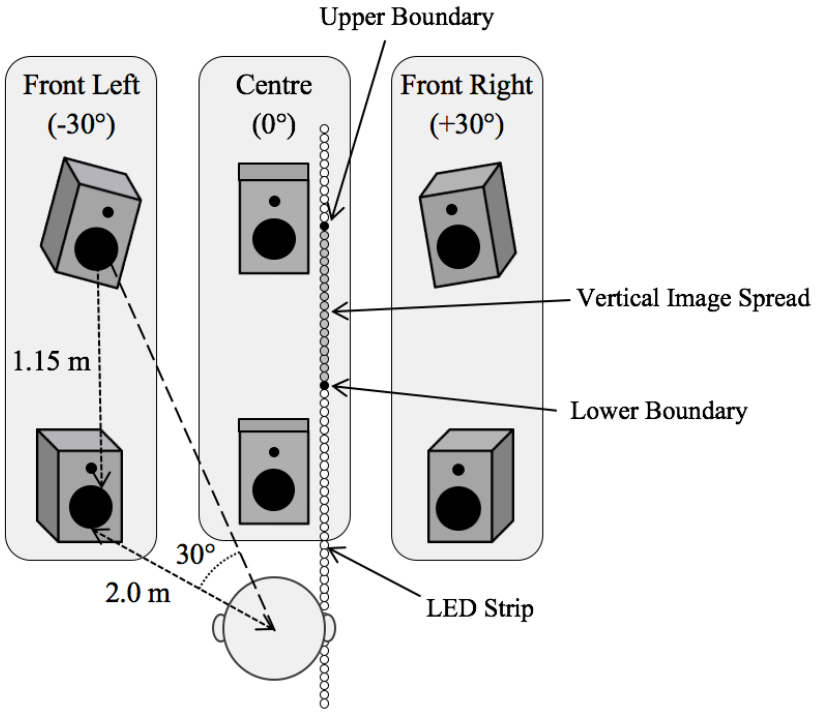

Figure 4.5 Loudspeaker setup for the absolute grading experiment, featuring a light emitting diode (LED) strip beside the 'Centre' 0° azimuth vertical loudspeaker pair for capturing responses.

In order to assess the absolute difference of VIS between the narrowest and broadest perceived samples, three stimuli from the first experiment were selected to test for each frequency band – these were the correlated stimulus ($ICCC_{avg}$ of 1.0), the monophonic stimulus (main-layer only) and the vertically decorrelated phase randomisation (PR) stimulus ($ICCC_{avg}$ of 0.1). PR was chosen over the complementary comb-filtering (CF) method as the first experiment demonstrated that, for every frequency band, PR with an $ICCC_{avg}$ of 0.1 consistently had the same or greater VIS than all other stimuli.

Nine subjects took part in the second experiment, all of who also participated in the first. In a single trial, the subject was asked to define the absolute upper and lower boundary of the auditory VIS, one stimulus at a time, using the LED strip and rotary controller connected to Max. There were 60 trials in total: three stimuli (decorrelated, correlated and monophonic) for ten frequency bands (63 Hz – 16 kHz octave-band and Broadband), tested at two loudspeaker azimuth positions (0° and ±30° (randomised between left and right)). The test was repeated twice

for each subject over two listening sessions of around 15-20 minutes each, and all 60 trials were randomised for each individual session. Repeating the test increases the statistical power of the test from 0.35 to 0.67, based on a two-tailed t-test with an effect size of 0.6 and α error probability of 0.05 (type I error i.e. false-positive), as calculated using G*Power 3.1 (Faul et al., 2007) – from this, the probability of a type II statistical error (false-negative) is 0.33 (β).

4.3.2 Results and Analysis

Results from the absolute vertical image spread (VIS) testing can be seen in Figure 4.6 below. Each stimulus condition features two box plots that represent the data for its lower and upper VIS boundaries – where the grey box plot on the left is the lower boundary and the white box plot on the right is the upper boundary. The line within the boxes signifies the median of subjects' responses, and the extent of the boxes indicates the 1st and 3rd quartiles (the interquartile range) of the data. To help quantify the spread of data around the median point, median absolute deviation (MAD) values were also calculated from the responses of each boundary position (shown in Table 4.3). Furthermore, to visualise the general spatial impact of the boundary positions between stimuli and frequency bands, the median overall VIS (the difference between the two boundaries) is displayed in Figure 4.7 below – these values were obtained by calculating the overall VIS of each individual response, then taking the median values from the calculated data.

Bonferroni-corrected Wilcoxon signed-rank tests have been conducted between stimuli within each octave-band, assessing the perceived difference of upper boundary position, lower boundary position and the overall VIS (the difference between the two boundaries). Further corrected Wilcoxon tests were also conducted between the 'Centre' (0°) and 'Front' (±30°) loudspeaker positions for the boundaries and overall VIS within each condition. In the following results, "Mono" refers to a signal from the lower main-layer loudspeaker only, "1.0" is the correlated stimulus and "0.1" is the vertically decorrelated stimulus.

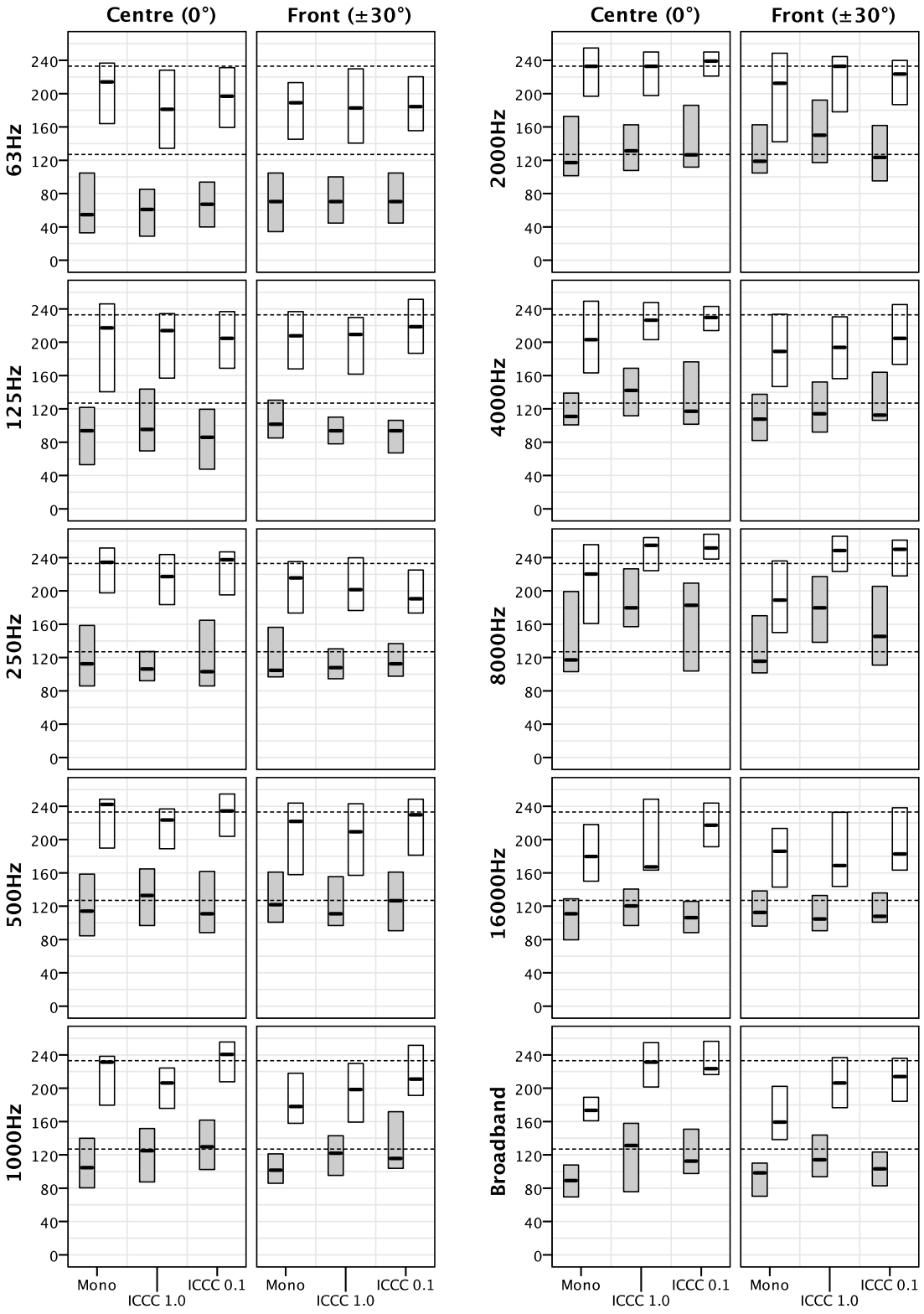

Figure 4.6 Box plots displaying the absolute location for the upper and lower boundaries of the auditory Vertical Image Spread (VIS) (cm), where each box features the median (the 2nd quartile) and interquartile range (1st to 3rd quartile) of a boundary position. The grey shaded boxes on the left show the lower boundary of a condition and the white boxes on the right show the upper boundary. The two dashed lines indicate the acoustic centres of the lower and upper loudspeakers.

Table 4.3 Median Absolute Deviation (MAD) (cm) of the absolute Vertical Image Spread (VIS) lower and upper image boundaries for each stimuli condition in the Centre (0°) and Front (±30°) positions.

Centre (0°)	63 Hz	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	16 kHz	BB
Mono Upper	25	36	23	8	16	22	41	39	31	11
Mono Lower	27	30	30	34	23	25	14	23	23	19
ICCC 1.0 Upper	45	27	30	16	19	25	19	14	17	23
ICCC 1.0 Lower	30	28	16	31	27	23	27	30	20	28
ICCC 0.1 Upper	34	33	17	23	17	14	13	16	25	13
ICCC 0.1 Lower	25	34	36	30	27	25	45	56	16	19

Front (±30°)	63 Hz	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	16 kHz	BB
Mono Upper	31	30	31	30	31	52	42	41	36	23
Mono Lower	33	17	28	31	16	25	27	22	19	22
ICCC 1.0 Upper	42	27	25	34	31	19	36	22	39	28
ICCC 1.0 Lower	27	14	17	17	22	36	25	36	20	23
ICCC 0.1 Upper	31	31	17	25	23	23	39	22	33	23
ICCC 0.1 Lower	28	22	20	33	42	30	9	39	11	19

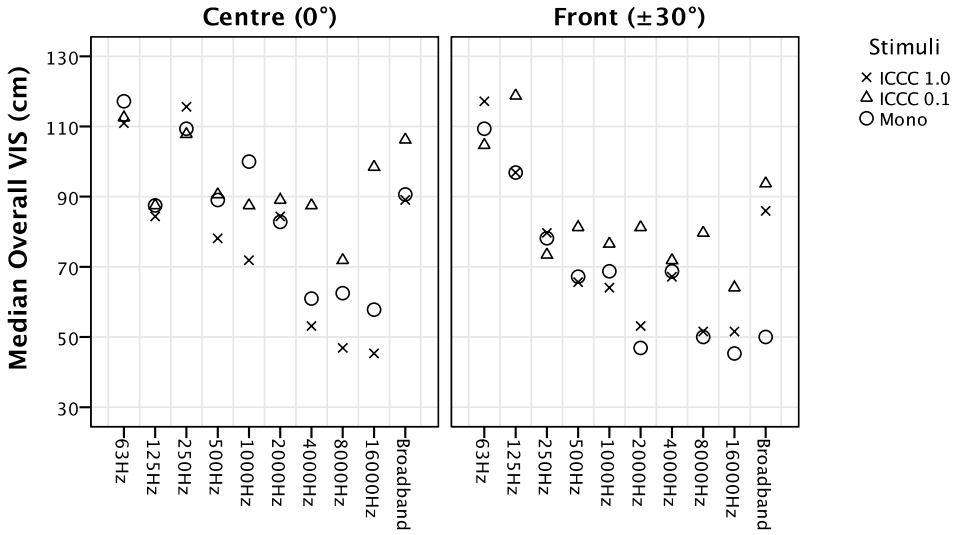

Figure 4.7 The absolute median of the overall Vertical Image Spread (VIS) for each condition (cm), where the raw overall VIS scores were calculated as the difference between the upper and lower boundaries for each individual subject response before averaging.

4.3.2.1 63 Hz – 500 Hz Bands

For octave-bands of 63 Hz to 500 Hz, the Wilcoxon tests revealed no significant change to the VIS boundary positions and the overall VIS between stimuli within each octave-band and loudspeaker position ($p > 0.05$). Likewise, for the same bands, there was little significant difference between the 0° and ±30° azimuth positions – the only stimulus showing a significant boundary and VIS change between loudspeaker angles was the 250 Hz '0.1' condition, with a

significantly narrower VIS at ±30° ($p = 0.03$). Looking at the overall VIS plots in Figure 4.7, it is seen that a similar trend occurs with all of the 250 Hz stimuli – when presented at the 0° azimuth, each stimulus has a vertical extent of around 110-120 cm, whereas from ±30°, the extent of each condition decreases to around 70-80 cm.

The box plots in Figure 4.6 show this decrease of the 250 Hz band is from the upper boundary shifting downwards in each case, with the lower boundary remaining fairly constant across all conditions and loudspeaker angles. Given that the three 250 Hz conditions have a relatively consistent overall VIS when presented at the same angle, it would suggest that this difference in perception between angles may be environmental, potentially due to a strong ceiling reflection increasing the upper boundary position in the median plane. In contrast, the 125 Hz octave-band appears to have a slight increase of overall VIS when presented at ±30° – this is most notable for the '0.1' condition, indicating a possible relationship with vertical decorrelation. Furthermore, Figure 4.7 also reveals that the median overall VIS of the 63 Hz octave-band stimuli is consistently large for each condition and loudspeaker angle (between 100-120 cm), supporting the minimal change of VIS seen for 63 Hz in the relative grading experiment. This supports the literature that states low frequencies are perceived as inherently broader than high frequencies (Cabrera & Tilley, 2003).

4.3.2.2 1 kHz Band

Wilcoxon results for the 1 kHz band show that, for both 0° and ±30° loudspeaker positions, the '0.1' upper boundary was significantly higher than the '1.0' upper boundary ($p < 0.01$), indicating an upward spread by decorrelation. Furthermore, '0.1' at ±30° also had significantly higher upper and lower boundaries than the 'Mono' stimulus ($p < 0.03$), demonstrating an upward image shift off-centre. As with 250 Hz, there appears to be some decrease of overall VIS for 1 kHz at ±30° compared to the 0° (most notably with the 'Mono' condition) (Figure 4.7); however, the Wilcoxon tests reveal these changes to be insignificant ($p > 0.05$).

4.3.2.3 2 kHz Band

Looking at the 2 kHz octave-band, there were no significant changes to boundaries or overall VIS at 0° ($p > 0.05$). From ±30°, '1.0' had a significantly higher lower boundary than both the '0.1' and 'Mono' conditions ($p < 0.03$) – this led to a significantly narrower VIS for '1.0' at ±30° compared to 0° ($p = 0.02$). Furthermore, the median absolute deviation (MAD) (Table 4.3) increases as the azimuth angle increases from 0° to ±30° for the '1.0' lower boundary (from 23 to 36 cm), as well as the 'Mono' upper boundary (from 22 to 52 cm), suggesting that off-centre presentation of the 2 kHz band may increase the difficulty of determining the upper and lower boundary positions.

4.3.2.4 4 kHz Band

The 4 kHz band showed no significant difference between stimuli boundaries within the 0° and ±30° loudspeaker positions ($p > 0.05$); however, both boundaries of '1.0' were significantly higher at 0° than ±30° ($p < 0.04$), indicating an upward image shift in the median plane. Despite this, there were no significant changes to the overall VIS, either within or between loudspeaker positions. It appears from Table 4.3 that this is probably due to a large disagreement amongst listeners – in particular, the upper boundaries at ±30° seem especially difficult to define (MADs = 36-42 cm). From 0°, the upper boundary of the 'Mono' condition also has a great deviation (MAD = 41 cm), which could be due to a differing perception of the pitch-height effect between subjects. It has previously been shown that the pitch-height effect (where higher pitched sounds are perceived higher in space than lower pitched ones) is evident with a 4 kHz octave-band in the median plane, when presented both monophonically and in vertical stereophony (Lee, 2016b; Wallis & Lee, 2015b). With regard to vertical decorrelation of the 4 kHz band, a considerable increase of overall VIS is seen from '1.0' to '0.1' at 0° azimuth (Figure 4.7); however, the lower boundary MAD of '0.1' (45 cm) is notably greater than '1.0' (25 cm), causing a statistically insignificant result due to disagreement amongst listeners ($p > 0.05$).

4.3.2.5 8 kHz Band

Both the 0° and ±30° results for the 8 kHz band show a significant upward shift for '1.0' from the 'Mono' condition ($p < 0.03$). Likewise, the '0.1' upper boundary was significantly higher than the 'Mono' condition for both positions ($p < 0.01$) – however, the '0.1' lower boundary was not significantly different ($p > 0.05$), due to a downward spread of VIS (though not to a significant level ($p > 0.05$)). Although there was no significant difference of overall VIS between positions or stimuli, Figure 4.7 shows that the '0.1' condition had the greatest VIS at both azimuth angles. There was also a relatively large lower boundary deviation for the '1.0' and '0.1' conditions at both loudspeaker angles (MADs = 30-56 cm), where the upper boundary deviation was comparatively small and located around the height loudspeaker (MADs = 14-22 cm) – this suggests that vertical stereophony elevates the upper boundary and auditory image towards the height loudspeaker, but defining the lower boundary becomes more ambiguous. In contrast, the upper boundary of the 8 kHz 'Mono' condition has a greater deviation than the lower boundary from both angles, which may be due to confusion from the 'directional band' or 'pitch-height' effect (Blauert, 1969/70; Cabrera & Tilley, 2003). Looking at the box plots in Figure 4.6, the upper quartile for both 'Mono' lower boundaries is spread upwards from the median (more so at 0°), indicating that some listeners perceived an elevated 'Mono' image, while others did not.

4.3.2.6 16 kHz Band

For the 16 kHz results at 0° azimuth, '0.1' had a significantly greater VIS than both the '1.0' and 'Mono' conditions ($p < 0.04$) – this was due to a significant upward movement of the upper boundary for '0.1' ($p < 0.03$). There were no significant differences between boundaries or overall VIS for stimuli from the ±30° angle ($p > 0.05$). However, as with the 0° position, a slight increase for '0.1' is also seen at ±30° (Figures 4.6 and 4.7). This was not a significant change as the MAD of the upper boundaries for all conditions at ±30° are relatively great (MADs = 33-36 cm), implying a difficultly with defining the 16 kHz octave-band upper

boundary at this angle. In comparison to the 8 kHz octave-band, it is observed that the boundaries for all 16 kHz stimuli are localised towards the lower loudspeaker at both 0° and ±30°, suggesting a lack of elevation effect when introducing 16 kHz in a height-channel.

4.3.2.7 Broadband

With the Broadband stimuli, there was some significant difference between boundaries for both loudspeaker positions. At 0° azimuth, the upper and lower boundaries of '0.1' and '1.0' were both significantly higher than the 'Mono' boundaries ($p < 0.01$), indicating an upward image shift when introducing a height-channel. Whereas from ±30°, only the upper boundary of '0.1' and '1.0' were significantly higher than the 'Mono' sample ($p < 0.02$), with the lower boundary remaining consistently below the lower loudspeaker position for all conditions. Furthermore, between loudspeaker conditions, the '0.1' upper boundary was significantly higher at 0° than at ±30° ($p < 0.01$). In terms of overall VIS, only '0.1' was significantly greater than the 'Mono' stimulus at ±30° ($p < 0.04$) (supporting the results seen in the first experiment), although Figure 4.7 also shows '0.1' as having the greatest VIS at 0° azimuth as well.

4.3.3 Discussion of Absolute Testing Results

The absolute results show no significant change to vertical image spread (VIS) boundaries or overall VIS for octave-bands of 500Hz and below. This is broadly in line with the results from the relative grading experiment (Section 4.2), where it is generally seen that the effect of inter-channel cross-correlation (ICC) on VIS occurs significantly around the 500 Hz octave-band and above. The only cases of decorrelation influencing a significant increase to the overall VIS (the difference between the upper and lower boundaries) are for the 16 kHz octave-band at 0° and the Broadband condition at ±30°. However, the median values in Figure 4.7 demonstrate that the '0.1' condition (the decorrelated stimulus) consistently had a greater overall VIS than '1.0' (the correlated stimulus) for all octave-bands of 500 Hz and above (as well as for the Broadband condition), supporting the relative grading results above.

In general, the changes of VIS appear to be fairly slight, with large deviations about the median point for many boundary positions (see the interquartile ranges in Figure 4.6 and the median absolute deviation (MAD) values in Table 4.3) – this is despite significant difference of VIS between the same conditions in the relative grading experiment (Section 4.2). These results suggest that perceiving changes to boundaries and vertical extent is probably easier when comparing stimuli relatively (as in the relative grading experiment), whereas absolutely defining the image boundaries independently is a rather difficult task. In particular, the 8 kHz octave-band at ±30° had the strongest relationship between ICC and VIS in the first experiment of this chapter (which is reflected by the absolute increase of overall VIS in Figure 4.7). However, a noticeably high deviation of lower boundary responses for both '1.0' and '0.1' (MADs = 36-39 cm) delivered a statistically insignificant VIS change between them.

Similarly, the Broadband results from the relative grading experiment demonstrated a significant relative VIS increase by decorrelation, although the absolute results only show an insignificant increase of overall VIS from '1.0' to '0.1' in Figure 4.7. It is possible that the absolute measures of the Broadband stimuli were judged based on the inherent large spread of low frequencies, as seen with the 63-250 Hz octave-bands in Figure 4.7 (i.e. the entire spectral image was considered); whereas the judgements in the relative grading could have been dictated by noticeable changes to VIS in the higher frequencies (e.g. within the 8 kHz octave-band). A study by Ferguson and Cabrera (2005) found that when a high frequency tweeter and low frequency woofer reproduced signals simultaneously (from vertically-spaced positions), subjects consistently localised the sound at the tweeter position, suggesting a dominance of high frequencies within spatial perception. Further investigation is required to determine whether changes of VIS at high frequencies take precedence over the VIS of low frequencies. Significant movements of upper and lower boundaries (not necessarily increasing VIS) may have also led to a relative perception of VIS change in the first experiment – the 1 kHz result is one case where the perception of VIS could have been dictated by changes to a single boundary or shift of image, rather than an increase to the absolute VIS.

For the 2 kHz and 8 kHz octave-bands, the '1.0' condition (the correlated stimulus) at ±30° appears to be elevated towards the height loudspeaker (in comparison to the 'Mono' condition) – the auditory image is then extended downwards from this position towards the lower loudspeaker following decorrelation. In the case of the 8 kHz band, the bias towards the height-layer loudspeaker is very strong at both 0° and ±30° when a height-channel is present (i.e. vertical stereophony). The results presented here suggest that the vertical localisation cues from an elevated 8 kHz octave-band source have dominance over a lower positioned correlated source, when both are presented simultaneously. A large spectral notch around 8 kHz (caused by the pinna from the main-layer loudspeaker direction) is likely to be the cause of this effect, resulting in more energy for the 8 kHz band from the height-channel direction (Lee, 2016b). Furthermore, the downward extension of the 8 kHz band following decorrelation suggests that both loudspeaker signals might be perceived independently, with the decorrelation process reducing the height dominance to 'un-mask' the main-layer loudspeaker. In contrast, all of the 16 kHz stimuli have a bias towards the main-layer loudspeaker, with a significant upward extension for the decorrelated stimuli. To investigate this further, a detailed analysis of the spectra for these key bands is presented in Chapter 5 (Section 5.3). The findings here somewhat support the hypothesis set out in Chapter 3 (Section 3.5.2), where the perception of the 'High' frequency band (4-16 kHz octave-bands) may have been vertically broad prior to decorrelation, due to the inherent vertical distribution of octave-bands based on the 'pitch-height effect' (Figure 3.11) (Cabrera & Tilley, 2003; Wallis & Lee, 2016; Lee, 2016b).

In the relative grading experiment, it was found that some 'Mono' conditions were perceived as having the same or greater VIS than vertically decorrelated stimuli. At 0° azimuth, this was the case with the 500 Hz, 1 kHz, 2 kHz and 16 kHz octave-bands, and at ±30°, all octave-bands from 500 Hz and above showed some increase of VIS for the 'Mono' condition. Considering the absolute results of the 0° bands first, the 500 Hz, 1 kHz and 16 kHz 'Mono' conditions all demonstrate an overall VIS greater than the '1.0' condition (Figure 4.7), supporting the relative grading results. However, with the 2 kHz band, the boundaries of the 'Mono' and '1.0'

condition both feature similarly great deviations (MADs = 23-25cm) – this would suggest that the absolute spread of the 2 kHz band stimuli was difficult for listeners to define, but when compared relatively, the differences may have been more apparent.

Looking at the ±30° absolute results for the 500 Hz octave-band and above, only the 500 Hz and 1 kHz bands show a very slight increase of overall VIS for the 'Mono' condition compared to '1.0'. Observing the median absolute deviation (MAD) of boundaries in Table 4.3, there is a clear inconsistency when grading the upper boundary for all 'Mono' stimuli at ±30° azimuth (except Broadband) (MADs = 30-52 cm). This would further imply a difficulty with defining absolute boundaries of VIS, where differences might be more obvious through a relative comparison of stimuli. It is not just the 'Mono' stimuli that see an increase of boundary deviation at ±30° compared to 0° – for instance, both the 4 kHz and 16 kHz octave-bands have noticeable increases to the vertical stereophonic upper boundary MADs at ±30°. Although these deviation results may be from differing perceptions or a difficulty defining the boundaries, it is also important to note that they could be related to the LED response method used during testing. Since the LED strip capturing a subject's responses was positioned in the centre (0° azimuth), subjects were tasked with translating the boundary locations from the ±30° presentation across to 0° – this may have been more difficult for some octave-bands or conditions than others, causing the apparent inaccuracy amongst listeners.

4.4 Practical Implications

The results from both the relative and absolute grading in the present chapter suggest that vertically decorrelating signals below the 500 Hz octave may not be necessary for VIS rendering. In general, the lower frequency bands (63-250 Hz) have an inherently broad VIS and are localised in similar positions between each condition (generally towards the lower loudspeaker) – this is regardless of whether they are presented monophonically or decorrelated vertically. Both of these points would suggest that an increase of VIS upwards might still be achieved without these frequencies e.g. for upmixing from 2D-to-3D loudspeaker formats. It is possible that decorrelating a high-pass filtered broadband signal vertically, while routing the low-pass component to the main-layer loudspeaker only, would have a similar effect to decorrelating the entire signal. This approach could have an impact on the tonal quality or clarity of a reproduction, by reducing the number of frequencies interacting at the ears.

Within the octave-band stimuli, vertical decorrelation appears to be particularly effective for the 8 kHz octave-band. These results are in agreement with the literature that states frequencies around 8 kHz are important for vertical localisation in the front median plane (Roffler & Butler, 1968a; Hebrank & Wright, 1974). A previous study by Chun, Kim, Choi, Jang and Lee (2011) has shown that boosting a band around 8 kHz alone can increase the vertical elevation of a broadband signal, based on Blauert's boosted band hypothesis (Blauert, 1969/70); so potentially a similar process could also control the vertical extent of an image as well. Having said that, there seems to be a general dominance of the height-channel signal when presenting the 8 kHz octave-band signals in vertical stereophony. As mentioned above, a reason for this could be due to an increase of energy for the 8 kHz band in the HRTF when presented from the height-channel direction (at +30° elevation) compared to the main-channel (0° elevation) (Lee, 2016b). A similar elevation effect for the 8 kHz octave-band (when presented in vertical stereophony) was also seen by Wallis and Lee (2015b, 2016), who investigated the height-channel gain reduction required to localise stereophonic octave-band conditions at the height of a monophonic

reference (the lower loudspeaker only). For the 8 kHz octave-band, the required upper channel level reduction was found to be around -9 dB (with respect to the unchanged level of the lower loudspeaker channel), in order to localise the auditory image at the position of the lower main-layer loudspeaker.

4.5 Conclusion

This chapter described a two-part investigation into the perception of vertical image spread (VIS) by vertical interchannel decorrelation. In both experiments, octave-band (centre frequencies: 63 Hz to 16 kHz) and broadband pink noise stimuli were tested to assess the frequency dependency of VIS at different azimuth angles to the listener.

For the first experiment, the relative VIS between stimuli was graded on a bipolar scale in multiple-comparison trials. Stimuli were presented through pairs of vertically-arranged loudspeakers, positioned at three azimuth angles to the listener (0°, ±30° and ±110°). Eight stimuli were compared for each frequency band and loudspeaker angle combination – these featured two decorrelation methods with three levels of average running interchannel cross-correlation coefficients ($ICCC_{avg}$) (0.1, 0.4 and 0.7), along with correlated ($ICCC_{avg}$ 1.0) and monophonic (lower loudspeaker only) cases. The two decorrelation methods assessed were all-pass filter phase randomisation (PR) and the amplitude-based complementary comb-filtering (CF). In the second experiment, absolute VIS was measured for the monophonic, $ICCC_{avg}$ 1.0 and $ICCC_{avg}$ 0.1 (PR) stimuli from the first experiment. The same frequency bands were also tested, but only for azimuth angles of 0° and ±30° due to practical reasons. Subjects defined the upper and lower boundaries of the VIS for each stimulus independently using a light-emitting diode (LED) strip.

The key findings from the two experiments are as follows:

- Interchannel cross-correlation (ICC) begins to have a significantly linear effect on VIS for frequencies around the 500 Hz octave-band and above at all azimuth angles; that is, VIS increases as the signal correlation between the loudspeakers decreases.
- The strongest correlation between ICC and VIS was for the 8 kHz band at ±30°.
- The strength of association between ICC and VIS appears to be divided into three directivity groups, as follows:

  - 500 Hz and 1 kHz had the strongest effect in the median plane (0° azimuth, where energy is equal in both ears).
  - 2, 4 and 16 kHz at ±110° azimuth (where head-shadowing is greatest).
  - 8 kHz and Broadband at ±30° azimuth (where both interaural differences and frontal HRTF filtering from the pinna are present).
- In general, there was very little difference between the two decorrelation methods, with PR having a slightly greater VIS in some cases.
- The 'Mono' condition had greater VIS than vertically decorrelated stimuli in some instances, possibly due to the 'pitch-height' or 'directional bands' effects – there may have also been a room effect (reflections) influencing some monophonic results.
- The absolute results demonstrate that significant changes to the boundaries occurred for stimuli above the 500 Hz octave-band, similar to the relative grading.
- Great deviations of boundary responses suggest an inherent difficulty with absolutely defining the image of an auditory event.
- A potential pitch-height or directional bands elevation effect may have caused greater deviations of perception, particularly with the 1 kHz, 4 kHz and 8 kHz bands.
- A strong bias towards the height loudspeaker was seen with the 8 kHz octave-band.
- The relative perception of VIS may have been influenced by changes to a single boundary or shift of image, rather than an extension of overall VIS e.g. 8 kHz.
- Lower frequencies tend to have a greater VIS than higher frequencies.

To investigate these findings further, all stimuli have been objectively analysed in Chapter 5 and discussed with relation to the results presented here. In particular, potential spectral cues of vertical decorrelation have been observed, in an attempt to understand the strength of association between higher frequencies and VIS perception (most notably the 8 kHz octave-band). Furthermore, interaural differences have been considered, given the apparent azimuth angle-dependency on the perception of some octave-bands. Lastly, binaural room impulse responses

of the listening room are analysed, in order to observe whether a room effect could have influenced the VIS results seen in the current chapter.

5 OBJECTIVE ANALYSIS OF VERTICALLY DECORRELATED OCTAVE-BAND PINK NOISE STIMULI[5]

This chapter follows directly on from Chapter 4 of the present thesis, where two subjective experiments have been described and the results discussed. The first of the two experiments looked at the relative perception of vertical image spread (VIS) by vertical decorrelation of octave-band stimuli (Section 4.2); and the second experiment assessed the absolute extent of VIS in space, using extreme stimuli conditions from the first (Section 4.3).

In the first subjective experiment of Chapter 4, subjects were asked to grade the relative vertical image spread (VIS) of stimuli against each other and a reference. Ten pink noise frequency bands were assessed: nine octave-band (with centre frequencies from 63 Hz to 16 kHz) and one broadband (20 Hz to 20 kHz). Within each frequency band, eight stimuli were presented in a multiple comparison – these consisted of three average-running interchannel cross-correlation coefficients ($ICCC_{avg}$ of '0.1', '0.4' and '0.7') from two decorrelation methods, along with a correlated condition ($ICCC_{avg}$ of '1.0', which was also the reference stimulus) and a monophonic condition (lower loudspeaker only). The two decorrelation methods tested were complementary comb-filtering (CF) (Lauridsen, 1954; Schroeder, 1958; Breebaart & Faller, 2007) and phase randomisation by all-pass filtering (PR) (Kendall, 1995). All stimuli were presented from vertically-arranged loudspeaker pairs at three discrete azimuth positions (0°, ±30° and ±110°), where the height-channel loudspeakers are elevated by +30°. This resulted in a total of 30 multiple comparison trials, one for each frequency band and azimuth angle combination.

In the second experiment of Chapter 4, the absolute VIS was measured for three extreme stimuli from the first experiment. These were the monophonic condition, the correlated condition and a vertically decorrelated condition (the PR method with an $ICCC_{avg}$ of 0.1). The same ten

[5] Gribben, C., & Lee, H. (2018). The Frequency and Loudspeaker-Azimuth Dependencies of Vertical Interchannel Decorrelation on Vertical Image Spread. *Journal of the Audio Engineering Society,* (accepted May 2018).

frequency bands were assessed and each stimulus was independently presented from both the 0° and ±30° azimuth angles. A vertical light emitting diode (LED) strip positioned by the 0° loudspeaker pair was used to help capture the responses, where subjects were asked the define the upper and lower boundary of the auditory image. Further details of the decorrelation methods and the testing conditions for both experiments can be found in Chapter 4.

Results from the first experiment indicate that a linear relationship between vertical interchannel cross-correlation (ICC) and vertical image spread (VIS) begins to develop around the 500 Hz octave-band and above. Little difference was also shown between the two decorrelation methods, indicating that changes of VIS might be controlled by the level of ICC. In general, the relationship between ICC and VIS for the 500 Hz and 1 kHz octave-bands was strongest in the median plane (0° azimuth, where energy is equal in both ears); for 2 kHz, 4 kHz and 16 kHz, the relationship was strongest at ±110° azimuth (where head-shadowing is greatest); and 8 kHz and Broadband were strongest at ±30° azimuth (where both interaural differences and frontal pinna filtering are present). The greatest statistical correlation between ICC and VIS from all conditions was with the 8 kHz octave-band from the ±30° position.

The second experiment broadly supported the results of the first, where significant changes to VIS boundaries were only seen for octave-bands above 500 Hz. There was some suggestion of a pitch-height (Lee, 2016b; Cabrera & Tilley, 2003) and/or directional bands (Blauert, 1969/70; Wallis & Lee, 2015a) effect causing deviations of VIS perception amongst listeners, where some subjects potentially perceived elevation, while others did not. Furthermore, an upward shift of an auditory image might have also been perceived as a change to VIS – this is particularly the case with the 8 kHz octave-band, where the vertical stereophonic image was biased towards the height-channel loudspeaker. In the objective analysis of Chapter 3 (Section 3.5.2), potential spectral cues of vertical decorrelation were observed around the 8 kHz and 16 kHz octave-bands in the median plane – it is possible that such cues influenced the elevation of the

8 kHz image seen in the absolute testing results, as well as contributed to the strong association between ICC and VIS at 8 kHz observed with the relative grading.

From the subjective results of Chapter 4, the following research questions are proposed:

- Do spectral cues exist that might contribute to the perception of VIS?
- Does the degree of vertical (de)correlation directly relate to the strength of cues?
- How do spectral cues differ between azimuth angles?
- Is there an interaural relationship that contributes to VIS at wider azimuths?
- Does the room have any effect on the perception of VIS?

In order to answer the above questions, the stimuli used in the subjective experiments of Chapter 4 have been binaurally synthesised for each individual loudspeaker condition (as detailed in Section 5.2 below), with their frequency spectra analysed and discussed in Section 5.3. Section 5.4 then looks at the interaural cross-correlation (IAC) of the binauralised stimuli, followed by Section 5.5 where binaural room impulse responses (BRIRs) of the listening room have been analysed.

5.1 Hypotheses

It is widely known that pinna spectral filtering from 3 kHz and above plays an important role in vertical localisation, particularly with frequencies around 8 kHz (Roffler & Butler, 1968a; Hebrank & Wright, 1974). Considering this, and taking into account the results above, it is hypothesised that similar spectral changes at high frequencies will also contribute to the perception of VIS in the median plane. The subjective results also suggest that interaural differences improve VIS perception off-centre (particularly for the 2 kHz octave-band and above). The literature demonstrates that the interaural level difference (ILD) is most effective for these frequencies (Section 1.1.1.1), suggesting that head-shadowing may have contributed to VIS perception at ±110°. From this, it is hypothesised that a relationship exists between the vertical interchannel cross-correlation and interaural differences, which is explored further in Section 5.4 below. Finally, in some cases, the monophonic condition had a greater perceived VIS than vertically decorrelated conditions – it is thought that this may be due to reflective energy in the listening room. In Chapter 3, there was suggestion of a floor reflection influencing the perception of the 'Low' frequency band; therefore, an additional hypothesis is that the ratio of early reflection energy to direct sound energy (ER/D) will reveal some association between early room reflections and the perception of VIS, particularly in the median plane.

5.2 Binaural Synthesis of Stimuli

For the objective analysis, all 240 stimuli from the first subjective experiment in Chapter 4 (Section 4.3) were convolved with two sets of binaural impulse responses. Each set features ten binaural impulses for the ten loudspeaker positions used during testing. These loudspeaker positions comprised five main-layer and five height-layer angles of incidence, based on Auro-3D 9.1 (with an additional centre height-channel) (Auro Technologies, 2015a) (Figure 4.3 in Section 4.3.1.1 of Chapter 4). From the main-layer, the azimuth angles were 0°, ±30° and ±110° at 0° elevation (in line with ear level), and the height-layer positions had the same azimuth angles with an elevation angle of +30° to the listening position. For both impulse response sets, three conditions were created for each stimulus and azimuth angle combination during the convolution process: the main-layer only (Equations 5.1 and 5.2), the height-layer only (Equations 5.3 and 5.4) and the layers combined (where the time-aligned main- and height-layer impulses were summed) (Equations 5.5 and 5.6) (Figure 5.1). Objective analysis on both the sets of convolved stimuli have been carried out and discussed below, in an attempt to gain theoretical insights into the subjective perception of VIS by vertical interchannel decorrelation.

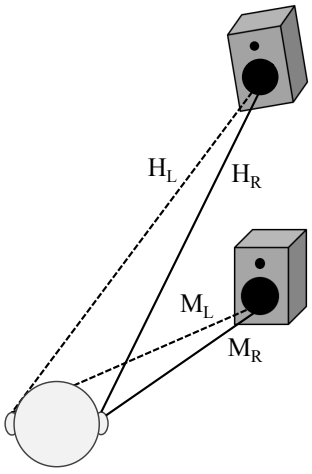

Figure 5.1 Diagram of the convolution process for a single loudspeaker pair. M_L and M_R are the left and right ear signals of the main-layer (lower) loudspeaker, and H_L and H_R are the left and right ear signals of the height-layer (upper) loudspeaker. Layer impulse are summed to replicate the test stimuli.

$$L_M = M_L * S_1 \qquad R_M = M_R * S_1 \qquad (5.1, 5.2)$$

$$L_H = H_L * S_2 \qquad R_H = H_R * S_2 \qquad (5.3, 5.4)$$

$$L_C = L_M + L_H \qquad R_C = R_M + R_H \qquad (5.5, 5.6)$$

where S_1 and S_2 are the two-channel stereophonic stimuli signals, M is 'Main', H is 'Height', and L_x and R_x are the left and right ear signals of the convolved stimuli.

The first set of impulse responses were taken from the MIT's anechoic head-related impulse response (HRIR) database, obtained using the KEMAR head and torso simulator (Gardner & Martin, 1994). These were applied to observe the effect of the HRTF and torso reflections on the stimuli signals, disregarding influence from the listening environment. Spectral analysis of the HRIR-convolved stimuli can be seen in Section 5.3, and calculation of the interaural cross-correlation coefficients (IACCs) for the HRIR-convolved stimuli are presented in Section 5.4.1 below.

The second set of impulse responses were captured from the loudspeakers in the listening room where the current experiments took place (Figure 4.3 of Section 4.3.1.1 in Chapter 4). Using the HAART impulse response capture toolbox (Johnson et al., 2015), binaural room impulse responses (BRIRs) were acquired for each of the ten loudspeakers independently, using the exponential sine sweep (ESS) method (Farina, 2000). To capture the impulses, a Neumann KU 100 dummy head was placed at the listening position, where the ear height was in line with the acoustic centre of the main-layer loudspeakers (0° elevation at a height of 1.27 m). Unlike the first set of impulse responses, the second did not feature a torso (thus producing no spectral notches due to the torso reflections below 3 kHz (Algazi et al., 2001)). Therefore, observations below around 3 kHz with these stimuli are likely to display effects of the listening room. Prior to the convolution, the listening room BRIRs were time and level aligned between the two loudspeaker layers, as with the alignment used during the subjective testing of Chapter 4.

IACCs are presented for the BRIR-convolved stimuli in Section 5.4.2 – spectral analysis was not conducted on the BRIR-convolved stimuli as the intention was to observe the effect of HRTF filtering, which is seen most clearly with the HRIR convolution.

5.3 Spectral Analysis of HRIR-Convolved Stimuli

In the subjective relative testing results of Chapter 4, there was a significant interchannel cross-correlation (ICC) effect on vertical image spread (VIS) from around the 500 Hz octave-band and above, where VIS increased as the ICC coefficient (ICCC) decreased. At high frequencies it is hypothesised that this change to VIS may be influenced by the head-related transfer function (HRTF) filtering of signals, particularly in the median plane (Hebrank & Wright, 1974). To investigate this, delta spectra of the frequency amplitude difference between the anechoic HRIR-convolved stimuli are presented below for each azimuth angle (0°, ±30° and ±110° in Figures 5.2, 5.6 and 5.10, respectively). Spectra were calculated using 4096 FFT-points with a frame length of 4096 samples – a Hann window was used on the frames with 50% overlap, and 1/6-octave Gaussian smoothing was applied to the FFT output.

In the delta plots, the broadband FFT spectra of the decorrelated signals ($ICCC_{avg}$ of 0.1, 0.4 and 0.7) have been subtracted from that of the correlated condition ($ICCC_{avg}$ of 1.0), to observe the differences of spectrum as correlation decreases. A positive amplitude difference indicates a boost of frequency for that particular ICC condition in comparison to $ICCC_{avg}$ 1.0. Both decorrelation methods are presented, with Phase Randomisation (PR) in the upper panel and Complementary Comb-Filtering (CF) in the lower. Spectral analysis is split by azimuth angle, with key frequency regions receiving further inter-layer analysis at each position. Since the spectral changes from vertical decorrelation are most apparent at higher frequencies, the FFT results below are presented with linear plots rather than logarithmic. This allows for a clearer observation of high frequency spectral changes, particularly within the 16 kHz octave-band.

5.3.1 0° Azimuth Spectral Analysis

Considering the 0° azimuth angle first, it can be seen in Figure 5.2 that there is very little change of spectrum between conditions below around 5 kHz for the CF method. On the other hand, the PR method shows some random magnitude fluctuation of frequencies below 5 kHz. These

spectral changes are mostly within ± 3 dB, although a larger notch can be seen at the 500 Hz octave-band. Observing the other delta spectra plots (Figures 5.6 and 5.10), the same spectral distortions are apparent at lower frequencies, indicating that these are within the stimuli source signal, rather than a change of HRTF response from ICC. Figure 5.3 confirms this, where the two decorrelated signals are summed for each ICC condition of the PR and CF methods (left and right panel, respectively).

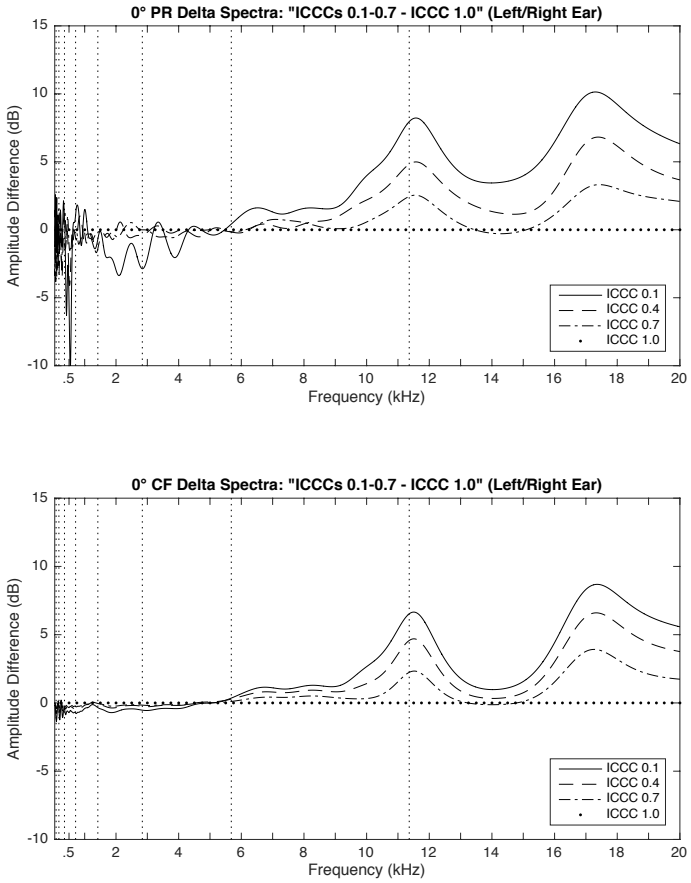

Figure 5.2 0° azimuth delta spectra of the FFT frequency amplitude difference between the HRIR-convolved correlated stimulus (ICCC 1.0) and the decorrelated stimuli (ICCCs 0.1-0.7). The vertical dotted lines signify the band limits of each octave-band. (Upper – PR Method; Lower – CF Method)

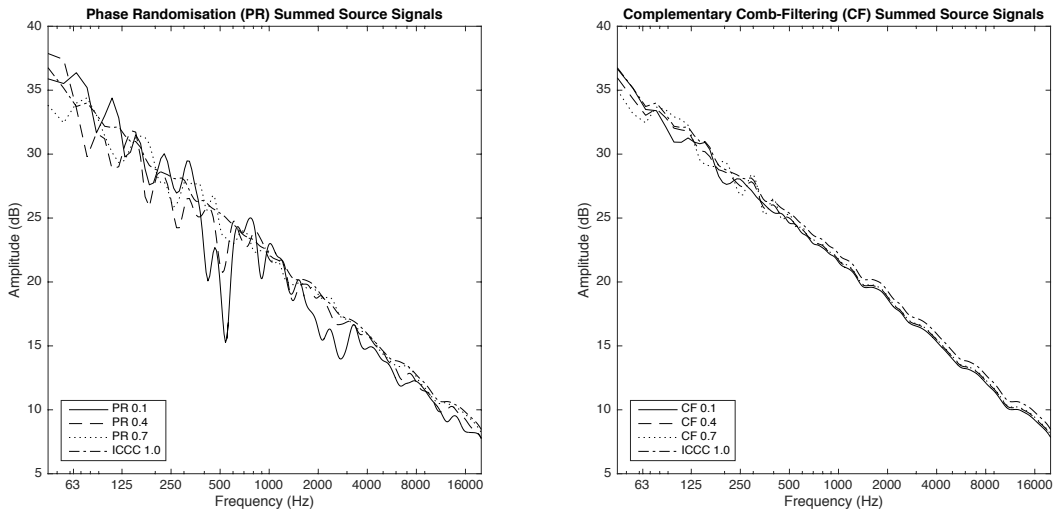

Figure 5.3 Summation of the two decorrelated output signals for both the PR (Left) and CF (Right) methods, displaying the sum of the actual broadband pink noise signals used during testing.

The ICCC 0.1 condition of the PR method in Figure 5.2 has clear notches around 500 Hz and 2 kHz, demonstrating a limitation of the PR method with regard to phase cancellation; whereas summation of the CF method decorrelated signals appear to reconstruct a broadband pink noise signal near perfectly. Since the greatest distortions appear to occur with the ICCC 0.1 condition of the PR method, it suggests that when larger degrees of phase variation are present between the two signals (i.e. more decorrelation), there is an increase of phase cancellation at middle to low frequencies when they are summed. Given the above observations, the low frequency distortion of the PR method shall not be considered in the following spectral analysis, as it is not an effect of ICC on HRTF filtering.

In terms of the ICC effect in Figure 5.2, a noticeable change to the spectrum begins around the 8 kHz octave-band and above for both decorrelation methods – that is, where a lower $ICCC_{avg}$ results in a general increase of high frequency energy. There are also clear spectral boosts at specific frequency points, which appear to increase as correlation decreases – these boosts strongly align between both methods, suggesting potential cues for the perception of vertical decorrelation in the median plane. The two greatest boosts are seen around 11-12 kHz and

17-18 kHz, which lie within the 8 kHz and 16 kHz octave-bands; consequently, both of these octave-bands are discussed further below.

5.3.1.1 8 kHz Band at 0°

Looking closer at the 8 kHz octave-band from 0° azimuth, Figure 5.4 displays the spectra for a monophonic main-channel signal, a monophonic height-channel signal, the correlated condition ($ICCC_{avg}$ 1.0) and two decorrelated conditions ($ICCC_{avg}$ 0.1 for PR and CF). The monophonic main- and height-channel signals have been level-matched with the vertical stereophonic stimuli – as a result, the monophonic main-channel signal also represents the monophonic condition used during the subjective testing, as described in Chapter 4.

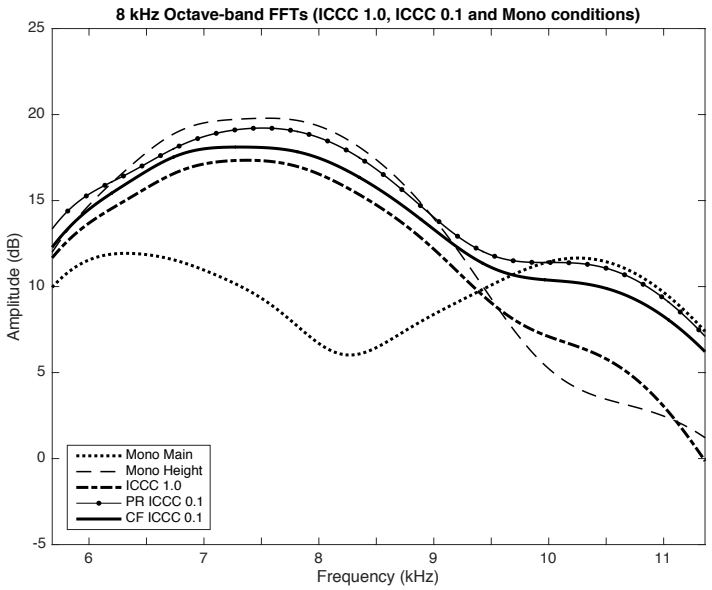

Figure 5.4 0° azimuth 8 kHz octave-band average FFT for the main-channel only, height-channel only and both channels combined for the ICCC 1.0 and ICCC 0.1 conditions.

In the subjective results of Chapter 4, it was noted that the absolute VIS position of the 8 kHz correlated condition was biased towards the height-channel loudspeaker at 0°, with decorrelation spreading the VIS downwards (Section 4.4.2.5). In Figure 5.4, a crossover between the main-channel and height-channel spectra can be seen around 9.5 kHz. The spectrum of the correlated condition appears to follow a similar downward trend to that of the height-channel

loudspeaker throughout the octave-band; however, the spectra of the decorrelated conditions ($ICCC_{avg}$ 0.1) are slightly boosted and follow the spectrum of the height-channel below 9.5 kHz, then switches to follow the main-channel above 9.5 kHz (i.e. the signal with the greater frequency amplitude). Taking into account the trend between ICC and spectral boosts in Figure 5.2 above (specifically around 10-11 kHz), it is recognised that the switch between loudspeaker dominance above 9.5 kHz is gradual and relates directly to the degree of ICC. This indicates that VIS in the median plane could potentially be controlled by frequencies around 10-11 kHz. It was also suggested in Chapter 4 that decorrelation of the 8 kHz octave-band might perceptually unmask the main-channel loudspeaker signal (causing the downward increase of VIS from the height-channel dominance) – the spectra presented here appear to somewhat support that hypothesis.

Further to the observations above, the large notch of the monophonic condition around 8 kHz is a known spectral cue of vertical localisation in the median plane (Hebrank & Wright, 1974; Lee, 2016b). As a source elevates upwards, the 8 kHz notch is gradually 'filled in', allowing for the auditory system to determine its vertical location. This is clearly demonstrated when comparing the main-channel and height-channel spectra in the plot above. Similarly, it is known that the pinna causes phase cancellation at higher frequencies (> 10 kHz) when sources are presented from height – this is shown by the decrease of energy around this region in the height-channel spectrum. Considering this, it is interesting to observe that both of these localisation cues appear to be 'filled in' with vertical interchannel decorrelation. It is assumed that less correlation between the loudspeaker signals results in less high frequency cancellation at the pinna (i.e. the cause of notches). From this, it might be said that vertical decorrelation removes these important elevation cues, causing a greater perception of vertical image spread (VIS) from an increase of vertical localisation ambiguity. Looking back at the delta spectra of the ICC effect in Figure 5.2, the spectral boosts are seen to increase in line with a decreasing ICC. However, rather than boosting the spectrum, this spectral change with ICC actually represents a 'filling in' of the vertical localisation notch and attenuation cues.

5.3.1.2 16 kHz Band at 0°

In general, the 16 kHz octave-band spectra of the decorrelated and main-channel conditions appear relatively similar throughout the band (Figure 5.5). However, when correlated signals ($ICCC_{avg}$ 1.0) are presented vertically, there seems to be some kind of phase cancellation and loss of frequencies, particularly above 16 kHz. This relates directly to the 'boosts' seen in Figure 5.2 above, where decorrelation reduces the degree of phase cancellation between the two correlated signals (similar to the effect observed for the 8 kHz band). This frequency cancellation of the correlated condition is probably due to the complex filtering and phase alteration that occurs at the pinnae (Hebrank & Wright, 1974), which is most prevalent at higher frequencies given the shorter wavelengths.

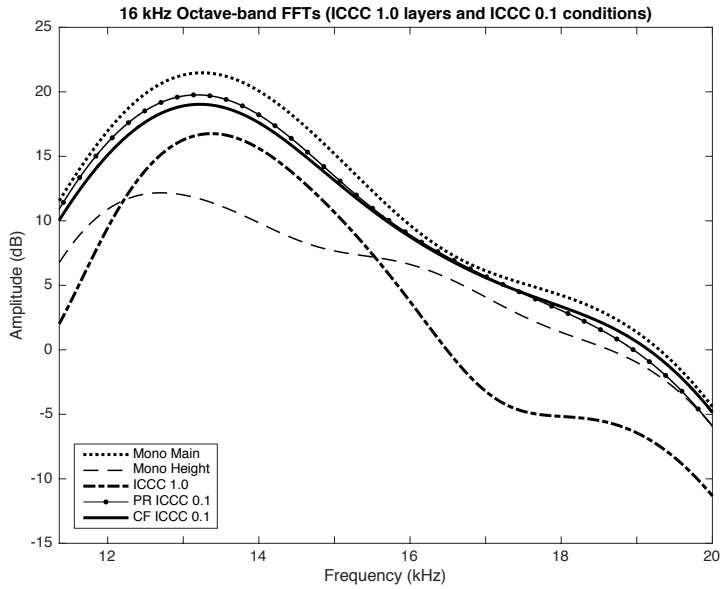

Figure 5.5 0° azimuth 16 kHz octave-band average FFT for the main-channel only, height-channel only and both channels combined for the ICCC 1.0 and ICCC 0.1 conditions.

Such a loss of frequencies may have caused the lower perception of VIS for the correlated condition seen in the subjective testing, due to a timbral 'thinning' effect or a decrease of the perceived loudness. This observation supports the idea that decorrelation can be beneficial to reducing high frequency spectral distortion, when presenting similar signals from multiple

directions. Since the monophonic main-layer condition also follows a similar trend to the decorrelated conditions, this may have led to the similar perception of VIS that was recorded for them both – it also implies a main-layer dominance, which was another observation made during the absolute testing. Unlike the 8 kHz band, there are no specific cues related to VIS perception within the 16 kHz band. This further suggests that VIS in the median plane is largely dictated by a 'filling in' of the localisation cues within the 8 kHz octave-band.

5.3.2 +30° Azimuth Spectral Analysis

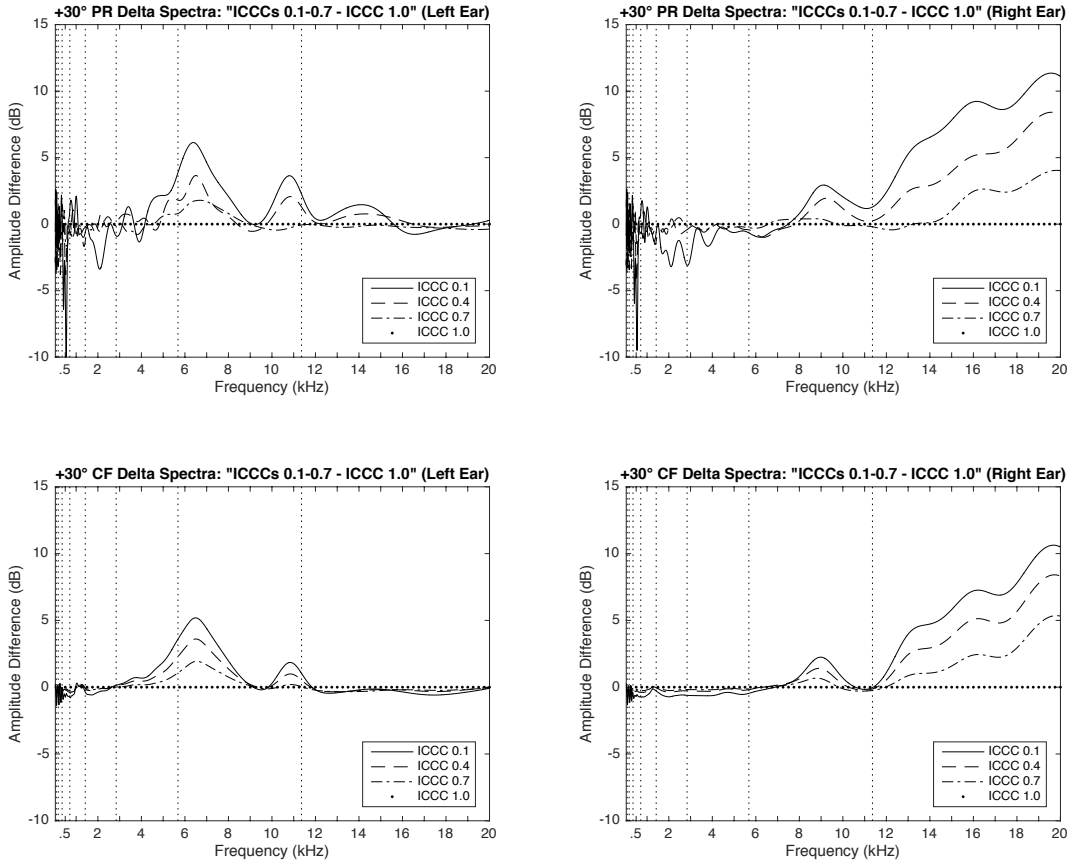

Figure 5.6 +30° azimuth delta spectra of the FFT frequency amplitude difference between the HRIR-convolved correlated stimulus (ICCC 1.0) and the decorrelated stimuli (ICCCs 0.1-0.7). The vertical dotted lines signify the band limits of each octave-band.
(Upper – PR; Lower – CF; Left – Contralateral Ear; Right – Ipsilateral Ear)

Now considering the stimuli from a +30° azimuth angle, large spectral differences can be seen between the left and right ears in Figure 5.6 above. With the contralateral left ear, the peaks of spectral boosts from decorrelation are seen around 6-7 kHz and 11 kHz (both of which are in the 8 kHz octave-band) – and there is also a slight increase of energy within the 4 kHz octave-band. In contrast, the ipsilateral right ear shows a boost around 8-10 kHz from decorrelation, as well as a general boost of high frequency energy above 12 kHz in the 16 kHz band. Considering these observations, the inter-layer relationship and changes of spectrum for the 4 kHz, 8 kHz and 16 kHz octave-bands have been investigated further below.

5.3.2.1 4 kHz Band at +30°

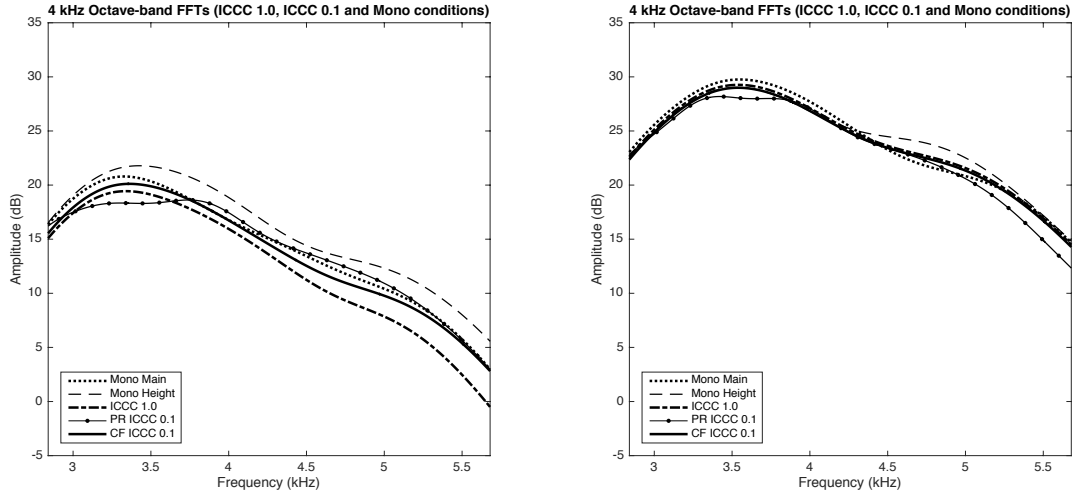

Figure 5.7 +30° azimuth 4 kHz octave-band average FFT for the main-channel only, height-channel only and both channels combined for the ICCC 1.0 and ICCC 0.1 conditions.
(Left – Contralateral Ear; Right – Ipsilateral Ear)

Inspecting Figure 5.7, the 4 kHz octave-band shows a small spectral boost above 4.5 kHz in the contralateral left ear for both decorrelated conditions ($ICCC_{avg}$ 0.1), when compared against the correlated condition ($ICCC_{avg}$ 1.0). Observing the ICC effect in Figure 5.6 above, it would appear that this slight boost increases as ICC decreases, particularly for the CF method – however with the PR method, there seems to be some slight spectral distortion, which may be related to the distortion of the summed source signals seen in Figure 5.3. In contrast for the ipsilateral right ear, the spectra of both the correlated and decorrelated conditions are very similar

throughout the octave-band for both methods. Since the spectra are fairly similar in each ear, it may have been an interaural phase relationship and/or room effect that dictated the significant perception of VIS seen in the subjective testing (Section 4.2.2). It is also noticed that the main-channel spectrum follows a similar trend to that of the stereophonic conditions, which could have influenced an increase of perceived VIS for the monophonic case in the subjective testing.

5.3.2.2 8 kHz Band at +30°

As with 0° azimuth, the 8 kHz octave-band from +30° appears to have a spectral crossover between the main-channel and height-channel spectra around 9.5 kHz (Figure 5.8). In general, the frequencies where the height-channel is dominant over the main-channel (< 9.5 kHz) have a greater amplitude than when the main-channel is dominant (> 9.5 kHz), which is the case in both ears. This increase of energy around 6-9 kHz could be the reason for the height-channel localisation dominance seen in the absolute testing (Section 4.4.2).

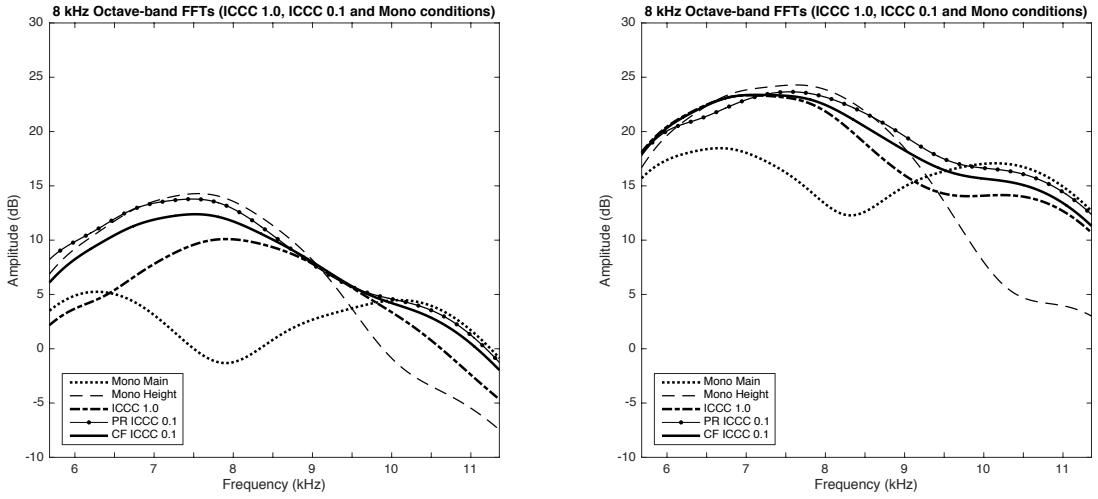

Figure 5.8 +30° azimuth 8 kHz octave-band average FFT for the main-channel only, height-channel only and both channels combined for the ICCC 1.0 and ICCC 0.1 conditions.
(Left – Contralateral Ear; Right – Ipsilateral Ear)

In terms of ICC effect, decorrelation appears to influence spectral boosts that differ between the two ears. However, as with the median plane, these spectral boosts are actually a 'filling in' of localisation cues (which is observed through comparison of the main-channel and height-

channel spectra). The effect of decorrelation on the degree of 'filling in' (i.e. the change between the correlated and decorrelated spectra) is particularly noticeable in the contralateral left ear below 9 kHz and above 10 kHz, while the ipsilateral right ear also has a slight change between 8-11 kHz. The observations here suggest that VIS could potentially be controlled by vertically decorrelating these specific frequency regions at ±30° azimuth. Furthermore, manipulation of these frequencies in binaural content may also have an effect on perceived VIS.

5.3.2.3 16 kHz Band at +30°

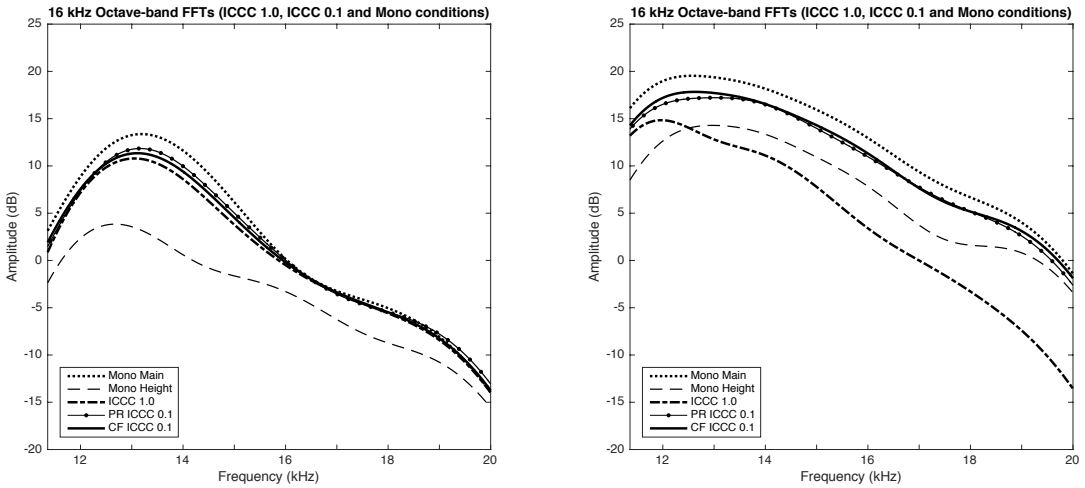

Figure 5.9 +30° azimuth 16 kHz octave-band average FFT for the main-channel only, height-channel only and both channels combined for the ICCC 1.0 and ICCC 0.1 conditions.
(Left – Contralateral Ear; Right – Ipsilateral Ear)

For the 16 kHz octave-band spectra in Figure 5.9 above, both the decorrelated and correlated conditions have very similar spectra in the contralateral left ear. On the other hand, the ipsilateral right ear demonstrates a general boost throughout the octave-band from decorrelation. A similar trend to the 0° spectra appears to occur, where the correlated signals cause frequencies to be 'cancelled out', with decorrelation reducing the amount of phase cancellation rather than boosting any frequencies. Again, it demonstrates the usefulness of decorrelation in applications where there is a risk of multiple signals interacting with one another at high frequencies. In general, the decorrelated condition follows a similar spectral trend to the main-channel only

spectra in both ears – this may have accounted for the increased perception of VIS with the monophonic condition in the subjective testing (Section 4.3.2.4).

5.3.3 +110° Azimuth Spectral Analysis

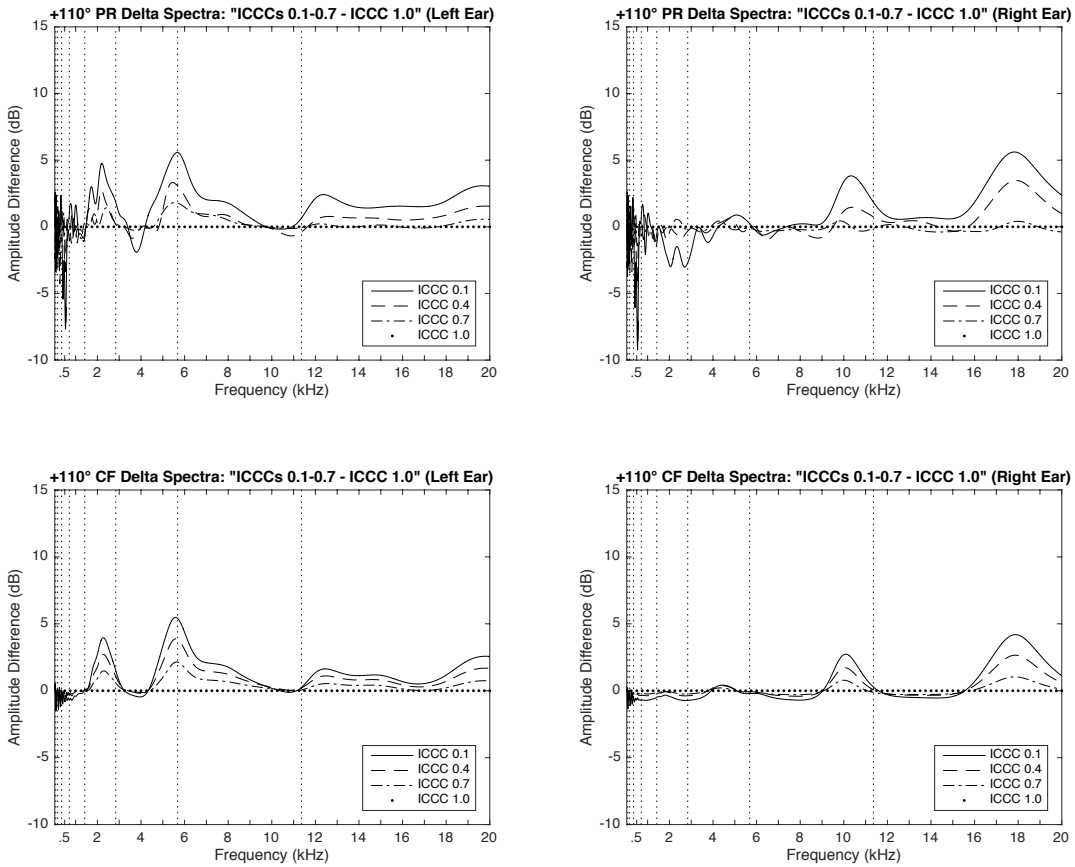

Figure 5.10 +110° azimuth delta spectra of the FFT frequency amplitude difference between the HRIR-convolved correlated stimulus (ICCC 1.0) and the decorrelated stimuli (ICCCs 0.1-0.7). The vertical dotted lines signify the band limits of each octave-band.
(Left – Contralateral Ear; Right – Ipsilateral Ear)

The +110° delta spectra in Figure 5.10 above show considerable differences of spectrum change between the left and right ear. In the contralateral left ear, spectral boosts from decorrelation can be seen at the 2 kHz octave-band, around 5-6 kHz (between the 4 kHz and 8 kHz octave-bands) and generally across the 16 kHz octave-band (peaks around 12 kHz and 19 kHz). Whereas, for the ipsilateral right ear, only two main boosts are seen at 10 kHz (8 kHz octave-

band) and 18 kHz (16 kHz octave-band). Given these observations, the inter-layer spectra and spectral changes from decorrelation have been investigated further for the 2 kHz, 4 kHz, 8 kHz and 16 kHz octave-bands below.

5.3.3.1 2 kHz Band at +110°

Observing the 2 kHz octave-band from +110° in Figure 5.11, there is no difference between the correlated and decorrelated spectra in the ipsilateral right ear. However, the contralateral left ear sees a general boost from decorrelation throughout the octave-band, with a shift towards the monophonic height-channel spectrum, suggesting a reduction of head-shadowing. This change in interaural level difference (ILD) is something that was initially discussed in the horizontal localisation section of the literature review (Section 1.1.1.2).

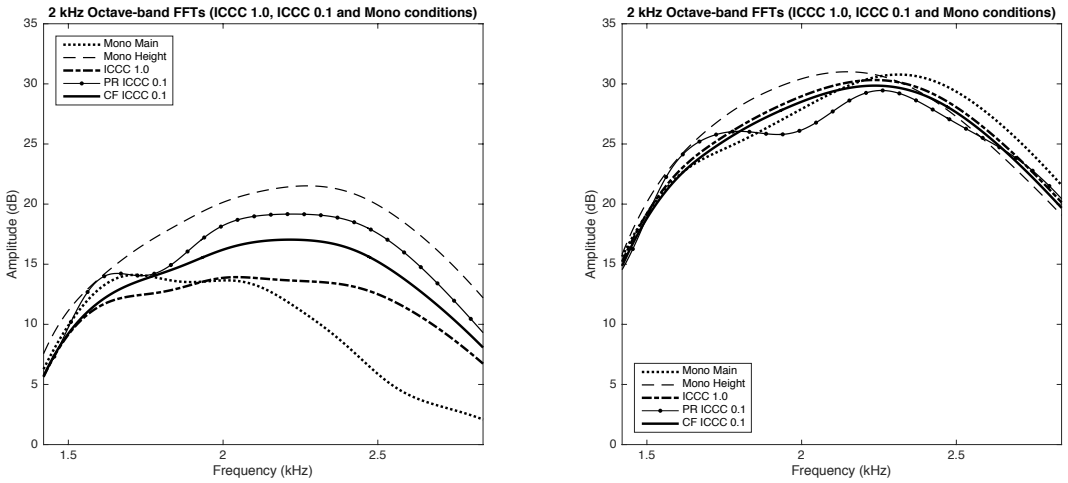

Figure 5.11 +110° azimuth 2 kHz octave-band average FFT for the main-channel only, height-channel only and both channels combined for the ICCC 1.0 and ICCC 0.1 conditions.
(Left – Contralateral Ear; Right – Ipsilateral Ear)

5.3.3.2 4 kHz Band at +110°

With the 4 kHz octave-band from +110° (Figure 5.12), there is very little difference between the spectra of the correlated and decorrelated conditions in the ipsilateral right ear, as with the 2 kHz band. However, decorrelation seems to cause a slight boost from around 4.5 kHz and above in the contralateral left ear. From Figures 5.11 and 5.12, it appears that a region between

2-4 kHz in the contralateral ear, and 3.5-5 kHz in the ipsilateral ear may relate to elevation at +110° – it is seen that the vertical stereophonic spectra (both correlated and decorrelated) are similar and positioned directly between that of the main- and height-channel spectra.

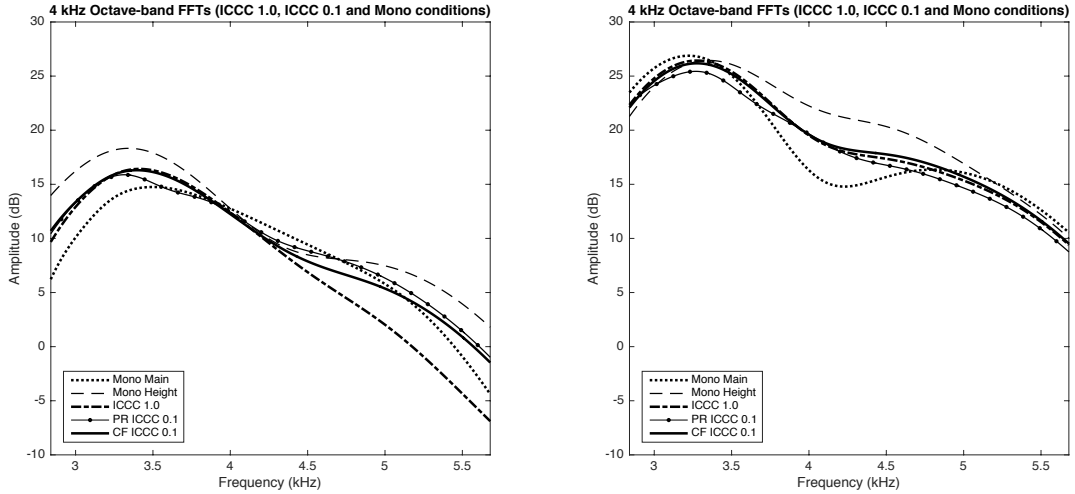

Figure 5.12 +110° azimuth 4 kHz octave-band average FFT for the main-channel only, height-channel only and both channels combined for the ICCC 1.0 and ICCC 0.1 conditions.
(Left – Contralateral Ear; Right – Ipsilateral Ear)

5.3.3.3 8 kHz Band at +110°

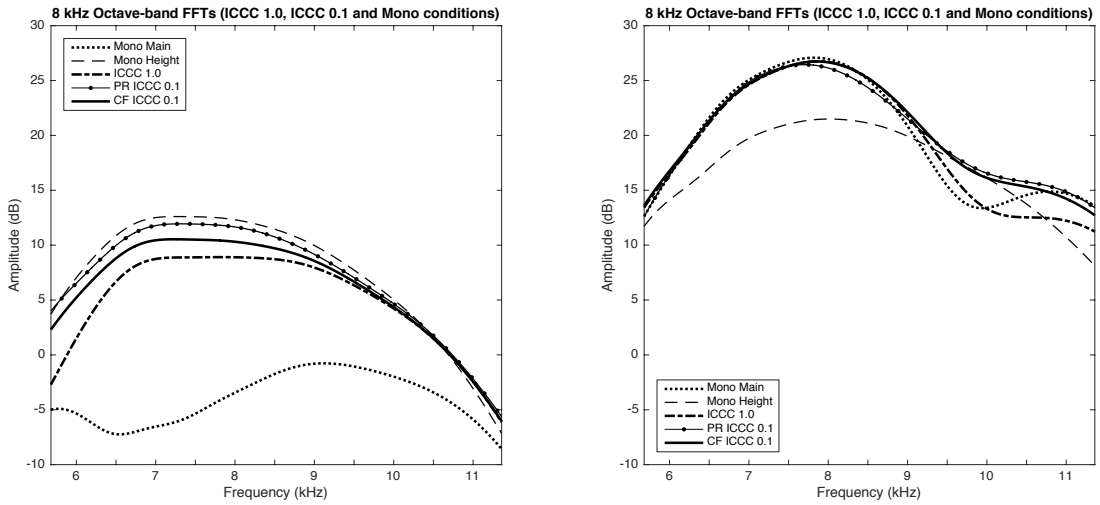

Figure 5.13 +110° azimuth 8 kHz octave-band average FFT for the main-channel only, height-channel only and both channels combined for the ICCC 1.0 and ICCC 0.1 conditions.
(Left – Contralateral Ear; Right – Ipsilateral Ear)

For the 8 kHz octave-band from +110° azimuth, Figure 5.13 above indicates a slight boost in the contralateral left ear below around 9.5 kHz, as well as a boost above 9.5 kHz in the ipsilateral right ear. In the right ear, it appears decorrelation reduces a slight notch that seems to be generated by a similar notch in the main-channel loudspeaker signal. This further indicates the effect vertical decorrelation can have by 'filling in' vertical localisation cues, as was observed for the 8 kHz octave-band at both 0° and +30° azimuth.

5.3.3.4 16 kHz Band at +110°

Looking at the 16 kHz octave-band spectra for +110° azimuth (Figure 5.14), it is seen that the contralateral left ear has a similar spectral response for both the correlated and decorrelated conditions. In the ipsilateral right ear, it appears that some frequency cancellation occurs for the correlated condition, as with the 16 kHz octave-band from 0° and +30° azimuth. In the case of +110°, a general spectral decrease is seen above 16 kHz for the correlated condition, which is then boosted by decorrelation of the signals (similar to the other azimuth angles).

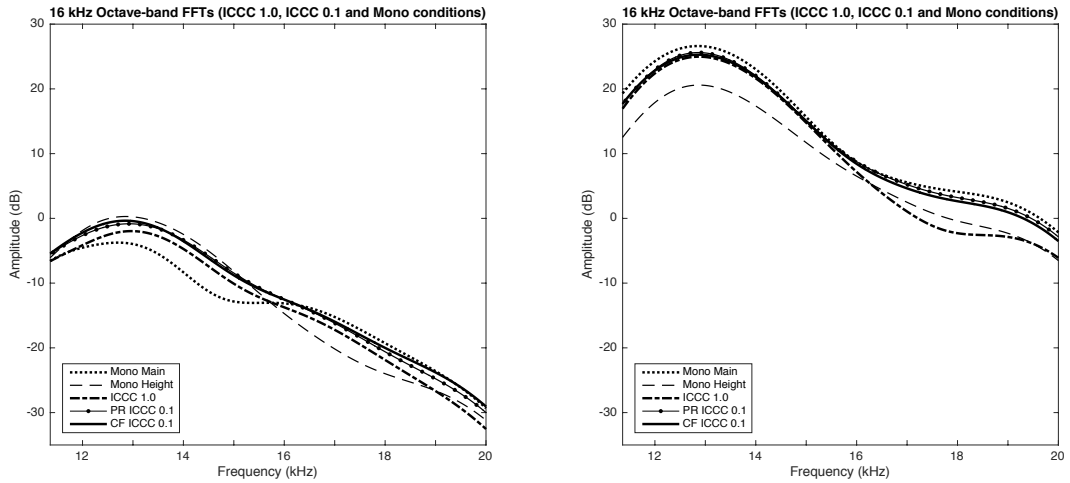

Figure 5.14 +110° azimuth 16 kHz octave-band average FFT for the main-channel only, height-channel only and both channels combined for the ICCC 1.0 and ICCC 0.1 conditions.
(Left – Contralateral Ear; Right – Ipsilateral Ear)

5.3.4 Discussion of Spectral Analysis Results

Taking into account the above spectral analysis of the HRIR-convolved stimuli signals, vertical interchannel decorrelation has a clear impact on spectral changes at higher frequency bands. At 0° azimuth (the median plane), variation of vertical ICC only affects the spectra for the 8 kHz and 16 kHz octave-bands. In the case of 8kHz, decorrelation appears to be related to a spectral boost around 10-11 kHz, where an increase of decorrelation seems to reduce the dominance of the height-channel, supporting the absolute testing results (Section 4.4.2). It is shown that this boost by decorrelation actually relates to the 'filling in' of a localisation notch cue from the height-channel position. On the other hand, the 16 kHz band shows some frequency cancellation at the ear with the correlated condition, where it is seen that decorrelation reduces the cancellation effect, rather than affecting vertical localisation cues. These observations suggest that energy in the 8 kHz band, particularly around 10-11 kHz, may be important for VIS perception in the median plane. This is in agreement with a widely accepted vertical localisation principle suggesting that accurate localisation in the median plane requires frequencies above 7 kHz (Roffler & Butler, 1968a). Further investigation of this frequency region may also demonstrate that its manipulation could control the degree of VIS within binaural audio. Previous research has already shown that manipulating the 8 kHz band in a broadband signal can control a source's perceived elevation (Chun et al., 2011), based on Blauert's boosted band hypothesis (Blauert, 1969/70).

With the 4 kHz octave-band at +30° and the 2 kHz and 4 kHz octave-bands at +110°, there is no spectral change in the ipsilateral right ear between the correlated and decorrelated conditions; whereas in the contralateral left ear, there are slight spectral boosts for each of these octave-bands. It was seen in the subjective results that the relationship between ICC and VIS was strongest for the 2 kHz and 4 kHz octave-bands at ±110° – this may have been influenced somewhat by the spectral boosts in the contralateral ear observed here, and is possibly related to the head-shadowing of the signal. The elevation of a source at +110° also seems to have an

impact on the spectra at the 2 kHz and 4 kHz octave-bands – a boost for the height-channel spectrum over the main-channel spectrum is seen between 1.5-4 kHz in the contralateral ear, and 3.5-5 kHz in the ipsilateral ear. Furthermore, the spectra of the stereophonic conditions are positioned between the main- and height-channel spectra in the same regions, suggesting that manipulation of these frequencies could potentially enhance vertical or binaural panning at +110° azimuth.

In contrast to the other high frequency bands, the 16 kHz octave-band has very little spectral change in the contralateral left ear for both the +30° and +110° azimuth angles. However, for the ipsilateral right ear, the 16 kHz octave-band displays some frequency loss with the correlation condition at both angles, presumably due to pinna filtering. This is similar to the effect seen in the median plane, and as with the 0° azimuth, vertical decorrelation of the 16 kHz octave-band causes the phase cancellation to reduce. The spectra of the 16 kHz band here demonstrate the benefit that decorrelation can have for reducing unwanted interaction between similar signals. It also implies that decorrelation could improve 3D panning between multiple loudspeakers, where the interaction of correlated signals may cause a similar frequency cancelling effect at the ears. However, decorrelation for this purpose may only be beneficial at higher frequencies – as Figure 5.3 demonstrates, greater degrees of phase decorrelation can cause visible spectral distortion at middle to lower frequencies.

For the 8 kHz octave-band at +30° and +110° azimuths, there are spectral boosts from decorrelation that differ in both the contralateral left ear and ipsilateral right ear. As with the 0° azimuth, it is seen that these boosts relate to the 'filling in' of vertical localisation notches, as dictated by the degree of decorrelation. At +30°, the boosts are observed below 8.5 kHz and around 10 kHz in the contralateral left ear, and between 8-11 kHz in the ipsilateral right ear. Whereas, at +110°, there is a general boost below 9.5 kHz in the contralateral left ear, and above 9.5 kHz in the ipsilateral right ear. Differences of spectra between the two ears indicate that an interaural relationship of spectral filtering may contribute to the perception of VIS, which

seems to be the case for other octave-bands as well. Given the interaural differences observed here, the following section examines a potential relationship between the interaural cross-correlation (IAC) and ICC, for both the HRIR-convolved and BRIR-convolved stimuli. It should also be noted that the spectral observations discussed above are specific to interchannel decorrelation between elevation angles of 0° and +30° – if other elevation angles were used, decorrelation could potentially affect different frequency points.

5.4 Interaural Cross-Correlation (IAC)

It is already established that the interaural cross-correlation coefficient (IACC) has a strong relationship with the perceived horizontal extent of a sound source (Hidaka et al., 1995). However, it is not currently known whether interaural differences also have a notable impact on the perception of vertical spatial extent. Judging from the spectral analysis in Section 5.3 above, there are clear interaural differences of spectral filtering at higher frequencies when the loudspeakers are off-axis. To follow this, consideration of IACC may further support an association between interaural differences and the perception of VIS. In particular, no spectral changes were evident for the 500 Hz or 1 kHz octave-bands from ICC change in Section 5.3, despite these octave-bands having a significant ICC effect in the subjective testing (Section 4.2.2).

Using the convolved stimuli described in Section 5.2, $IACC_{avg}$ values have been determined for each of the convolved conditions using Equations 5.7 and 5.8 below, taking the average of time-varying IACCs calculated over 50 ms-long windows, with 1 ms lag to account for interaural time delay (ITD) (Hidaka et al., 1995). Analysis of the $IACC_{avg}$ results has been separated into the anechoic HRIR-convolved stimuli (Section 5.4.1) and the BRIR-convolved stimuli which simulate the listening room (Section 5.4.2), in order to consider the impact of decorrelation on $IACC_{avg}$ both without and with room reflections.

$$IACF(\tau) = \frac{\int_{-\infty}^{\infty} x_{left}(t)x_{right}(t+\tau)dt}{\sqrt{\left[\int_{-\infty}^{\infty} x_{left}^2(t)dt\right]\left[\int_{-\infty}^{\infty} x_{right}^2(t)dt\right]}} \tag{5.7}$$

$$IACC = max|IACF(\tau)| \qquad \text{where } |\tau| \leq 1ms \tag{5.8}$$

5.4.1 IAC Results – HRIR-Convolved Stimuli

A summary of the $IACC_{avg}$ results for the anechoic HRIR-convolved stimuli can be seen in Table 5.1 below. The table features $IACC_{avg}$ values for the extreme interchannel cross-correlation (ICC) conditions of '1.0' and '0.1' (for both decorrelation methods) and the monophonic

stimuli used in the subjective experiments of Chapter 4. The $IACC_{avg}$ values are presented for each frequency band at the three azimuth angles (0°, +30° and +110°, assuming symmetry between the left and right directions). Given the stimuli were convolved with symmetrical anechoic HRIR impulses, results for the centre position of 0° (the median plane) display full correlation between the two ears for all stimuli ($IACC_{avg}$ = 1.0) – as the azimuth angle increases off-centre, the level of $IACC_{avg}$ begins to decrease for higher frequencies.

Table 5.1 Interaural Cross-Correlation Coefficient averages from 50 ms windows ($IACC_{avg}$). KEMAR Anechoic HRIR-convolved stimuli (with the main- and height-layers combined)

	Centre (0°)				**Front Right (+30°)**				**Rear Right (+110°)**			
	Mono	**1.0**	**0.1 (PR)**	**0.1 (CF)**	**Mono**	**1.0**	**0.1 (PR)**	**0.1 (CF)**	**Mono**	**1.0**	**0.1 (PR)**	**0.1 (CF)**
63 Hz	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98
125 Hz	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98
250 Hz	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.98	0.98	0.98	0.98
500 Hz	1.00	1.00	1.00	1.00	0.98	0.98	0.97	0.98	0.97	0.98	0.95	0.96
1 kHz	1.00	1.00	1.00	1.00	0.93	0.95	0.94	0.94	0.96	0.96	0.94	0.94
2 kHz	1.00	1.00	1.00	1.00	0.97	0.99	0.92	0.95	0.79	0.96	0.58	0.65
4 kHz	1.00	1.00	1.00	1.00	0.93	0.93	0.81	0.81	0.79	0.89	0.79	0.78
8 kHz	1.00	1.00	1.00	1.00	0.90	0.72	0.63	0.65	0.37	0.80	0.44	0.50
16 kHz	1.00	1.00	1.00	1.00	0.86	0.66	0.58	0.62	0.51	0.83	0.45	0.47
Broadband	1.00	1.00	1.00	1.00	0.91	0.90	0.79	0.83	0.48	0.73	0.46	0.63

Observing the effect of vertical ICC on $IACC_{avg}$ at +30° and +110°, a comparatively noticeable decrease of $IACC_{avg}$ (> 0.1) from ICCC 1.0 to 0.1 is seen for the 4 kHz octave-band and above at +30°, and for the 2 kHz band and above at +110°. That is, as correlation between the vertical pair of loudspeakers decreases, so does correlation between the ear-input signals, which is in line with the spectral changes observed in Section 5.3 above. Blau (2002) determines that the just noticeable difference of the IACC is 0.038, suggesting that these changes in IACC are perceivable. It is also noted from Table 5.1 that the monophonic conditions at +110° have a similar $IACC_{avg}$ to the decorrelated conditions, which may have accounted for the increased perception of VIS for the monophonic stimuli when both conditions were level-matched, as observed in the subjective results (Section 4.2.2). It is thought that the lower $IACC_{avg}$ of the monophonic condition seen here is caused by a greater head-shadowing effect at +110°; whereas the low $IACC_{avg}$ value for the decorrelated condition is likely controlled by the

decorrelated height-channel signal in the contralateral ear, as a result of head shadowing from main-channel – this hypothesis is discussed further in Section 5.4.1.1 below.

5.4.1.1 The Head-Shadowing Effect

The above $IACC_{avg}$ values imply that the head may play an important role in VIS perception off-axis, through the shadowing of the main-channel signal to the contralateral ear. This head-shadowing effect is clearly demonstrated in Table 5.2, where the interaural level differences (ILDs) of the HRIR-convolved stimuli are presented for each layer independently, as well as the correlated and decorrelated conditions. ILD has been calculated as the difference of RMS between the two ear input signals, taken as the average RMS of 50 ms windows over the signals' duration. To support the discussion, Table 5.3 also displays the $IACC_{avg}$ values for the main- and height-layers independently at both +30° and +110° azimuth.

Table 5.2 Interaural Level Differences (ILD) (dB) of the KEMAR HRIR-convolved stimuli for the main-layer, height-layer, ICCC 1.0, ICCC 0.1 (PR) and ICCC 0.1 (CF) at +30° and +110° azimuth positions.

	Front Right (+30°)					**Rear Right (+110°)**				
	Main-Layer	**Height-Layer**	**ICCC 1.0**	**ICCC 0.1 (PR)**	**ICCC 0.1 (CF)**	**Main-Layer**	**Height-Layer**	**ICCC 1.0**	**ICCC 0.1 (PR)**	**ICCC 0.1 (CF)**
63 Hz	-0.7	-0.6	-0.7	-0.7	-0.6	-1.0	-0.7	-0.8	-1.0	-0.8
125 Hz	-0.5	-0.5	-0.5	-0.6	-0.5	-0.9	-0.8	-0.8	-1.0	-0.8
250 Hz	-1.0	-0.7	-0.8	-0.8	-0.8	-1.6	-1.2	-1.4	-1.4	-1.4
500 Hz	-1.7	-1.2	-1.5	-1.5	-1.5	-2.7	-1.9	-2.3	-2.2	-2.3
1 kHz	-3.3	-2.9	-3.2	-3.1	-3.1	-3.9	-3.0	-3.6	-3.3	-3.5
2 kHz	-3.7	-2.9	-3.6	-3.2	-3.4	-8.1	-4.6	-7.5	-5.1	-5.9
4 kHz	-4.6	-3.8	-5.1	-4.5	-4.4	-5.4	-4.2	-5.0	-4.6	-4.8
8 kHz	-6.6	-5.1	-6.6	-5.2	-5.5	-12.8	-4.4	-7.9	-6.9	-7.2
16 kHz	-3.9	-5.6	-2.3	-3.7	-4.0	-14.4	-9.9	-13.0	-12.9	-12.5
Broadband	-3.8	-3.3	-3.9	-3.7	-3.6	-6.8	-4.2	-6.1	-4.9	-5.2

Table 5.3 Interaural Cross-Correlation Coefficient averages from 50 ms windows ($IACC_{avg}$) of the KEMAR anechoic HRIR-convolved stimuli Main-Layer vs. Height-Layer $IACC_{avg}$ at +30° and +110° azimuth positions

	Front Right (+30°)		**Rear Right (+110°)**	
	Main-Layer	**Height-Layer**	**Main-Layer**	**Height-Layer**
63 Hz	0.99	0.99	0.99	0.99
125 Hz	0.99	0.99	0.98	0.99
250 Hz	0.99	0.99	0.98	0.99
500 Hz	0.98	0.98	0.97	0.97
1 kHz	0.93	0.97	0.96	0.96
2 kHz	0.97	0.97	0.79	0.95
4 kHz	0.93	0.94	0.79	0.93
8 kHz	0.90	0.97	0.37	0.88
16 kHz	0.86	0.81	0.51	0.84
Broadband	0.91	0.91	0.48	0.89

Inspecting the difference between the main- and height-layer ILD values in Table 5.2, there is a slight decrease of ILD for the height-layer at +30° in most cases (i.e. an increase of level in the contralateral ear). However, at +110°, the difference of ILD between the main- and height-layers is clear for the 2kHz octave-band and above, demonstrating the head-shadowing effect. This is also reflected in the $IACC_{avg}$ calculations of the main- and height-layers at +110° azimuth (Table 5.3), where the interaural cross-correlation is noticeably higher for the height-channel than the main-channel with the 2 kHz octave-band and above. These results broadly agree with the bands at which the spectral differences occur in Section 5.3 above, however, they also suggest that the HRTF filtering might also have an effect on interaural cross-correlation, most notably at wider azimuth angles.

Given the head-shadowing of the main-channel signal at +110° azimuth, it is thought that the contralateral (far) ear is largely influenced by the height-channel signal, whereas the ipsilateral (near) ear is a sum of both the height- and the main-channel signals. This is particularly the case for the 2 kHz and 8 kHz octave-bands, where the attenuation of the main-layer signal in the contralateral ear is around 8-13 dB, whereas the height-layer signal is only attenuated by around 4-5 dB (Table 5.2). When correlated signals are presented from both the main- and height-layer loudspeakers (ICCC = 1.0), it results in relatively high correlation between both ears (as well as a decrease of ILD) – this can be seen in the $IACC_{avg}$ results of Table 5.1. However, when the two loudspeaker signals are decorrelated (ICCC = 0.1), the correlation between the two ears also decreases, given the height-channel dominance in the contralateral ear. This is the cause of the relationship between IACC and vertical decorrelation at higher frequencies seen in Table 5.1, and potentially contributes to the perception of VIS at wider azimuth angles. Although the results here are unable to confirm a direct relationship between IACC, ILD and VIS, there is certainly cause for further investigation. It may be found that the hearing system is able to assume an auditory source's vertical location and spread around the head based on interaural cues from frequency-dependent head-shadowing.

5.4.1.2 Comparison of Decorrelation Methods

In Table 5.1, differences of IACC between the two decorrelation methods under testing (Complementary Comb-Filtering (CF) and Phase Randomisation (PR)) are broadly the same across all frequency-bands for ICCC 0.1, indicating that they may operate in a similar manner. A strong similarity of perceived VIS was also evident between methods in the subjective testing (Section 4.2.2), as well as with the spectral analysis of higher frequencies in Section 5.3. Considering this, Figure 5.15 below displays the difference of FFT frequency amplitudes (within the 8 kHz octave-band) between the two outputs of each method for ICCC 0.1 (4096 FFT-points). Here it is seen that amplitude variations between the two channels are present for both methods. Given that the all-pass filters used with the PR approach are time-invariant, these amplitude differences are maintained throughout the signal, much like for the CF method. Although the differences for PR are lower in magnitude and less regular (more random) than the CF method, the combination of amplitude and phase variations may contribute to a greater perception of VIS by decorrelation, as was occasionally the case with the PR method in the subjective testing (Section 4.2.2.1).

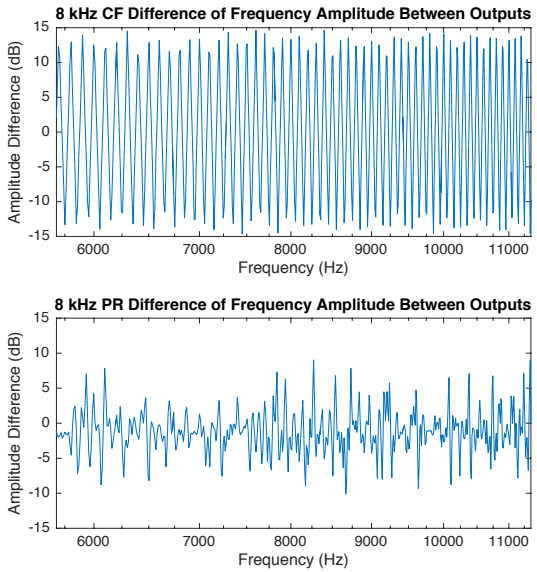

Figure 5.15 Difference of spectral magnitude between the two outputs of both decorrelation methods (8 kHz octave-band; Upper – Complementary Comb-Filtering; Lower – Phase Randomisation)

5.4.2 IAC Results – BRIR-Convolved Stimuli

The IACC and spectral analysis of the anechoic HRIR-convolved stimuli have indicated some trends with VIS for around the 2 kHz octave-band and above. However, the subjective results also revealed significant changes to VIS with the 500 Hz and 1 kHz octave-bands, particularly in the median plane (0° azimuth). Since there are no interaural differences or spectral filtering from the HRTF for these bands, it suggests that some room effect may have had an influence on the perception of VIS. In order to examine this, Table 5.4 below presents the $IACC_{avg}$ data of the BRIR-convolved listening room stimuli.

Table 5.4 Interaural Cross-Correlation Coefficient averages from 50 ms windows ($IACC_{avg}$) BRIR-convolved stimuli, captured using a Neumann KU 100 in the listening room

	Centre (0°)				Front Right (+30°)				Rear Right (+110°)			
	Mono	1.0	0.1 (PR)	0.1 (CF)	Mono	1.0	0.1 (PR)	0.1 (CF)	Mono	1.0	0.1 (PR)	0.1 (CF)
63 Hz	1.00	1.00	1.00	1.00	0.98	0.98	0.99	0.95	0.98	0.98	0.98	0.98
125 Hz	0.99	1.00	1.00	0.99	0.94	0.96	0.94	0.93	0.91	0.92	0.85	0.92
250 Hz	0.97	0.97	0.98	0.97	0.92	0.90	0.92	0.90	0.86	0.88	0.92	0.83
500 Hz	0.81	0.89	0.77	0.83	0.57	0.64	0.70	0.61	0.76	0.78	0.70	0.73
1 kHz	0.83	0.89	0.83	0.82	0.68	0.74	0.72	0.64	0.58	0.62	0.51	0.53
2 kHz	0.84	0.89	0.85	0.86	0.73	0.81	0.71	0.68	0.28	0.36	0.27	0.28
4 kHz	0.88	0.89	0.87	0.87	0.73	0.80	0.75	0.73	0.40	0.34	0.24	0.28
8 kHz	0.66	0.74	0.68	0.70	0.37	0.47	0.34	0.39	0.21	0.22	0.22	0.20
16 kHz	0.86	0.85	0.81	0.80	0.54	0.28	0.30	0.28	0.25	0.19	0.15	0.20
Broadband	0.95	0.97	0.96	0.94	0.79	0.81	0.81	0.77	0.70	0.73	0.73	0.71

Looking at the 0° results in Table 5.4, it is evident that the $IACC_{avg}$ begins to decrease around the 500 Hz octave-band and above ($IACC_{avg} < \sim 0.9$), which is in line with the VIS change seen in the subjective testing results (Section 4.2.2). Specifically, for the 500 Hz and 1 kHz octave-bands, it is observed that the ICCC 0.1 decorrelated conditions have a slight, yet noticeable, decrease of $IACC_{avg}$ in comparison to the ICCC 1.0 correlated condition (> 0.05). This suggests a potential association between vertical ICC and $IACC_{avg}$ at these bands, which could be related to an increased decorrelation of reflections when presented in a non-anechoic environment. There is also a decrease in $IACC_{avg}$ for the 500Hz and 1 kHz monophonic samples compared to ICCC 1.0 (similar to that of ICCC 0.1) – this further corresponds with the subjective testing results, where the monophonic and ICCC 0.1 stimuli had a similar perceived VIS, with both

significantly greater than the ICCC 1.0 condition. The lower $IACC_{avg}$ seen for the monophonic and decorrelation conditions here could have invoked a greater VIS, where a lower $IACC_{avg}$ may increase general spaciousness (both horizontally and vertically).

Looking at the +30° and +110° $IACC_{avg}$ results in Table 5.4, as with the anechoic HRIR-convolved stimuli, there is a decrease of IACC at +110° compared to +30°, presumably due to the head-shadowing effect discussed in Section 5.4.1.1. Additionally, as would be expected, the introduction of room reflections decreases the level of IACC considerably across all stimuli from around the 500 Hz octave-band and above – this is in contrast with the anechoic HRIR-convolved stimuli, where the two ear signals were almost correlated below the 4 kHz octave-band for +30° azimuth, and below the 2 kHz octave-band at +110°. This further supports the notion that room reflections may have influenced VIS perception for the 500 Hz and 1 kHz octave-bands.

5.4.3 Discussion of IAC Results

Differences between the two sets of IACC results with the HRIR- and BRIR-convolved stimuli imply that the perception of VIS might be dependent on both head shadowing and the effect of the room. At frequencies of around the 2 kHz octave-band and above, the anechoic HRIR-convolved stimuli results tend to align with both the subjective results and spectral analysis at +30° and +110°, where a decrease of vertical ICCC appears to decrease IACC. Considering the well-known relationship between IACC and horizontal image spread (HIS) (Zotter & Frank, 2013; Hidaka et al., 1995), this result initially suggests that vertical decorrelation might generate an increase of HIS as well as VIS for the high frequency bands. Reversely, it could also be hypothesised that IACC may be a useful perceptual cue for the perception of VIS (in addition to HIS) at off-axis loudspeaker positions – this requires further study to verify.

On the other hand, the BRIR IACC results in Section 5.4.2 for the 500 Hz and 1 kHz octave-bands indicate a potential relationship between VIS and a decrease in $IACC_{avg}$ in the median plane, presumably due to the influence of room reflections. Statistical analysis of the subjective

data also indicated that the correlation between ICCC and VIS with the 500 Hz and 1 kHz octave-bands was strongest in the median plane (Section 4.2.2.3). The decrease of $IACC_{avg}$ in the median plane is likely due to the effect of multiple room reflections summing at the ears from different directions, given that the ears are symmetrical, i.e., the same direct sound signal is received at both ears. Furthermore, it is thought that early reflections may have caused the decrease of $IACC_{avg}$ seen with the monophonic condition, due to greater early reflection energy for the monophonic condition than the vertical stereophonic conditions, when both are level-matched – this is explored in Section 5.5 below.

5.5 Ratio of Early Reflection Energy to Direct Energy (ER/D)

5.5.1 ER/D at 0° Azimuth

Taking into account the $IACC_{avg}$ results for the BRIR-convolved stimuli in Section 5.4.2 above, a relationship has been suggested between room reflections and the perception of VIS in the median plane. This is particularly the case for the 500 Hz and 1 kHz octave-bands, where both the monophonic and vertically decorrelated conditions had an increase of VIS. To investigate this further, Table 5.5 below displays the ratio of early reflection energy (2.5 – 80 ms) to direct sound energy (< 2.5 ms) (ER/D ratio) at 0° azimuth, calculated for both the main-layer loudspeaker BRIR (the monophonic condition) and the summed result of the time-aligned main- and height-layer BRIRs (vertical stereophony).

Table 5.5 Ratio of Early Reflection Energy (2.5-80 ms) to Direct Energy (< 2.5 ms) (ER/D) (dB) BRIRs at +0°, captured using a Neumann KU 100 in the listening room.

Centre (0°)	Left Ear		Right Ear	
	Main-Layer Only	Layers Summed	Main-Layer Only	Layers Summed
63 Hz	29.7	31.7	29.3	31.2
125 Hz	8.0	7.4	8.8	7.9
250 Hz	3.0	0.5	3.7	1.1
500 Hz	-1.5	-5.6	0.5	-2.7
1 kHz	-3.5	-6.8	-3.5	-7.3
2 kHz	-6.8	-8.6	-6.6	-8.2
4 kHz	-10.3	-12.5	-10.8	-12.9
8 kHz	-2.3	-4.8	-5.5	-6.3
16 kHz	-13.5	-12.4	-14.6	-14.0
Broadband	-6.8	-7.8	-6.2	-7.5

Although the listening room has relatively low levels of reflective energy, due to a short decay time (RT = 0.25 s) and acoustic treatment compliant with ITU-R BS.1116 (ITU-R, 2015a), there remains a clear increase of ER/D at 500 Hz and 1 kHz for the main-layer only condition in Table 5.5 (with around 3-4 dB more early reflection energy). It seems a reason for this could be that, when the time-aligned signals of a vertical stereophonic loudspeaker pair are summed at the ear, the direct sound level doubles; however, the relative level of early reflections remains low due to differing lengths of reflection paths and a staggering of arrival. The first floor reflection from the main-layer loudspeaker arrives at the ears around 3.5 ms after the direct sound,

whereas, the first floor reflection from the height-layer arrives around 5.5 ms after the direct sound. In order to level match the SPL LAeq of the monophonic stimuli with the doubling of direct sound in the vertical stereophonic samples, the overall output level of the main-layer loudspeaker would have been raised to match that of the stereophonic condition, thus increasing both the direct sound energy *and* reflection energy (to a greater level than the average of the staggered delays). This is illustrated in Figure 5.16, where impulses that represent the direct signal and first floor reflections are drawn for both the vertical stereophonic condition and the level-matched monophonic condition.

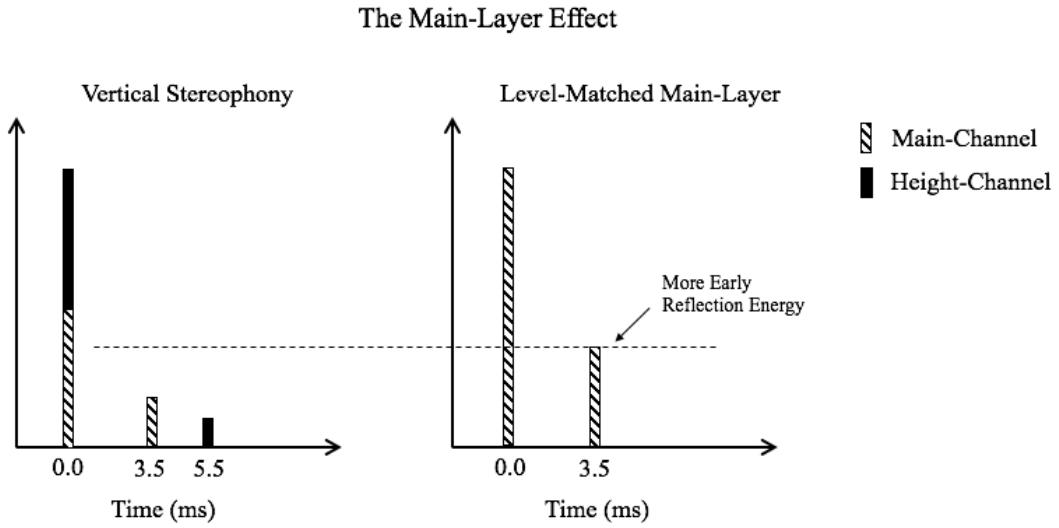

Figure 5.16 Illustration of the 'Main-Layer Effect', which suggests that the summed main- and height-layer signals have less early reflective energy than a level-matched monophonic main-layer only signal.

Further demonstration of the 'main-layer effect' with real room impulses can be seen in Figure 5.17 below. This figure displays the 500 Hz and 1 kHz octave-bands, comparing the main-layer only condition and the condition of the summed layers, where both signals have been RMS level-matched (to replicate the SPL LAeq level-matching of stimuli in the subjective testing of Chapter 4). Greater early reflective energy is clearly seen for the monophonic condition compared to the summed condition, as was hypothesised above. It is possible that, when comparing the monophonic main-layer stimuli directly against the correlated stereophonic stimuli (ICCC

1.0) under controlled conditions, this increase in early reflection energy may be perceivable and contribute to an increase of VIS, possibly by decreasing IACC.

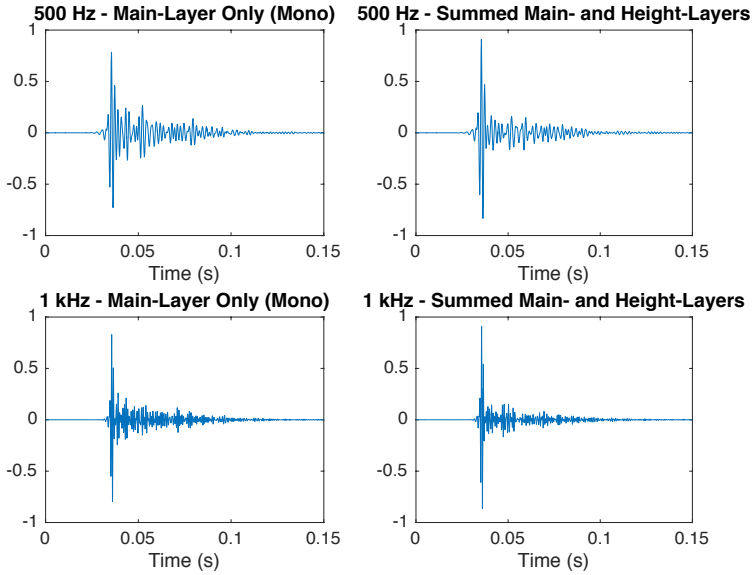

Figure 5.17 'The Main-Layer Effect' – 500 Hz and 1 kHz octave-band filtered binaural room impulse responses, with RMS level-matching between the mono (main-layer only) and summed conditions.
Left – Main-layer only impulse response (monophonic)
Right – Summed main- and height-layer impulse responses

In the subjective testing, the absolute VIS results for 500 Hz and 1 kHz at 0° showed that the ICCC 0.1 and monophonic conditions had a slight increase of spread in opposite directions compared to the coherent ICCC 1.0 condition (Section 4.3.2). It is possible that the decrease in $IACC_{avg}$ for these stimuli (Table 5.5) influenced an ambiguous perception of greater extent in all directions, given the weakness and inaccuracy of vertical localisation at these frequencies. This is similar to the hypothesis proposed in Section 5.4.1.1 above, where the head-shadowing of a monophonic stimulus at high frequencies produces a low $IACC_{avg}$, potentially leading to a general increase of both VIS and HIS. An illustration of possible perception of VIS by IACC is presented in Figure 5.18 below. The left of the image displays the typical understanding of the IACC effect on horizontal spread, whereas the right shows how a decrease of IACC for the 500 Hz and 1 kHz octave-bands might have been perceived.

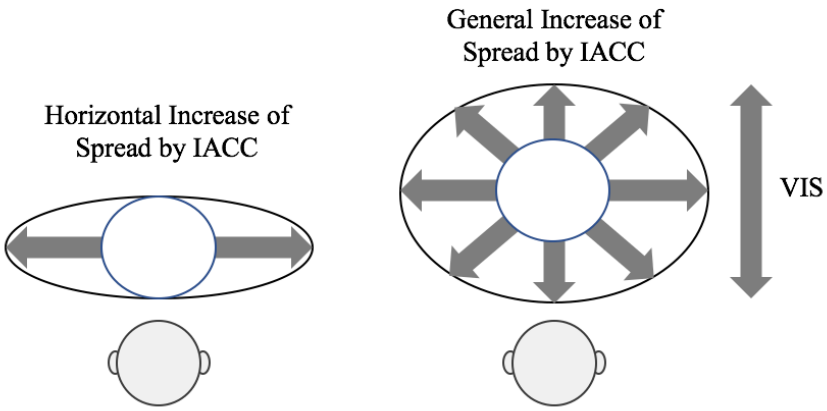

Figure 5.18 The potential perception of vertical image spread (VIS) change from a decrease in IACC.

The results in Table 5.5 also demonstrate that as frequency decreases, the ER/D ratio tends to increase – and for the 63 Hz, 125 Hz and 250 Hz octave-bands, the reflective energy is reported as greater than the direct sound energy. This room effect may be the reason for the inherent broad perception of lower frequencies in space, due to less distinction between direct and secondary sound signals in a reverberant room; however, it should also be noted that this effect could be exclusive to the impulse responses of the specific listening room in question. Although the effect may be room-dependent, it could potentially cause an increased perception of listener envelopment (LEV) for 2D main-layer only content, when SPL level-matched and compared directly against 3D stimuli – this is something to be explored further.

5.5.2 ER/D at +30° and +110° Azimuth

Looking at the same ER/D calculations for the +30° and +110° positions (Tables 5.6 and 5.7, respectively), it can be seen that ER/D increases for the contralateral left ear in comparison to the ER/D for the ipsilateral right ear – this is likely due to a decrease of direct energy in the contralateral left ear from the head-shadowing effect described in Section 5.4.1.1. The difference of ratio values between the two ears becomes noticeably greater (> 10 dB) for the 4 kHz octave-band and above at +30°, and the 2 kHz octave-band and above at +110°, which

corresponds with the decrease of IACC and ILD from head-shadowing at these frequencies (as discussed in Section 5.4 above).

Table 5.6 Ratio of Early Reflection Energy (2.5-80 ms) to Direct Energy (< 2.5 ms) (ER/D) (dB). BRIRs at +30°, captured using a Neumann KU 100 in the listening room.

Front Right (+30°)	Contralateral Left Ear		Ipsilateral Right Ear	
	Main-Layer Only	Layers Summed	Main-Layer Only	Layers Summed
63 Hz	29.1	31.2	28.7	29.3
125 Hz	9.9	8.7	8.8	7.3
250 Hz	2.1	-0.6	2.7	-1.0
500 Hz	1.7	-0.4	-2.5	-4.1
1 kHz	0.9	-1.1	-6.1	-7.4
2 kHz	-3.3	-5.8	-9.1	-11.6
4 kHz	-3.1	-5.8	-14.8	-17.5
8 kHz	5.7	4.3	-7.3	-9.2
16 kHz	-2.4	4.5	-15.1	-14.1
Broadband	0.0	-1.2	-9.2	-10.7

Table 5.7 Ratio of Early Reflection Energy (2.5-80 ms) to Direct Energy (< 2.5 ms) (ER/D) (dB). BRIRs at +110°, captured using a Neumann KU 100 in the listening room.

Rear Right (+110°)	Contralateral Left Ear		Ipsilateral Right Ear	
	Main-Layer Only	Layers Summed	Main-Layer Only	Layers Summed
63 Hz	23.9	25.9	20.8	21.9
125 Hz	12.5	6.3	8.0	4.3
250 Hz	2.4	0.0	1.5	-0.6
500 Hz	1.1	-1.1	-4.5	-6.3
1 kHz	-0.1	-1.0	-6.4	-8.5
2 kHz	7.3	3.2	-7.1	-9.7
4 kHz	1.0	2.9	-11.1	-10.5
8 kHz	11.7	9.3	-12.8	-12.0
16 kHz	9.8	13.5	-13.7	-11.6
Broadband	3.0	1.6	-10.0	-10.2

As with the 0° position, the main-layer only condition at +30° and +110° also shows some increase of early reflective energy in relation to direct sound, compared to when the layers are combined. From these azimuths, the main-layer effect is most prominent for the 250 Hz, 2 kHz and 4 kHz bands at +30° (~2-4 dB more early reflection energy in both ears), and 125 Hz and 2 kHz at +110° (~3-6 dB more energy). For both 250 Hz at +30° and 125 Hz at +110°, this increase of main-layer early reflection energy coincides with an increased perception of VIS for the monophonic stimuli at these bands in the subjective results (Section 4.2.2).

5.5.3 Discussion of ER/D Results

Although the above ER/D values are specific to the acoustic condition of the listening room used, the results suggest that there may be a relationship between the perception of VIS and the level of early reflection energy in an enclosed space. This is particularly the case for lower frequencies when there are weak or no vertical localisation cues (Roffler & Butler, 1968a). In the case of this investigation, it appears that the early reflection energy may be dictated by the arrival of first floor reflections. It is possible that an increase of early reflection energy has a direct impact on decreasing $IACC_{avg}$, generating a greater sense of source extent both horizontally and vertically. During the listening tests described in Chapter 4, subjects were asked to ignore changes to the HIS of a source and only focus on the VIS; therefore, it is not known whether changes to HIS were also perceived. Alternatively, the hearing mechanism could potentially determine the phase difference between the initial direct sound and the arrival of the first reflection from each source, allowing the brain to interpret each discrete source's spatial position based on the subconscious visual inspection and spatial mapping of a room. It is understood that time differences for frequencies below around 1.5 kHz are interpreted on the fine phase structure of a signal (Blauert, 1997), which may have aided the cognition of signal directivity for the continuous low frequency noise sources.

Furthermore, responses in a previous research investigation suggest that a single ceiling reflection can increase the perceived VIS of an auditory event (Robotham et al., 2016), indicating the effect a room's acoustic can have on vertical spatial perception. From the absolute VIS results in the subjective testing (Figure 4.6 of Section 4.2.2), it can be seen that there is a downward extension of VIS for the 500 Hz and 1 kHz monophonic conditions in the median plane. This could be related to an increase of floor reflection energy from the 'main-layer effect', as discussed above. Similarly, the broadband monophonic stimulus also displays some downward extension at both 0° and ±30° azimuth, would could be down to a similar reason.

5.6 Conclusion

This chapter describes objective analysis of binauralised stimuli from the subjective listening tests detailed in Chapter 4. Two sets of binauralised stimuli were created: 1) the stimuli signals convolved with anechoic head-related impulse responses (HRIRs) from the MIT KEMAR database; and 2) the stimuli signals convolved with binaural room impulse responses (BRIRs) captured in the listening room. Firstly, the HRIR-convolved stimuli were analysed spectrally, in order to observe the HRTF effect from vertical interchannel decorrelation. Both sets of binauralised stimuli were then analysed for interaural cross-correlation (IAC). Lastly, the raw BRIRs from the listening room were analysed to observe the ratio of early reflection to direct energy (ER/D).

The main findings of the objective analysis are as follows, based on vertical decorrelation between a main-layer loudspeaker at 0° and a height-layer loudspeaker elevated by +30°:

- For the 500 Hz and 1 kHz bands, VIS appears to be influenced by room reflections, as demonstrated by comparing the HRIR- and BRIR-convolved stimuli, i.e., changes were seen with the BRIR stimuli that were not present for the HRIR stimuli.
- At higher frequencies, filtering from the head-related transfer function (HRTF) seems to have the most influence on VIS perception (either through head-shadowing or pinna filtering). This was observed with the HRIR-convolved stimuli and varies depending on the azimuth angle of incidence, as follows:
    - In the median plane (0° azimuth), spectral changes that coincide with decorrelation occur in the 8 kHz and 16 kHz octave-bands.
    - From ±30° azimuth, the contralateral ear input signal is filtered in the 4 kHz and 8 kHz bands, while the ipsilateral ear input signal is filtered in the 8 kHz and 16 kHz bands.
    - At ±110° azimuth, filtering occurs in the 2-16 kHz octave-bands for the contralateral ear, and in the 8 kHz and 16 kHz bands with the ipsilateral ear.

  - With the 8 kHz band, vertical decorrelation appears to 'fill in' the vertical localisation notch cues, which seems to be a contributor to VIS perception.
  - For the 16 kHz band, a phase cancelling effect occurs when the vertical signals are correlated (ICCC = 1.0), with decorrelation reducing the degree of phase cancellation at the ears.
- For the 500 Hz and 1 kHz bands, there is a trend between IACC and VIS at 0° (seen with the BRIR-convolved stimuli), possibly from a greater decorrelation of reflections.
- For the higher frequency bands, as the azimuth angle increases, the IACC of the monophonic main-layer signal decreases due to greater head-shadowing (demonstrated with both the HRIR- and BRIR-convolved stimuli).
- When a coherent height-channel signal is introduced above the main-layer at +110°, IACC increases due to more direct signal in the contralateral ear. As ICC is decreased, IACC also decreases, which may contribute to the perception of VIS.
- Analysing the BRIRs, it is seen that the monophonic main-layer signal had greater early reflection energy than vertical stereophonic signals, potentially causing an increase of VIS for the monophonic condition – this has been called the 'main-layer effect'.
- This 'main-layer effect' is most apparent for the 500 Hz and 1 kHz bands in the median plane, which could also be linked to the decrease of IACC seen for these bands.

These objective results provide further insights into the perception of vertical interchannel decorrelation. Taking both the subjective results and objective analysis into account, it is clear that vertical image spread (VIS) change is mostly influenced by higher frequencies. In particular, spectral cues in the 8 kHz octave-band demonstrate a strong association with VIS perception. From these results, it is assumed that vertical interchannel decorrelation of low frequencies provides little contribution towards increasing VIS. Considering this, it is possible that vertical decorrelation may control / increase VIS with the exclusion of lower frequencies all together. To explore this further, Chapter 6 presents an experiment where complex stimuli have been 'high-pass decorrelated'.

6 HIGH-PASS FILTERED VERTICAL INTERCHANNEL DECORRELATION OF COMPLEX SOURCES

This chapter describes a two-part subjective experiment investigating the perceptual effect of vertically decorrelating just higher frequencies, looking at both the vertical image spread (VIS) and tonal quality (TQ) of decorrelated stimuli. In Chapters 4 and 5, it was found that a significant increase of VIS occurs when decorrelating pink noise octave-bands with centre frequencies of 500 Hz and above – that is, where VIS increased as the interchannel cross-correlation (ICC) decreased. Since these VIS changes were most apparent at higher frequencies, it is of interest to determine whether vertically decorrelating a high-pass filtered signal has a similar effect to decorrelating the broadband signal, i.e., where high frequencies are decorrelated between a main- and height-layer loudspeaker, while low frequencies are routed monophonically to the main-layer loudspeaker only. In particular, Chapter 5 demonstrated that potential spectral cues from vertical decorrelation feature in the 8-16 kHz octave-bands at 0° azimuth, the 4-16 kHz octave-bands at ±30° azimuth and the 2-16 kHz octave-bands at ±110°. If these spectral cues are dominant for VIS perception, then it may not be necessary to decorrelate lower frequencies at all. Furthermore, considering the application of 2D-to-3D upmixing, it is of interest to observe the impact vertical decorrelation has on perceived TQ – for example, high-pass decorrelation may reduce phase cancellation at lower frequencies, resulting in a more accurate representation of the source signal.

From these considerations, the main research questions for this investigation were as follows:

- Does vertical interchannel decorrelation of high-pass frequencies have a comparable effect on VIS to broadband decorrelation?
- At which high-pass frequency cut-off does VIS reach maximum extent?
- What impact does vertical decorrelation have on TQ?
- Are the subjective perceptions of VIS and TQ source-dependent?

In order to answer the above questions, stimuli were decorrelated using the all-pass filter phase randomisation method that featured in Chapters 4 and 5. Decorrelation was performed with varying high-pass cut-off frequencies, where only higher frequencies were decorrelated between a vertically-arranged loudspeaker pair, and the lower frequencies are routed to the main-layer loudspeaker only. To provide a practical context, the stimuli under testing are ambient complex sound sources, in addition to broadband pink noise – and the same three azimuth angles as Chapter 4 were assessed (0°, ±30° and ±110°).

6.1 Experimental Hypotheses

From the results of the previous subjective experiments in Chapter 4, it is hypothesised that decorrelation of frequencies below the 500 Hz octave-band in a broadband signal will provide little increase to the perception of vertical image spread (VIS). In the absolute testing, it was found that the lower frequency bands (centre frequencies: 63-250 Hz) were perceived as inherently broad for the monophonic condition, with no significant increase of VIS following decorrelation. Significant changes to the absolute VIS upper and lower image boundaries were only seen for octave-bands above 500 Hz. Similarly, in the relative VIS testing results, a direct relationship between the interchannel cross-correlation (ICC) and VIS was revealed for the 500 Hz octave-band and above. The strongest relationship was seen for the 8 kHz octave-band at ±30° azimuth, which is also known to be an important band for vertical localisation from the front (Roffler & Butler, 1968a; Hebrank & Wright, 1974). Given the strength of the relationship around 8 kHz, it is also hypothesised that only decorrelating the 8 kHz octave-band and above in a broadband signal will produce a perceivable increase of VIS from the monophonic condition, providing the source signal has adequate energy within this region. For the 500 Hz and 1 kHz octave-bands, the relationship between ICC and VIS seems to be related to early reflections in the listening room. Considering this, it is further hypothesised that decorrelation of frequencies down to these bands will also contribute to an increased perception of VIS; however, it is thought that there will be no significant difference of VIS between broadband decorrelation and high-pass decorrelation of only the 500 Hz octave-band and above.

In terms of tonal quality (TQ), it was noticed in Chapter 5 that the all-pass filter decorrelation method generates greater distortion at lower frequencies, when the two decorrelated signals are summed together (as would be the case at the ear input). This is presumably due to phase cancellation from greater degrees of phase variation between the two channels, and was particularly noticeable for the 500 Hz octave-band and below. From the above observation, it is hypothesised that the decorrelation of these lower frequencies will have a detrimental effect on

perceived TQ, when compared against an unprocessed monophonic source. Considering this and the VIS hypotheses above, it is thought that high-pass decorrelation may increase VIS, whilst having reduced impact on the perceived TQ degradation, which would make for a desirable scenario. Furthermore, the phase distortions might only be perceivable for certain source types (e.g. broadband pink noise); therefore, various complex stimuli have been assessed during this part of the study, in order to assess decorrelation in a practical context.

6.2 Experimental Design

6.2.1 Physical Setup

The loudspeaker format used for the current subjective testing is the same as that in Chapter 4, based on Auro-3D 9.1 with the addition of a centre height-channel (Figure 6.1) (Auro Technologies, 2015a). As with the octave-band pink noise experiment, all stimuli were presented through vertically arranged loudspeaker pairs at three azimuth angles to the listener: 0°, ±30° and ±110°. This resulted in the use of ten Genelec 8040A loudspeakers (Frequency response: 48 Hz – 20 kHz (±2 dB)) during testing. The five main-layer loudspeakers were positioned at a distance of 2 m from the listener, with the height-layer loudspeakers positioned directly above at an elevation angle of +30° (vertically spaced by 1.15 m).

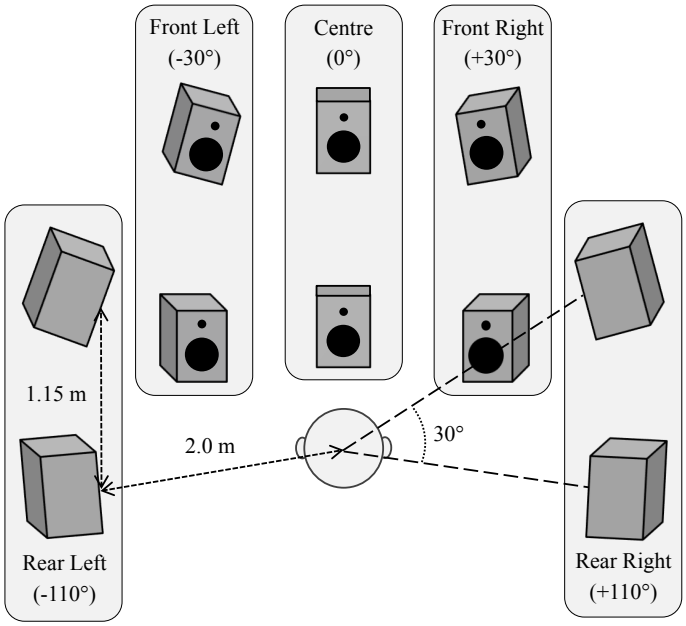

Figure 6.1 Physical loudspeaker setup used during testing (based on Auro-3D 9.1 (Auro Technologies, 2015a) with an additional Centre height-channel). Five main-layer loudspeakers positioned 2 m from the listener at ear height with azimuth angles of 0°, ±30° and ±110°. Five upper height-layer loudspeakers elevated directly above its main-layer pair by +30° to the listener.

Testing was conducted at the University of Huddersfield in a critical listening room that fulfils the specification of ITU-R BS.1116-3 (ITU-R, 2015a) (6.2m x 5.6m x 3.8m; RT = 0.25 s;

NR 12). Time and level alignment were applied between the two loudspeaker layers, to compensate for interlayer difference of signal arrival at the listening position. An acoustically transparent curtain was also used to obscure the loudspeakers from view, so as to avoid visual bias during testing.

6.2.2 Stimuli Creation

The aim of the current experiment is to look at vertically decorrelated higher frequencies only, by varying the cut-off frequency of high-pass decorrelation. It is of importance to observe the effect in a practical context, therefore, both noise and ambient complex stimuli were assessed during the experimentation. The source signals comprised broadband Pink Noise and five anechoic samples: Male Speech, Cello, Acoustic Guitar, a Drumkit and a String Quartet. Each of the anechoic source signals were chosen for their unique characteristics, in order to assess a wide variety of complex signals. The Drumkit sample was used for its transient and repetitive nature, which also covers a broad range of frequencies (kick, snare and high-hat). Male Speech is also relatively broadband and is often used for detecting unwanted artefacts within audio processing, due our inherent familiarity and sensitivity to the human voice. The plucked Acoustic Guitar provides both transient and melodic characteristics, where multiple notes are played simultaneously with varying dynamics. In contrast, the Cello sample has a relatively slow attack with rich harmonic overtones for each note and a moving melodic line. Lastly, the String Quartet provides an ensemble performance with similar characteristics to the Cello, however, the instruments cover a broader range of high frequencies (cello, viola and violin), with multiple parts performing different melodies simultaneously.

In order to generate ambient signals, as would be the case in a practical upmixing application, the same artificial reverb (RT = 2 s) was applied to the Male Speech and musical sources (but not Pink Noise). The reverb was applied using the ‘ReaVerb’ plug-in in Reaper, and the dry signal was removed from the output of the reverb to reproduce the ambient wet signal only. The waveform and frequency response of the reverb used can be seen in Figure 6.2 – these were

obtained by passing a unit sample through the reverb plug-in to generate the reverb's impulse response. Artificial reverb was chosen over convolution with a real room impulse response so that more control could be had over the frequency response of the output. It was important to maintain high frequency energy as much as possible, in order to fully assess the effect of cut-off frequency with high-pass decorrelation. In Figure 6.2, the FFT for the artificial reverb is compared against that of an impulse response captured beyond the critical distance in St Paul's Concert Hall at the University of Huddersfield (i.e. capturing mostly ambience with little direct sound) – it is clearly seen that the artificial reverb has more high frequency energy above 1 kHz, as well as more low frequency energy below 60 Hz.

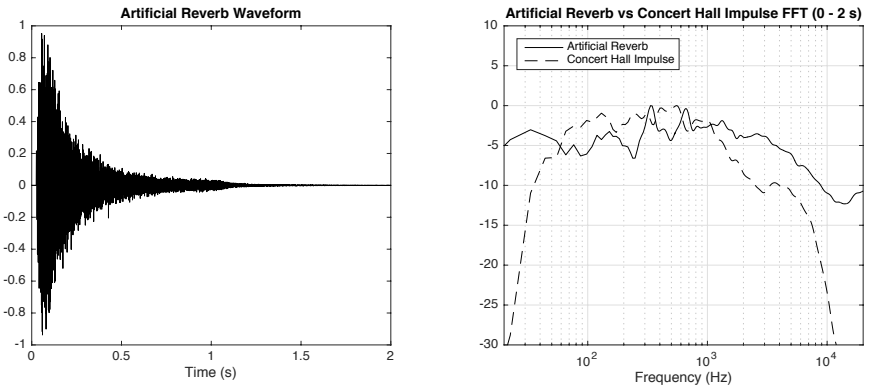

Figure 6.2 Waveform and FFT (long-term average with 4096 points, 4096 sample frame length, 50% overlapping and 1/6th-octave smoothing) of the artificial reverb used to create the ambient stimuli, with the FFT compared against that of a real concert hall impulse beyond the critical distance.

The waveforms and FFT plots of the resultant ambient stimuli (after being processed with the reverb) are presented in Figure 6.3 below, where the vertical dotted lines signify the octave-band limits. The cut-off frequency of the decorrelation was determined by octave-band lower limits, for octave-bands of centre frequencies 63 Hz to 8 kHz. This resulted in 8 high-pass cut-off frequency conditions, as follows: 44 Hz (Broadband), 88 Hz, 177 Hz, 355 Hz, 710 Hz, 1420 Hz, 2840 Hz and 5680 Hz. A condition decorrelating the 16 kHz octave-band only was not included as the complex source signals noticeably lack energy in this region (in particular, see the FFTs of the Cello, Acoustic Guitar and Male Speech ambient sources in Figure 6.3).

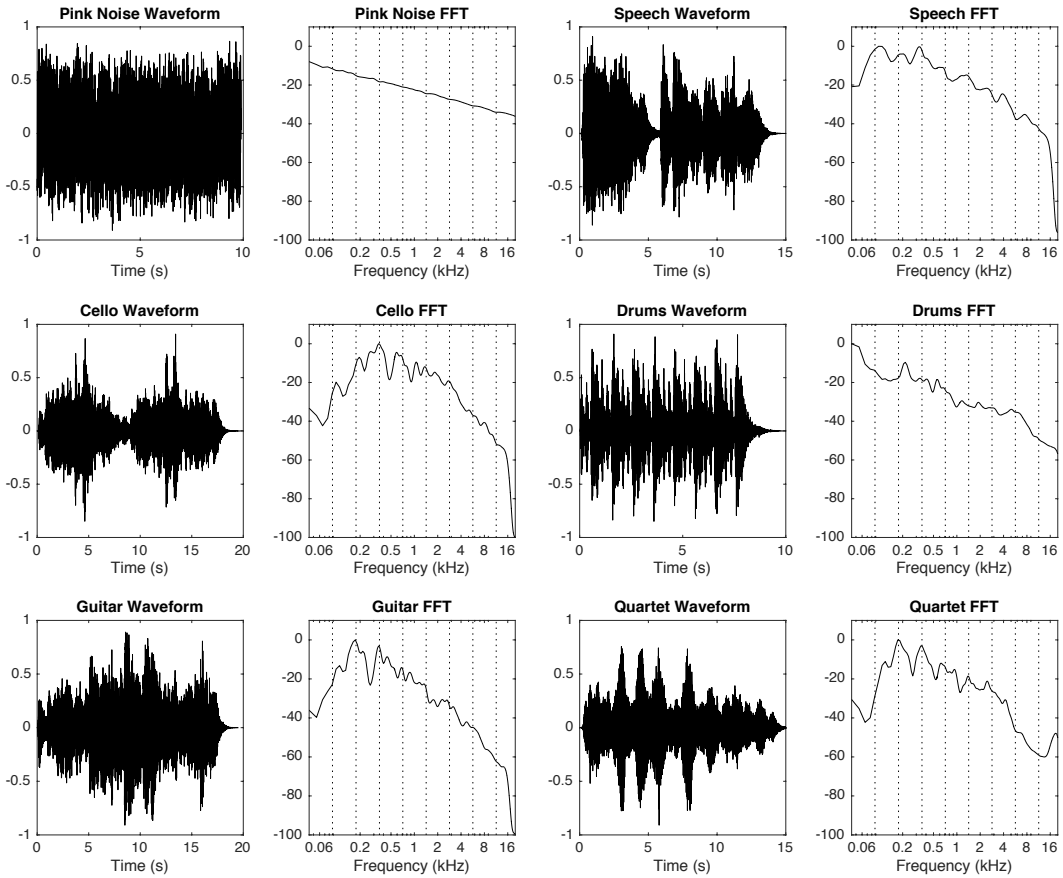

Figure 6.3 Waveforms and FFTs (long-term average with 4096 points, 4096 sample frame length, 50% overlapping and 1/6th-octave smoothing) of the broadband pink noise and ambient test stimuli.

For decorrelation, each of the six source stimuli were filtered into octave-bands (centre frequencies from 63 Hz to 16 kHz) using 16th-order linear phase Butterworth band-pass filters (96 dB/octave). The octave-band filtered signals were then decorrelated independently using the all-pass filter phase randomisation method, as featured in Chapters 4 and 5. This approach involves convolving the stimuli signals with random number sequences of length 30 ms (which equates to convolution with a short burst of noise). Further details on the decorrelation process can be found in Chapter 4 (Section 4.2.1.3). An interchannel cross-correlation coefficient

(ICCC) less than 0.3 was achieved for each octave-band filtered signal (of every source stimulus). Through experimentation it was found that attempting to achieve lower levels of ICCC was not possible for all source signals – 0.3 was chosen as it was an easily achieved threshold across all samples and conditions. ICCC has been calculated as the 50 ms windowed average over the entire signal duration ($ICCC_{avg}$), and the $ICCC_{avg}$ values for the octave-bands of each source are displayed in Table 6.1 below.

Table 6.1 Octave-band interchannel cross-correlation coefficients ($ICCC_{avg}$)

	63 Hz	**125 Hz**	**250 Hz**	**500 Hz**	**1 kHz**	**2 kHz**	**4 kHz**	**8 kHz**	**16 kHz**
Pink Noise	0.25	0.23	0.17	0.19	0.12	0.13	0.05	0.04	0.03
Speech	0.23	0.19	0.23	0.22	0.22	0.15	0.13	0.15	0.21
Cello	0.20	0.26	0.25	0.24	0.23	0.18	0.20	0.15	0.13
Drumkit	0.21	0.22	0.21	0.20	0.20	0.16	0.16	0.14	0.17
Guitar	0.16	0.16	0.22	0.24	0.21	0.25	0.25	0.21	0.19
Quartet	0.21	0.26	0.25	0.25	0.19	0.21	0.22	0.17	0.11

The two decorrelated output signals were RMS level-matched with the input signal, then attenuated by -3 dB, to maintain consistent energy between the input and output. Broadband stimuli signals were reconstructed by summing the decorrelated and non-decorrelated octave-band signals together. For each high-pass condition, the original monophonic octave-band signals that had not been decorrelated were routed to the lower main-layer loudspeaker only, while one decorrelated signal was routed to the main-layer, and the other to the height-layer – this process can be seen in Equations 6.1 and 6.2 below. The -3 dB attenuation of the decorrelated signals ensured that the energy of the summed main- and height-layer signals at the ear remained consistent with the original monophonic condition. In other words, the general energy distribution between the monophonic and decorrelated conditions was mostly identical, other than inherent spectral distortions that may occur through the decorrelation process. The resultant $ICCC_{avg}$ of the reconstructed stimuli are presented in Table 6.2 below, indicating that all conditions were at most 0.3 $ICCC_{avg}$.

$$S_M = \sum_{j=1}^{k} m_j + \sum_{i=1}^{n} (x_i * 0.707) \quad (6.1)$$

$$S_H = \sum_{i=1}^{n} (y_i * 0.707) \quad (6.2)$$

where S_M is the main-layer loudspeaker signal, S_H is the height-layer loudspeaker signal, x_i and y_i are a pair of decorrelated octave-band signals, m_j is an unprocessed monophonic octave-band signal, n is the number of decorrelated octave-bands and k is the number of unprocessed octave-bands (where $k = 9 - n$, based on nine octave-bands from 63 Hz to 16 kHz)

Table 6.2 Average-running interchannel cross-correlation coefficients ($ICCC_{avg}$) of each stimulus condition.

	Broadband	**125 Hz +**	**250 Hz +**	**500 Hz +**	**1 kHz +**	**2 kHz +**	**4 kHz +**	**8 kHz +**
Pink Noise	0.08	0.03	0.03	0.03	0.02	0.02	0.01	0.01
Speech	0.13	0.12	0.09	0.07	0.03	0.02	0.01	0.01
Cello	0.15	0.15	0.15	0.16	0.08	0.05	0.04	0.01
Drumkit	0.10	0.08	0.07	0.03	0.03	0.02	0.02	0.02
Guitar	0.27	0.27	0.29	0.21	0.07	0.04	0.02	0.01
Quartet	0.12	0.12	0.13	0.09	0.05	0.05	0.03	0.01

The conditions within each source were all SPL level-matched (LAeq) (Table 6.3). Each source's SPL level was determined through informal listening of the test stimuli, in order to match the perceived loudness between sources (ranging from 68 to 73 dBA). The SPL for each source was set in this way due to their differing performances. For example, the Cello, Acoustic Guitar and String Quartet samples had small pauses at the end of each phrase, causing a reduction in average SPL (LAeq), whereas the Drumkit loop was tight and had continuous energy throughout. As a result, if each source were matched to the same average SPL, the Drumkit would be perceived as quieter than the other samples. Informal loudness matching was seen as a compromise to compensate for the differences of average energy between the source signals.

Table 6.3 Playback SPL (LAeq) of stimuli for each source

	Pink Noise	Male Speech	Cello	Drumkit	Acoustic Guitar	String Quartet
SPL	71 dBA	71 dBA	73 dBA	68 dBA	72 dBA	72 dBA

6.2.3 Testing Procedure

The present subjective experiment consists of two parts – the first assesses vertical image spread (VIS) of the stimuli described above, and the second looks at tonal quality (TQ) of the same stimuli. Stimuli were presented in multiple comparison trials using an adapted version of the MUSHRA format (ITU-R, 2015b), based on the reasons detailed in Section 3.2.4. Each trial featured nine stimuli, consisting of the eight decorrelated stimuli described in Section 6.2.1 (with the varying high-pass decorrelation cut-off frequencies) and a monophonic condition, where the original source stimulus was reproduced from the lower main-layer loudspeaker only. The monophonic condition was included to represent the practical application of 2D-to-3D upmixing, where a monophonic main-layer signal would be decorrelated vertically to generate new height-channel signals.

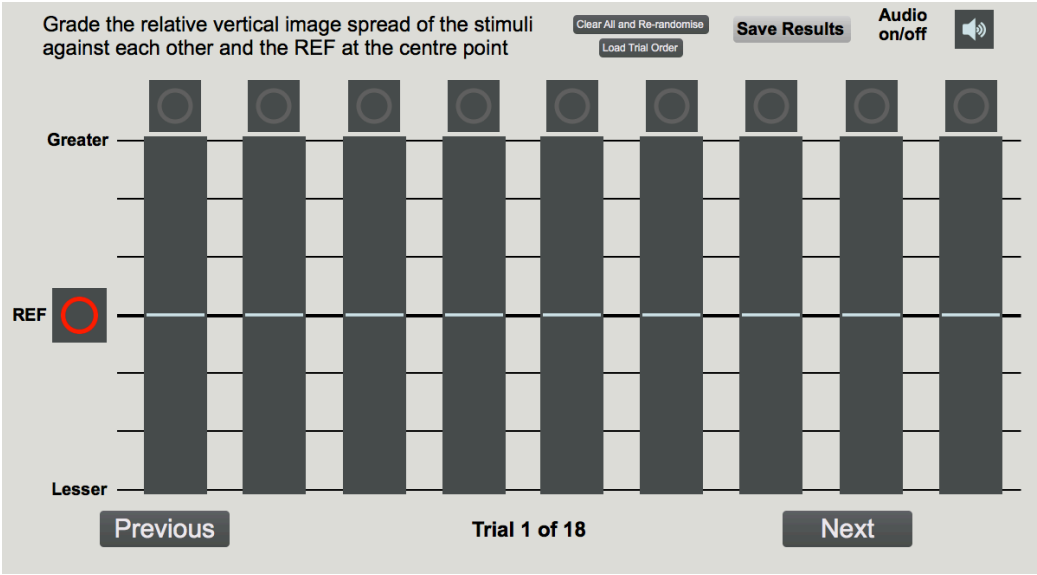

Figure 6.4 Graphical user interface used during the subjective testing.

A total of 18 trials were conducted for each attribute (VIS and TQ) – one for every source and loudspeaker azimuth angle combination (6 x 3). An adaptation of HULTI-GEN was used to create the graphical user interface (GUI) for testing, as shown in Figure 6.4 above (Gribben & Lee, 2015) (Appendix A). Each of the nine stimuli under testing were looped in synchrony and the user could freely switch between them throughout. Subjects were asked to grade the stimuli against each other and a reference stimulus (the monophonic condition). Sliders for grading were on a continuous bipolar scale, ranging from -30 to 30, with the reference located at 0 on the grading scale. If the perception of an attribute was greater or better than the reference, subjects were to grade above 0 on the scale, and if it was lesser or worse, subjects graded below. A familiarisation stage was carried out before testing, introducing the listener to a broadband pink noise source with varying degrees of ICCC, in order to demonstrate changes of VIS. The 18 trials of each attribute were split over two sessions, totalling four sittings of around 30 minutes each to complete both parts.

For the TQ part of testing, the term 'Tonal Quality' was deliberately left open to the listener's own subjective interpretation – this broadly relates the grading of TQ in this instance to a subject's own personal preference. A sheet was given to each subject with various opposing terms and their definitions, in order to give them an idea of what TQ might be interpreted as. These terms were as follows: 'Clear / Muddy', 'Natural / Unnatural', 'Full / Thin', 'Hard / Soft', 'Bright / Dark' and 'Loud / Quiet', along with 'Distortion' and 'Phasiness'. Subjects were asked to write comments (name specific terms) for the highest and lowest graded stimuli of each trial, to indicate why they graded them the way they did – it was also made clear that the written responses were not limited to the terms listed on the sheet.

6.2.4 Subjects

The two parts of the subjective experiment were conducted at two separate times due to the scheduling of the listening room – the vertical image spread (VIS) test was carried out first, followed by the tonal quality (TQ) test at a later date. In total, 17 subjects took part the VIS part

of the experiment, and 13 of the same subjects participated in the TQ part – the other 4 subjects were unavailable at the time of TQ testing, and it was considered best not to include any new subjects for consistency. All subjects reported normal hearing and were familiar with exercises that involve the critical listening of spatial audio. A minimum sample size of 13 subjects results in a statistical power of 0.51 for the experiment, based on a two-tailed t-test with an effect size of 0.6 and α error probability of 0.05 (type I error i.e. false positive), as calculated using G*Power 3.1 (Faul et al., 2007). This indicates that the probability of a type II statistical error (false-negative) is 0.49 (β) i.e. there is a chance that a significant result will be reported as insignificant. Despite this, Bech and Zacharov (2006) state that 5-15 subjects are sufficient for critical listening exercises if the listeners are experienced, as they are in the current test. It is therefore considered that a minimum of 13 listeners in both parts is acceptable.

6.3 Experiment Part 1: Vertical Image Spread (VIS) Results

The relative vertical image spread (VIS) results can be seen in Figure 6.5 below – all data was normalised in accordance with ITU-R BS.1116-3 (ITU-R, 2015a) (as described in Section 3.3) and analysed in SPSS. Shapiro-Wilk tests for normality indicated that the data of each condition was not always normally distributed – as a result, non-parametric statistical tests have been performed on all conditions. The graphs below display the median VIS with notch edge bars (a non-parametric equivalent of 95% confidence intervals (McGill et al., 1978)). In the plots and following results, 'XXX Hz +' indicates decorrelation of the XXX Hz octave-band and above.

Friedman repeated measure tests assessed the effect of the high-pass cut-off frequency on VIS perception. These tests included the eight decorrelated conditions (excluding the monophonic condition), and were performed for each source stimulus and loudspeaker angle combination. If a significant cut-off frequency effect was observed, post-hoc Wilcoxon tests with Bonferroni correction were performed to ascertain where the significant difference occurs. Further Wilcoxon tests with Bonferroni correction were conducted between the monophonic condition and each decorrelated condition, in order to reveal any significant difference of VIS for potential upmixing applications.

6.3.1 Pink Noise VIS Results

The Friedman results for the Pink Noise source indicate a significant effect of cut-off frequency on VIS for all three azimuth angles (0°, ±30° and ±110°) ($p < 0.01$). Post-hoc Wilcoxon tests with Bonferroni correction reveal that there was no significant difference between any decorrelated conditions at 0° azimuth ($p > 0.05$). Observing the graph in Figure 6.5, a general increase of VIS can be seen as the high-pass cut-off decreases to '1 kHz +', before levelling out – this suggests that decorrelation below the 1 kHz octave-band point may not be necessary (i.e. the 63-500 Hz octave-bands). Additional Wilcoxon tests at 0° show that all decorrelated stimuli had significantly greater VIS than the monophonic condition ($p < 0.04$).

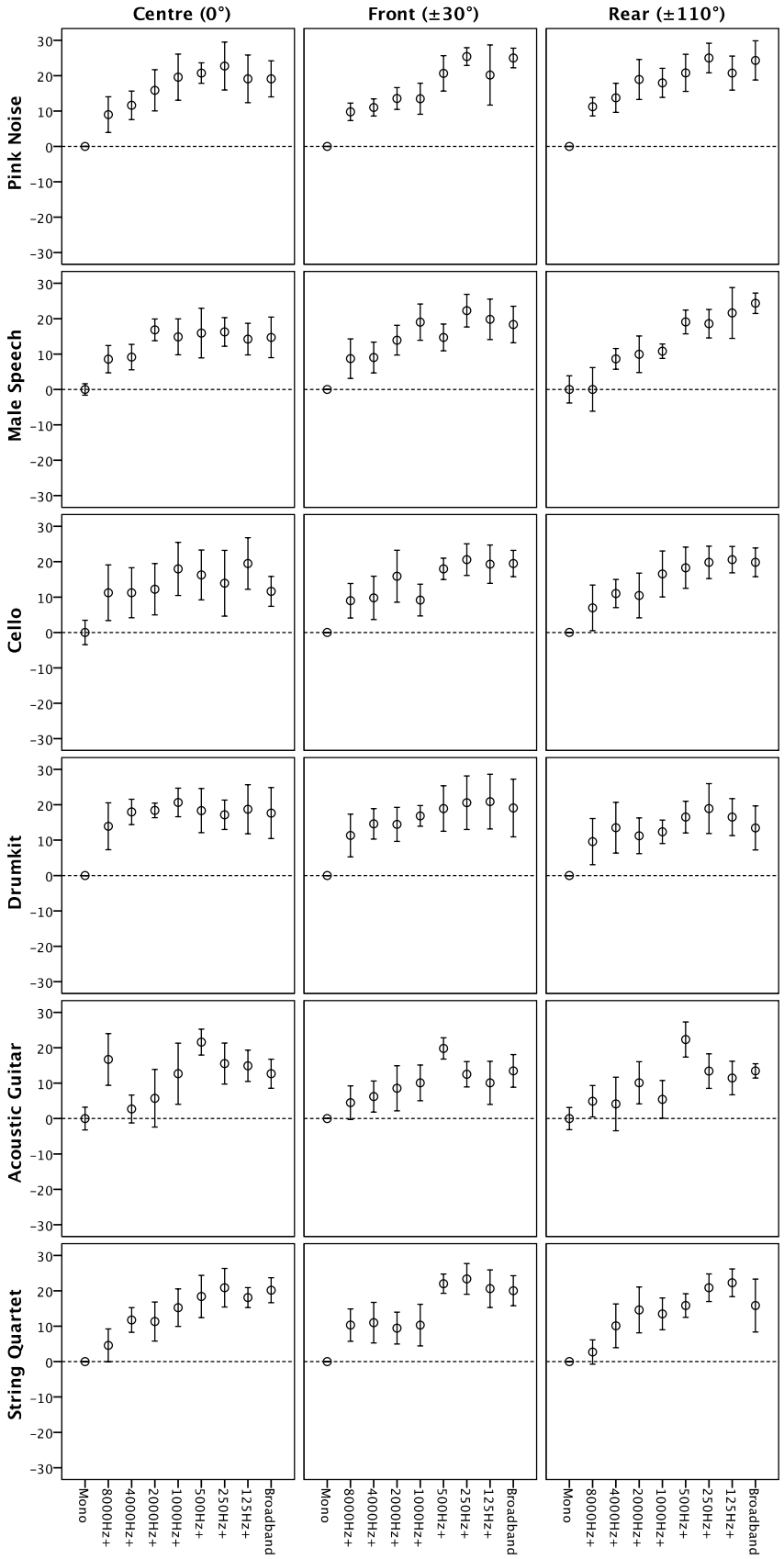

Figure 6.5 Relative vertical image spread (VIS) results (median score with 95% confidence).

For the Pink Noise at ±30° azimuth, Wilcoxon tests with Bonferroni correction indicate that the '250 Hz +' decorrelation condition produces significantly greater VIS than the '1 kHz +', '4 kHz +' and '8 kHz +' conditions ($p < 0.03$). Wilcoxon tests also demonstrate that each decorrelated condition had significantly greater VIS than the monophonic condition ($p < 0.02$).

From ±110° azimuth, the Bonferroni-corrected Wilcoxon tests on the Pink Noise results reveal that the '125 Hz +' condition had significantly greater VIS than the '8 kHz +' condition ($p = 0.03$); furthermore, '250 Hz +' was significantly greater than '2 kHz +', '4 kHz +' and '8 kHz +' ($p < 0.03$). As with the 0° and ±30° angles, additional Wilcoxon tests show that all the decorrelated conditions had significantly greater VIS than the monophonic condition ($p < 0.01$).

6.3.2 Male Speech VIS Results

Friedman tests on the Male Speech VIS data reveal a significant high-pass cut-off frequency effect for the ±30° and ±110° azimuth angles ($p < 0.01$), but not 0° ($p > 0.05$). Although there is no significant difference between the decorrelated conditions at 0°, Bonferroni-corrected Wilcoxon results indicate that every decorrelation condition had significantly greater VIS than the monophonic condition ($p < 0.05$), except for '4 kHz +' ($p > 0.05$).

Between the decorrelated Male Speech conditions at ±30° azimuth, Wilcoxon tests show that the 'Broadband' condition was significantly greater than '4 kHz +' ($p < 0.03$), and both '125 Hz +' and '1 kHz+' were significantly greater than '8 kHz +' ($p < 0.03$). Further Wilcoxon tests demonstrate that most decorrelation conditions were significantly greater than the monophonic condition at ±30° ($p < 0.03$), with the exception of '4 kHz +' and '8 kHz +' ($p > 0.05$).

From the ±110° azimuth angle, the pairwise Wilcoxon tests between the Male Speech stimuli demonstrated that the 'Broadband' condition had significantly greater VIS than the '1 kHz +', '2 kHz +', '4 kHz +' and '8 kHz +' conditions ($p < 0.03$). Furthermore, the '125 Hz +' condition was significantly greater than '8 kHz +' ($p < 0.03$); '250 Hz +' was significantly greater than

‘4 kHz +’ and ‘8 kHz +’ ($p < 0.03$); and ‘500 Hz + was significantly greater than ‘2 kHz +’ ($p < 0.03$).

6.3.3 Cello VIS Results

Results from Friedman tests on the Cello data indicate a significant high-pass cut-off effect for the ±110° azimuth angle ($p < 0.01$), but not the 0° and ±30° azimuths ($p > 0.05$). Wilcoxon tests on the 0° data suggest that only the ‘125 Hz +’ was significantly greater than the monophonic condition ($p = 0.03$). Whereas, from ±30°, the ‘Broadband’, ‘125 Hz +’, ‘250 Hz +’, ‘500 Hz +’ and ‘4 kHz +’ conditions all had significantly greater VIS than the monophonic condition ($p < 0.04$). For the ±110° data, there was no significant difference between any of the decorrelation conditions, following Bonferroni correction ($p > 0.05$) – however, all decorrelated conditions had significantly greater VIS than the monophonic condition ($p < 0.02$).

6.3.4 Drumkit VIS Results

The Friedman results for the Drumkit sample indicate a significant high-pass cut-off effect for the ±30° and ±110° azimuth angles ($p < 0.05$), but not from 0° ($p > 0.05$). Following Bonferroni correction, the Wilcoxon test results show no significant difference between any decorrelated conditions for all azimuth angles ($p > 0.05$). However, when compared against the monophonic condition, all decorrelated conditions had significantly greater VIS from all angles ($p < 0.05$), apart from 2 kHz at ±110° ($p > 0.05$).

6.3.5 Acoustic Guitar VIS Results

Friedman tests on the Acoustic Guitar data demonstrate a significant high-pass cut-off effect for all azimuth angles ($p < 0.05$). From 0°, ‘500 Hz +’ had significantly greater VIS than ‘4 kHz +’ ($p < 0.03$), and the ‘500 Hz +’ condition was the only decorrelated condition with significantly greater VIS than the monophonic condition ($p < 0.01$). With the ±30° results, the ‘500 Hz +’ condition was significantly greater than ‘1 kHz +’ and ‘8 kHz +’ ($p < 0.03$); and for ±110°, ‘500 Hz +’ was significantly greater than ‘1 kHz +’, ‘2 kHz +’ and ‘8 kHz +’ ($p < 0.03$).

From both ±30° and ±110° azimuth, the 'Broadband', '250 Hz +' and '500 Hz +' conditions all had significantly greater VIS than the monophonic condition ($p < 0.05$).

6.3.6 String Quartet VIS Results

The Friedman results for the String Quartet sample show a significant high-pass cut-off effect on decorrelation for all azimuth angles ($p < 0.01$). For 0° azimuth, Bonferroni-corrected Wilcoxon tests indicate that the 'Broadband' condition was significantly greater than '4 kHz +' ($p < 0.01$); and '250 Hz +' was also significantly greater than the '8 kHz +' condition ($p < 0.03$). Furthermore, the 'Broadband', '125 Hz +', '250 Hz +' and '500 Hz +' conditions all had significantly greater VIS than the monophonic condition at 0° ($p < 0.05$).

With the String Quartet at ±30° azimuth, the Wilcoxon results show that '125 Hz +', '250 Hz +' and '500 Hz +' had significantly greater VIS than '4 kHz +' and '8 kHz +' ($p < 0.03$); 'Broadband' was significantly greater than '8 kHz +' ($p < 0.01$); and both '125 Hz +' and '250 Hz +' were significantly greater than '1 kHz +' and '2 kHz +' ($p < 0.03$). Additional Wilcoxon tests indicate that all decorrelation conditions had significantly greater VIS than the monophonic condition at ±30° ($p < 0.05$).

For ±110° azimuth, the 'Broadband', '125 Hz +', '250 Hz +' and '500 Hz +' String Quartet conditions had significantly greater VIS than '8 kHz +' ($p < 0.03$); and both '125 Hz +' and '250 Hz +' were significantly greater than '2 kHz +' and '4 kHz +' ($p < 0.03$). Furthermore, Wilcoxon tests reveal that most decorrelation conditions were significantly greater than the monophonic condition ($p < 0.05$), with the exception of '2 kHz +' and '4 kHz +' ($p > 0.05$).

6.3.7 Discussion of the VIS Results

From the above results, no significant VIS difference was seen between the 'Broadband', '125 Hz +', '250 Hz +' and '500 Hz +' conditions for all source signals and azimuth angles. Moreover, the '500 Hz +' condition also had significantly greater VIS than the monophonic condition for all sources and azimuth angles, except the cello source at 0° azimuth. This

suggests that vertical decorrelation of only the 500 Hz octave-band and above can significantly increase VIS to that of broadband decorrelation in the majority of cases. Observing the graphs in Figure 6.5, it is seen that the Cello sample generally has greater error bars from the 0° azimuth position, indicating that grading of this stimulus may have been difficult at this azimuth angle. A reason for this could be that the Cello is the only musical source with a monophonic melody, whereas the Acoustic Guitar and String Quartet samples have multiple parts playing at different frequencies, which potentially excite multiple VIS cues simultaneously. For example, an inherent vertical spread may have occurred due to different frequencies (melodic parts) being perceived from different heights, i.e., the pitch-height effect (Cabrera & Tilley, 2003; Lee, 2016b). This hypothesis has also been suggested in Section 3.5.2 of the present thesis (see Figure 3.11)

All samples except the Acoustic Guitar see some cases of significant VIS increase for cut-off frequencies higher than the 500 Hz octave-band – this indicates that decorrelation of even higher frequencies alone can have a significant impact on VIS perception, but the effectiveness appears to be largely source-dependent. In order to compare the distribution of frequency energy between the different source signals, octave-band RMS values (normalised to 0 dB for the 1 kHz octave-band) are presented in Table 6.4 below. It is seen that the Acoustic Guitar sample has greatest energy in the 500 Hz octave-band, and relatively lower energy in the 2 kHz and 4 kHz bands compared to the other samples, which may have resulted in the insignificant change of VIS for higher cut-off frequencies that was observed with this sample.

Table 6.4 Octave-band RMS values for the source signals (dB), normalised to 0 dB at the 1 kHz octave-band

	63 Hz	**125 Hz**	**250 Hz**	**500 Hz**	**1 kHz**	**2 kHz**	**4 kHz**	**8 kHz**	**16 kHz**
Pink Noise	0.2	0.1	0.1	0.1	0.0	0.0	-0.1	-0.2	-0.4
Speech	-1.9	1.0	2.5	1.8	0.0	-1.4	-1.9	-5.4	-10.4
Cello	-20.4	-9.9	0.8	1.4	0.0	-1.0	-5.2	-11.0	-17.8
Drumkit	4.9	-0.3	4.6	3.0	0.0	0.0	-0.2	-1.2	-5.6
Guitar	-13.0	0.5	1.6	2.4	0.0	-3.4	-5.5	-9.5	-16.8
Quartet	-16.4	0.9	2.5	1.3	0.0	-0.3	-2.8	-12.0	-12.1

Further to the above observations, it was also revealed in the subjective results that the '500 Hz +' condition had significantly greater VIS than the '8 kHz +' condition for the Acoustic Guitar and String Quartet samples. Looking again at the RMS values in Table 6.4, it is seen that these samples have noticeably less energy in the 8 kHz and 16 kHz octave-bands than the other sources, which would explain the subjective VIS results. Given that the greatest energy for the Guitar, Cello and String Quartet is around the 250 Hz and 500 Hz octave-bands, it is considered that decorrelation of at least the 500 Hz octave-band and above is required to achieve maximum levels of VIS for these (and similar) musical sources. On the other hand, for the Pink Noise and Drumkit sources, there was no significant difference between 'Broadband' decorrelation and '8 kHz +' decorrelation for all azimuth angles – as might be expected, Table 6.4 shows that these two samples have the greatest level of energy in the 8 kHz and 16 kHz bands. The '8 kHz +' condition for both the Pink Noise and Drumkit samples was also perceived as significantly greater than the monophonic condition for all azimuth angles.

The results here suggest that significant VIS increase can be achieved from only decorrelating the 8 kHz and 16 kHz octave-bands, provided the source signal has relatively great energy within these octave-bands. This highlights the importance of the 8 kHz octave-band in VIS perception. From the objective signal analysis in Chapter 5, it is suggested that the VIS perception of the 8 kHz band may work in a similar way to vertical localisation from the front, where notches are 'filled in' and the spectrum is boosted at specific frequency points (Roffler & Butler, 1968a; Hebrank & Wright, 1974). In the absolute VIS testing (Chapter 4), it was observed that the 8 kHz octave-band was biased towards the upper height-layer loudspeaker, when presented in vertical stereophony. This biasing effect could have also influenced the perception of VIS for the '8 kHz +' condition (i.e. through a shift of the 8 kHz band upwards from the monophonic condition). If this were the case, it is possible that simply routing the frequencies of the 8 kHz octave-band and above to the height-channel may significantly increase VIS, without the need for decorrelation. A technique known as perceptual band allocation (PBA) already works on a similar principle, where octave-bands are routed to either the main- or height-layer

to increase VIS, based on their inherent vertical localisation in space (Lee, 2016b). It has previously been shown that routing frequencies of the 2 kHz octave-band and above to the height-layer significantly increases VIS – perhaps a higher cut-off could also be used for generating VIS, provided the source signal has sufficient high frequency energy.

Between loudspeaker positions, a similar relationship between high-pass cut-off frequency and VIS is seen for all azimuth angles. That is, where VIS generally increases as the high-pass cut-off decreases (i.e. the bandwidth of decorrelation increases). The subjective experiment in Chapter 4 indicated that vertical decorrelation of octave-band pink noise is perceived differently depending on the frequency and azimuth angle. Since the results presented here are broadly similar for each angle, it suggests that the vertical decorrelation of multiple octave-bands may each contribute to the same broadband perception of VIS, regardless of presentation location. Furthermore, when the number of decorrelated octave-bands increases, the extent of perceived VIS generally increases too, indicating that VIS takes cues from multiple frequencies simultaneously. From this, it might be assumed that decorrelation of the 8 kHz and 16 kHz octave-bands still provides some contribution to the perception of broadband vertical decorrelation, even when samples lack high frequency energy. It would be interesting to perform an experiment where the 8 kHz and 16 kHz bands are excluded from vertical decorrelation to explore this further.

6.4 Experiment Part 2: Tonal Quality (TQ) Results

The results for the relative tonal quality (TQ) part of the investigation can be seen in Figure 6.6 below. All data was normalised in accordance with ITU-R BS.1116-3 (ITU-R, 2015a) (as described in Section 3.3) and analysed in SPSS. Shapiro-Wilk tests indicated that not all conditions had normally distributed data, therefore, non-parametric statistical tests have been performed on all groups of data. In the graphs, the median TQ scores are plotted with notch edges (a non-parametric equivalent of 95% confidence intervals (McGill et al., 1978)). As with the VIS results, Friedman repeated measure tests were conducted on the data, in order to observe the effect of the high-pass cut-off frequency on TQ. If a significant high-pass cut-off effect was revealed, post-hoc Wilcoxon tests with Bonferroni correction were carried out between the decorrelated stimuli to determine any significant difference. Wilcoxon tests were also conducted between the decorrelated conditions and the monophonic condition, to assess for significant difference of TQ from the unprocessed reference.

6.4.1 TQ Qualitative Responses

A summary of listeners' qualitative responses for the stimuli with the 'best' and 'worst' TQ are presented in Tables 6.5 and 6.6 below, respectively. For each term, the total number of occurrences is displayed both with and without the broadband pink noise responses on the right of the tables – it is thought that the exclusion of pink noise is useful, as it is unusual to judge noise signals in terms of TQ. It is anticipated that the tally of terms presented in Tables 6.5 and 6.6 are used to provide a general indication of the attributes that listeners used to judge TQ, rather than be a statistical analysis of semantic terms. The discussions surrounding these terms are a result of comparing individual subject's terms against their subjective results, in order to determine the 'best' and 'worst' stimuli that the attributes refer to for each listener.

Table 6.5 Summary of the qualitative responses for the sample with the best tonal quality from all trials combined (totals displayed both with and without the broadband pink noise responses)

	Pink Noise	Cello	Acoustic Guitar	Drumkit	Male Speech	String Quartet	Total	Total wo PN
Full	15	18	21	15	13	19	**101**	**86**
Clear	4	16	18	13	21	20	**92**	**88**
Natural	8	15	10	16	9	11	**69**	**61**
Bright	2	8	13	8	3	10	**44**	**42**
Soft	6		3	1	1	3	**14**	**8**
Loud			2	2	2	6	**12**	**12**
Warm		2	1		2	4	**9**	**9**
Bassy	1	2	2	1		1	**7**	**6**
Rich			2	1		1	**4**	**4**
Balanced	1			1		2	**4**	**3**
Intelligible					3		**3**	**3**
Neutral	3						**3**	**0**
Smooth	3						**3**	**0**
Thin		1					**1**	**1**
Livelier				1			**1**	**1**

Table 6.6 Summary of the qualitative responses for the sample with the worst tonal quality from all trials combined (totals displayed both with and without the broadband pink noise responses)

	Pink Noise	Cello	Acoustic Guitar	Drumkit	Male Speech	String Quartet	Total	Total wo PN
Muddy	3	15	22	6	21	17	**84**	**81**
Phasey	20	9	2	20	14	7	**72**	**52**
Thin	12	15	9	5	9	15	**65**	**53**
Unnatural	5	7	4	14	12	3	**45**	**40**
Dull	1	7	6	2	3	6	**25**	**24**
Hard	4	3	1	3	4	4	**19**	**15**
Harsh	6	2	4	3	2	1	**18**	**12**
Distorted	2	4	2	3	4	2	**17**	**15**
Metallic	2	2		4	2		**10**	**8**
Too Full	1		1	2	1		**5**	**4**
Dark	1		1		1	2	**5**	**4**
Too Bright	2		1			2	**5**	**3**
Quiet	1		1	1			**3**	**2**
Bassy			2	2			**4**	**4**
Distant			1	2			**3**	**3**
Loud		1			1		**2**	**2**
Soft					1	1	**2**	**2**
Reverberant		1					**1**	**1**
Twangy			1				**1**	**1**
Rough				1			**1**	**1**
Narrow	1						**1**	**0**

Observing the results for the best samples (excluding pink noise), both ‘Full’ and ‘Clear’ are the most regularly occurring terms, and for the ‘worst’ samples, the most common terms are ‘Muddy’, ‘Phasey’ and ‘Thin’. The combination of ‘Clear’ and ‘Muddy’ suggests that TQ judgements were largely based on the clarity of the stimulus, while ‘Full’, ‘Phasey’ and ‘Thin’ imply a phase cancelling effect for some stimuli. In particular, ‘Full’ and ‘Thin’ refer to the

level of low frequency energy – considering this, it suggests that some low frequency distortion may be present. Furthermore, observing the qualitative results when the pink noise responses are included, 'Full' and 'Clear' are still the terms most referred to for the best samples; however, for the worst samples, there is a large increase of the term 'Phasey' (from 52 to 72 occurrences). It is assumed that this does not refer to the monophonic reference stimulus, which further suggests that phase cancellation occurs when the decorrelated signals are summed at the ears (this has been confirmed by looking at the worst graded sample for each individual listener). It was previously observed in the spectral analysis of Section 5.3 that summing distortion from decorrelation is most evident at lower frequencies – this effect has been discussed further in Section 6.5 below.

6.4.2 Pink Noise TQ Results

Looking towards the subjective results, Friedman tests for the pink noise data indicate a significant high-pass cut-off frequency effect at all azimuth angles ($p < 0.05$). However, the Bonferroni-corrected Wilcoxon tests demonstrate that only the 'Broadband' condition had significantly worse TQ than the '4 kHz +' condition at 0° azimuth ($p < 0.03$). Additional Wilcoxon tests show that the 'Broadband', '125 Hz +', '250 Hz +' and '500 Hz +' conditions all had significantly worse TQ than the monophonic condition at 0° ($p < 0.05$); and from ±30°, the '125 Hz +' was also perceived as significantly worse than the monophonic condition ($p < 0.04$). Comments relating to all of the conditions with significantly worse TQ refer to the stimuli as 'Phasey' and 'Thin' – as suggested above, this potentially indicates a loss of frequencies when the decorrelated signals are summed at the ear. Although it is difficult to subjectively judge the TQ of pink noise (due to its unnatural features), it is clear that as the high-pass cut-off frequency decreases (i.e. as more low frequencies are decorrelated), the perceived TQ appears to decrease almost linearly, with a greater variation of responses as the azimuth angle increases.

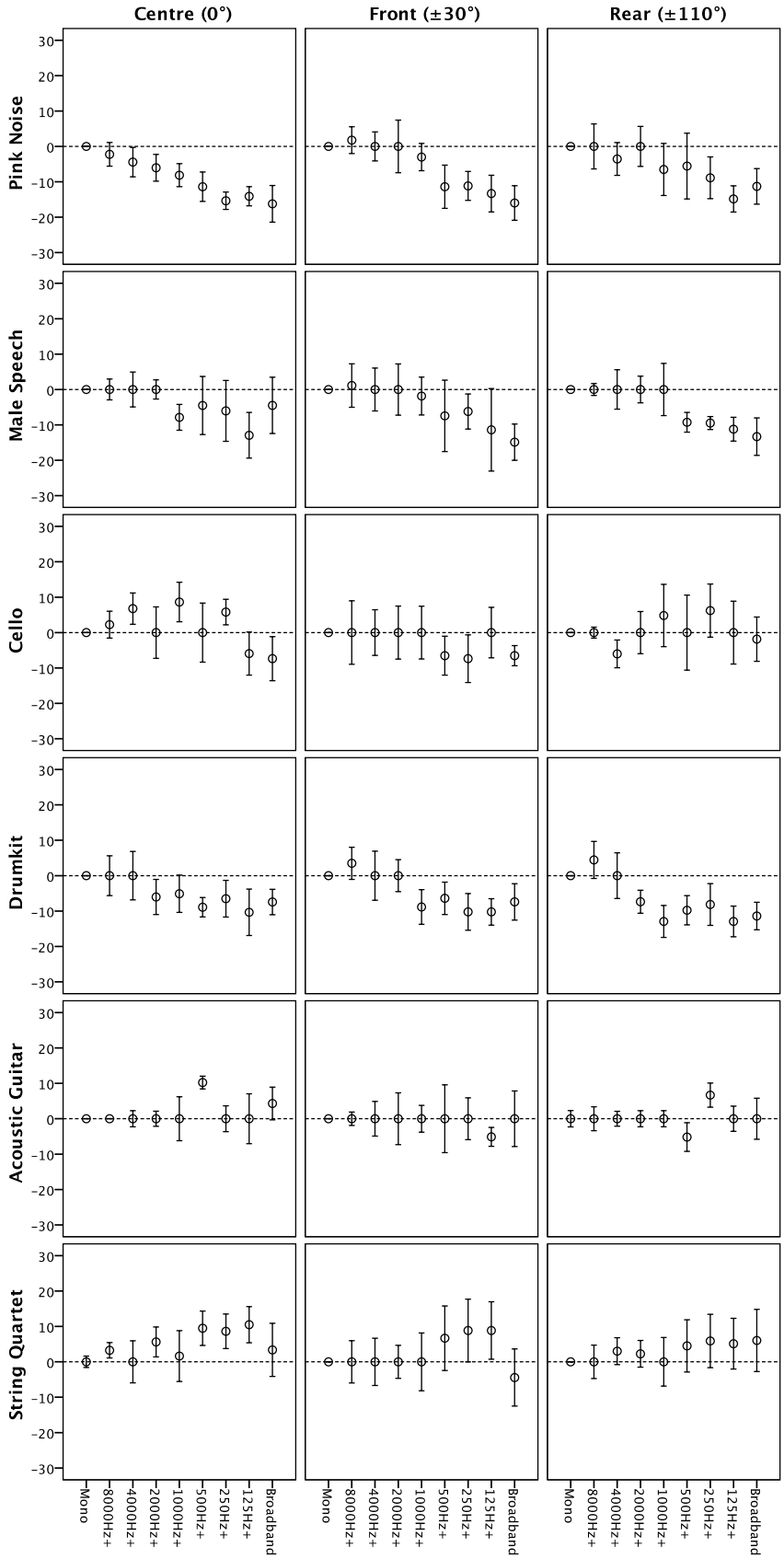

Figure 6.6 Relative tonal quality (TQ) results (median score with 95% confidence).

6.4.3 Male Speech TQ Results

The Friedman results for the male speech sample indicate that the high-pass cut-off frequency has a significant effect on TQ at ±110° azimuth ($p < 0.05$), but not 0° and ±30° ($p > 0.05$). However, the Bonferroni-corrected Wilcoxon tests show no significant difference between the decorrelated conditions from all azimuth angles ($p > 0.05$). Compared against the monophonic condition, 'Broadband' at ±110° is the only decorrelated condition to be graded as significantly worse ($p < 0.03$). Comments for the 'Broadband' stimulus refer to terms such as 'Muddy', 'Distorted', 'Unnatural' and 'Dull' (when looking at subjects' individual responses), suggesting that decorrelation of lower frequencies may have an impact on the intelligibility of speech sources. Observing the plots in Figure 6.6, a general decrease of TQ is seen as the high-pass cut-off frequency decreases, similar to the Pink Noise sample – however, unlike the Pink Noise source, this decrease of TQ appears to be most prominent from ±110° azimuth.

6.4.4 Cello TQ Results

With the Cello source, the Friedman results show no significant high-pass cut-off frequency effect at all azimuth angles ($p > 0.05$). There was also no significant difference between the monophonic condition and all of the decorrelated conditions ($p > 0.05$). Looking at the graphs in Figure 6.6, a slight improvement of TQ is seen with some decorrelation conditions at both 0° and ±110° azimuth. The most common terms for the Cello sample with the best TQ are 'Full', 'Clear' and 'Natural', suggesting there may be some tonal benefit to decorrelation.

6.4.5 Drumkit TQ Results

The Friedman results for the Drumkit sample indicate a significant high-pass cut-off frequency effect on TQ at all azimuth angles ($p < 0.05$). However, the Wilcoxon tests show no significant difference between any of the decorrelation conditions, following correction ($p > 0.05$). Additional Wilcoxon tests also demonstrate that the 'Broadband', '125 Hz +' and '1 kHz +' conditions had significantly worse TQ than the monophonic condition at ±110° ($p < 0.05$). As with the pink noise and male speech sources, a general decrease of TQ as the cut-off frequency

decreases can also be seen in Figure 6.6 for the Drumkit sample. The comments for the Drumkit stimuli with significantly worse TQ were also similar to that of the Pink Noise and the Male Speech samples: 'Phasey', 'Thin', 'Unnatural', 'Distorted' and 'Muddy'. For the total tally of responses in Table 6.6, the most common term for the worst sample was 'Phasey' (20 occurrences) followed by 'Unnatural' (14 occurrences) – this further implies that vertical decorrelation is causing a noticeable loss of frequencies.

6.4.6 Acoustic Guitar TQ Results

Friedman analysis of the acoustic guitar data reveal a significant high-pass cut-off frequency effect for the 0° azimuth angle ($p < 0.05$), but not ±30° and ±110° ($p > 0.05$). The post-hoc Wilcoxon tests suggest that the difference between the decorrelated conditions are not significant, following correction ($p > 0.05$). Observing the graphs in Figure 6.6, it is seen that most conditions see no change of TQ from the monophonic reference. Some slight increase of TQ is observed for the '500 Hz +' condition at 0° and '250 Hz +' at ±110°.

6.4.7 String Quartet TQ Results

The Friedman results for the String Quartet sample show a significant effect of the high-pass cut-off frequency for the 0° azimuth angle ($p < 0.05$), but not ±30° and ±110° ($p > 0.05$). From 0°, Wilcoxon tests reveal that the '125 Hz +' condition had significantly better TQ than both the '8 kHz +' and monophonic conditions ($p < 0.04$). In contrast to the other samples, TQ generally appears to increase slightly as the high-pass cut-off frequency decreases (i.e. when more low frequencies are decorrelated). The subject comments for the significant '125 Hz +' condition suggest that decorrelation made the sample 'Warmer', 'Softer', 'Livelier', 'Fuller' and 'Clearer'. This increase of TQ could be due to the musical nature of the source, where the frequency content from multiple parts varies over time, potentially leading to a richer sound.

6.4.8 Discussion of the TQ Results

In general, high-pass vertical decorrelation appears to have little negative effect on TQ – there is no significant difference between the monophonic unprocessed condition and the '2 kHz +', '4 kHz +' and '8 kHz +' conditions for all sources. Apart from the Pink Noise stimuli, only the Male Speech and Drumkit samples saw a significant decrease of TQ from decorrelation (both at ±110°). These results could be related to the inherent nature of speech and drums; for example, the familiarity of speech and transient response of the drums may have revealed signal degradation more clearly. Observing the FFTs in Figure 6.3, it is also seen that both the Male Speech and Drumkit sources have relatively greater low frequency content than the other samples – this is further reflected by the octave-band RMS calculations in Table 6.4 above. In the spectral analysis of Chapter 5 (Section 5.3), it was noticed that summing the decorrelated signals resulted in a distorted spectrum, particularly at lower frequencies – this was presumably due to large opposing shifts of phase in either signal. When summing the broadband decorrelated signals of the present stimuli, a similar distortion is also seen. Figure 6.7 shows the delta plot of the difference between the monophonic and broadband decorrelated conditions, representing the monophonic FFT spectra taken from the summed decorrelation FFT spectra. It is observed that the distortions vary for each source. However, there appears to be a general tendency for greater distortion between the 63 Hz and 500 Hz octave-bands. Given the increased lower frequency energy in the Pink Noise, Male Speech and Drumkit samples, it is possible that these low frequency distortions were more noticeable than with other samples, causing a decrease of perceived TQ. Relating this back to the qualitative comments for TQ in Table 6.6, it is seen that the most common term for the worst Pink Noise and Drumkit samples is 'Phasey' – which further suggests that the degradation of TQ seen for these stimuli is related to the phase relationship when summing at the ear (particularly at lower frequencies). For the worst Male Speech stimuli, the most common terms are 'Muddy' and 'Phasey', suggesting that this decorrelation distortion may also have an impact on intelligibility and clarity.

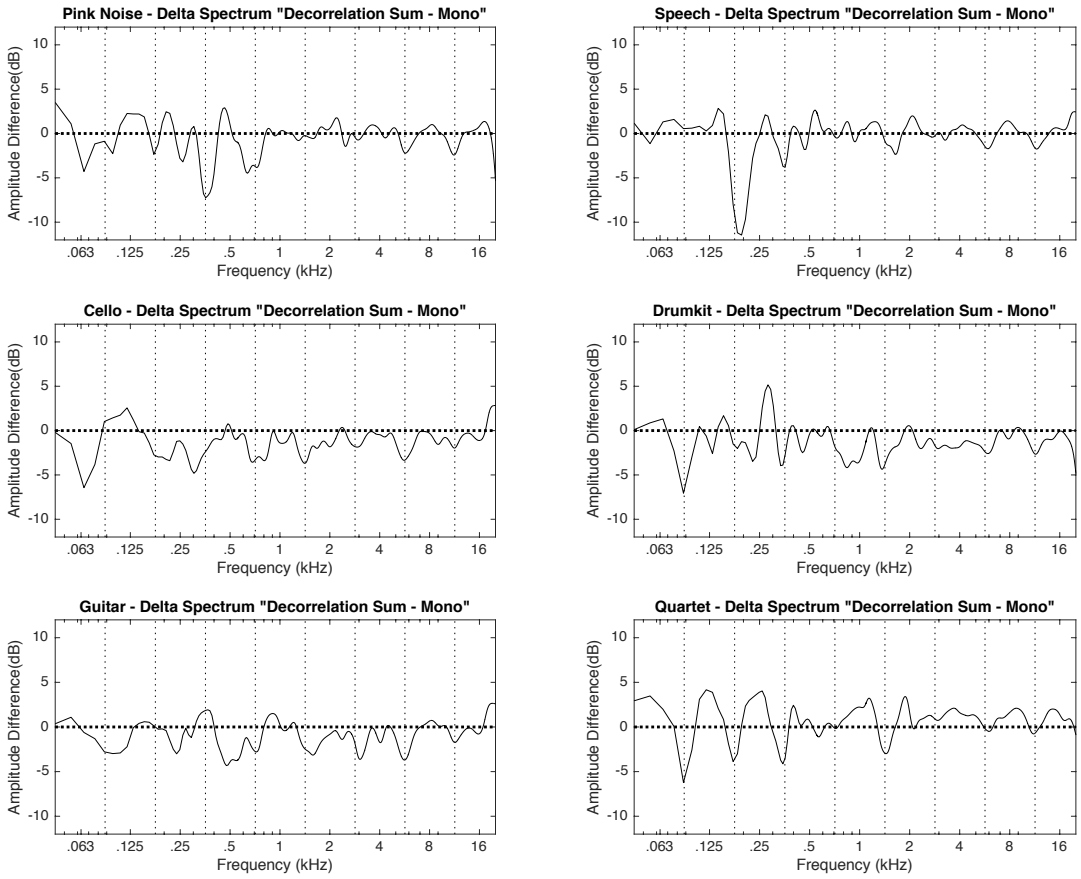

Figure 6.7 Delta spectra: 'Summed broadband decorrelated signals - unprocessed monophonic signal'

With the Cello, Guitar and String Quartet samples, there was some evidence of improved TQ following decorrelation. Observing spectrograms of the samples in Figure 6.8 (displaying the short-time Fourier transform (STFT) over time), the time-varying change in spectrum is clearly seen for the musical sources, and there are also clear harmonic tones above the fundamental frequency of the notes (particularly with the Cello). In contrast, the Male Speech and Drumkit samples have mostly consistent or repeating spectrums over time, particularly at lower frequencies. It is possible that the negative effects of decorrelation on TQ (i.e. the distortion seen in Figure 6.7) are less perceivable for polyphonic musical sources with moving parts; whereas for relatively steady-state broadband sources with repetitive or familiar patterns (e.g. Speech and

Drums), the low frequency distortion causes a noticeable degradation of TQ. This suggests that for application on a wide variety of sources, 2D-to-3D upmixing may benefit from high-pass decorrelation, in order to reduce the low frequency summing distortion when using the all-pass filter approach.

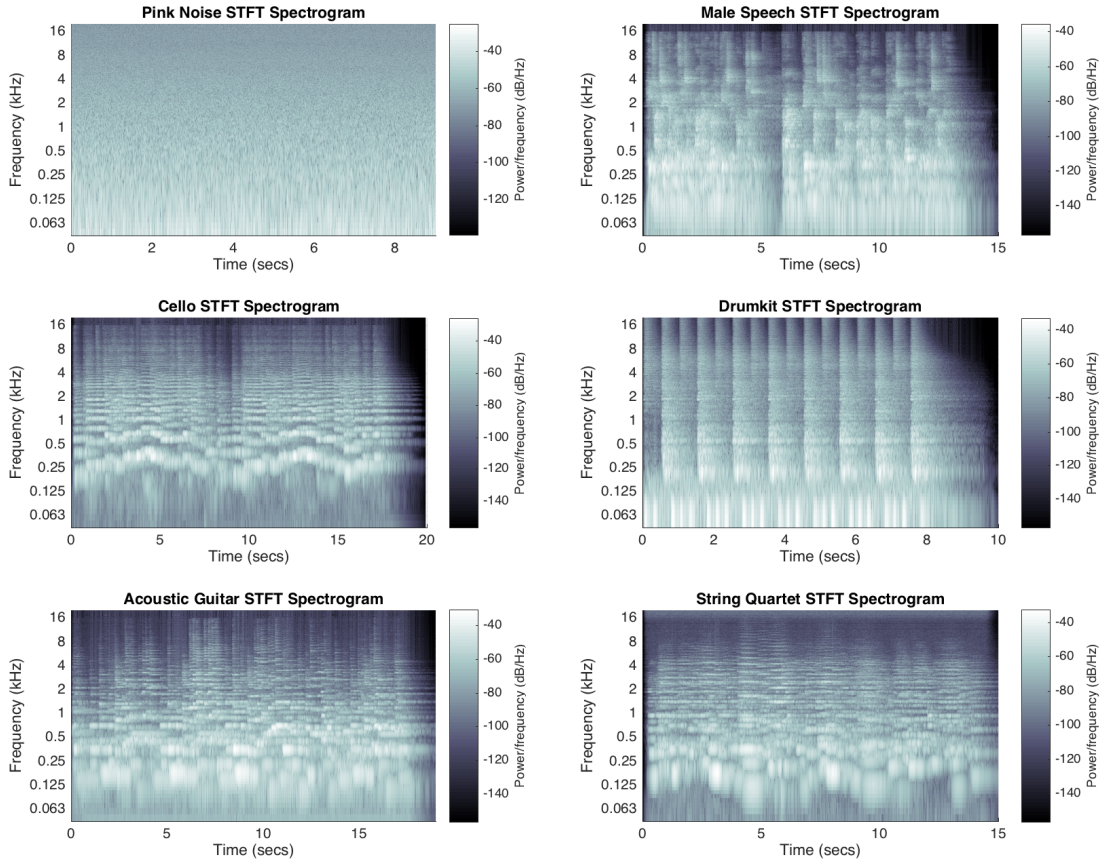

Figure 6.8 Spectrogram showing the short-time Fourier transform (STFT) of the ambient source signals, 4096 FFT-points calculated with a frame length of 1024 samples and 50% overlapping windows

6.5 Overall Discussion of Results

The vertical image spread (VIS) testing results appear to suggest that the VIS increase from high-pass decorrelation peaks around the 500 Hz octave-band cut-off, with decorrelation of lower frequencies resulting in an insignificant change to VIS in comparison. This is in line with the subjective results of the octave-band decorrelation tests in Chapter 4 (Section 4.2.2), where no significant increase of VIS was seen for octave-bands with centre frequencies of 63 Hz, 125 Hz and 250 Hz. These results imply that the decorrelation of lower frequencies may not be necessary in terms of increasing overall VIS, and when compared against a monophonic reference, a significantly greater level of VIS can be achieved from just decorrelating the higher frequencies in broadband signals (similar to that of broadband decorrelation).

Further to this, the tonal quality (TQ) part of the experiment seems to indicate that signals with great and steady low frequency energy experience a loss of TQ (i.e. the Male Speech, Drumkit and Pink Noise samples). It is thought that this decrease of TQ is related to summation distortion when using the all-pass filter phase randomisation method, where greater distortion is seen at lower frequencies. Given this and the VIS findings, it is proposed that the '500 Hz +' condition might be a suitable compromise for increasing vertical extent, while reducing the low frequency distortion effect. In some instances, the TQ of the musical sources improved through decorrelation of lower frequencies, where the spectrum is varied over time due to the musical melody. Results from Robotham et al. (2016) indicate that the inclusion of a single ceiling reflection can lead to a preferred perception of musical sources – this may be similar to the effect observed here. For the '500 Hz +' condition, the musical samples were graded as similar or slightly greater than the monophonic unprocessed reference, implying that decorrelating the 500 Hz octave-band and above could potentially be applied universally to increase VIS, with little detrimental effect to the perceived TQ. Having said that, it should also be noted that the samples used in this experiment were ambient to represent an upmixing scenario. It is possible

that if dry or anechoic signals were used, the detrimental effects of vertical decorrelation on TQ might be more apparent.

6.6 Conclusions

A two-part listening test has been conducted, investigating the effect of high-pass decorrelation, with varying degrees of cut-off frequency. The first part observed relative vertical image spread (VIS) between stimuli, and the second looked at the perceived tonal quality (TQ) differences from vertical decorrelation. The high-pass cut-off frequencies were based around eight octave-bands with centre frequencies from 63 Hz to 8 kHz, where the lower limit of each octave band defined the cut-off frequency. For example, the '1 kHz +' condition was decorrelation of the 1 kHz octave-band and all octave-bands above.

Decorrelation was applied to each octave-band independently, then stimuli were constructed by routing the two signals of a decorrelated octave-band between a main- and height-layer loudspeaker pair, with the unprocessed octave-bands routed to the main-layer loudspeaker only (i.e. the octave-bands below the cut-off). A monophonic condition was also included as a reference stimulus, where all octave-bands were routed to the lower main-layer loudspeaker only. The decorrelation technique used was the all-pass filter phase randomisation method, which was also implemented in Chapter 4 of this study. Six source signals were tested: Broadband Pink Noise, Male Speech, Cello, Drumkit, Acoustic Guitar and String Quartet. All stimuli were presented from the same three azimuth angles as assessed in Chapter 4 (0°, ±30° and ±110°).

The key findings from the listening tests are as follows:

- The '125 Hz +', '250 Hz +' and '500 Hz +' high-pass conditions had a similar perceived VIS to 'Broadband' decorrelation for all sources and azimuth angles.
- The Pink Noise and Drumkit samples had greater energy at 8-16 kHz, resulting in the '8 kHz +' condition having significantly greater VIS than the monophonic sample.
- Summing two phase-decorrelated (all-pass filtered) signals results in spectral distortion, most notably at lower frequencies.

- Samples with strong and steady low frequency energy (Pink Noise and Drumkit) had a significant decrease of TQ with decorrelation of lower octave-bands (63-125 Hz).
- Low frequency decorrelation of musical sources (where the frequency content varies over time i.e. Cello, Acoustic Guitar and String Quartet) sees an improvement of TQ.
- For all sources except Pink Noise at 0°, the '500 Hz +' condition has no significant difference in TQ from the monophonic reference.

These results demonstrate the source-dependency of vertical interchannel decorrelation on both VIS and TQ perception, as well as the importance of the 8 kHz octave-band as seen in the previous subjective experiments of Chapter 4. From the VIS and TQ results presented here, it appears that high-pass decorrelation of the 500 Hz octave-band and above significantly increases VIS, with no significant degradation to the TQ of complex stimuli. Considering this, the following experiment in Chapter 7 compares both broadband decorrelation and '500 Hz +' decorrelation in a practical upmixing scenario – and similar to the present experiment, both the spatial effect (listener envelopment (LEV)) and TQ have been assessed.

7 2D-TO-3D UPMIXING BY VERTICAL INTERCHANNEL DECORRELATION

This chapter describes a two-part experiment that investigates the perception of vertical interchannel decorrelation in the application of 2D-to-3D upmixing. To assess the spatial effect of this application, the first part looks at the perception of listener envelopment (LEV) between 2D and 3D upmixed stimuli; while the second part assesses the subjective tonal quality (TQ) of the same stimuli, similar to that of the second listening test in Chapter 6. In the previous chapters, the concern has been on vertical decorrelation at discrete azimuth angles to the listener. Potential cues at higher frequencies were found in Chapters 4 and 5, which seem to contribute to the perception of vertical image spread (VIS) (from around the 2 kHz octave-band and above, depending on the angle of incidence). However, these observations were not made in a practical upmixing scenario, and the effects of vertical decorrelation must also be investigated within an established surround sound system (i.e. presented from multiple azimuth angles simultaneously). Consequently, the present experiment looks at vertical decorrelation of multiple ambient signals, using both phase-based and amplitude-based decorrelation methods, in order to generate 3D content for the commercial Auro-3D 9.1 format (Auro Technologies, 2015a). Furthermore, it was found in Chapter 6 that when vertically decorrelating frequencies of the 500 Hz octave-band and above only, the perception of vertical image spread (VIS) was similar to decorrelating the broadband signal, with little detrimental effect to the TQ. As a result, both broadband and high-pass decorrelation have been assessed for each decorrelation method, in order to observe whether upmixing low frequencies is necessary.

From the above considerations, the research questions for this investigation are as follows:

- What influence does vertical decorrelation of ambience in 3D formats have on LEV?
- Does the decorrelation method have an impact on the effectiveness of upmixing?
- Is there a perceptual difference between broadband and high-pass decorrelation?

- What impact does 2D-to-3D upmixing by decorrelation have on TQ?
- Are the subjective perceptions of LEV and TQ source-dependent?

In order to answer these questions and to provide a practical scenario, 5-channel test stimuli (5.1 content) were created with both direct and ambient components. The ambient signals were then decorrelated between the main- and height-layer loudspeakers of Auro-3D 9.1, in order to create the 3D upmixed stimuli. The level of direct sound energy and ambience energy were kept constant between the original 2D stimulus and the 3D upmixed stimuli, so that only the effects of vertical decorrelation on the ambient signals can be observed. A total of three decorrelation methods have been assessed: the phase-based all-pass filter technique used in Chapters 4 and 6, along with another phase-based approach and an amplitude-based method – details of which can be seen in Section 7.2.1.5 below.

7.1 Experimental Hypotheses

Since the previous experiments of the present thesis have focused on vertical decorrelation at discrete azimuth angles, it is unknown whether similar effects will be perceivable in a 2D-to-3D upmixing scenario (i.e. when decorrelated signals are presented simultaneously from multiple angles). Some studies in the literature indicate that LEV is related to the energy of late reflections arriving from above, as well as from lateral directions (Furuya et al., 1995; Furuya et al., 2001; Furuya et al., 2007). It is therefore hypothesised that the inclusion of artificially decorrelated ambience from elevated positions will also increase the sensation of LEV.

Considering the differences between amplitude- and phase-based decorrelation, a phase-based approach relies on slight phase differences between two signals. It may be found that when many similar ambient signals are reproduced simultaneously (of which the phase is already randomised), these perceptual cues become confused and break down. For example, if VIS perception is reliant on the phase relationship between a specific pair of signals, the inclusion of additional phase-decorrelated signals at the ear may reduce the effectiveness of discrete interchannel VIS cues. In contrast, it is thought that an amplitude-based method could provide perceptually clearer interchannel differences between multiple signals. With amplitude-based methods, each decorrelated pair is to be processed using the exact same pair of filters, from which the regular amplitude differences are generated. If each pair were processed identically in 2D-to-3D upmixing, the interchannel amplitude differences would effectively become interlayer amplitude differences. Given this, it is hypothesised that, if a difference between the decorrelation techniques is apparent, the amplitude-based method will outperform the phase-based methods.

The results of Chapters 4, 5 and 6 suggest the importance of high frequencies in vertical decorrelation. Chapter 6 demonstrated that sources with greater energy in the 8 kHz and 16 kHz octave-bands resulted in a greater perception of vertical image spread (VIS), when these octave-bands were vertically decorrelated alone. As a result, it is hypothesised that the effectiveness

of upmixing by vertical decorrelation is also reliant on sufficient high frequency energy in the source signals. Furthermore, it is thought that if upmixing by decorrelation proves effective, then both broadband and high-pass decorrelation will perform similarly (provided the source signal has adequate high frequency information). The results of Chapter 6 also suggest that signals with a greater low frequency energy result in lower TQ when subjected to broadband decorrelation, due to low frequency spectral distortion when the signals are summed at the ear. From this, it is lastly hypothesised that high-pass decorrelation will improve TQ for sources with greater energy in the low frequencies (63 Hz – 250 Hz octave-bands), whilst providing a similar perception of LEV.

7.2 Experimental Design

7.2.1 Physical Setup

During testing, the commercial 3D loudspeaker format being upmixed to was Auro-3D 9.1 (Figure 7.1) (Auro Technologies, 2015a), which consisted of nine Genelec 8040A loudspeakers (Frequency response: 48 Hz – 20 kHz (±2 dB)) surrounding the listener. With this format, five main-layer loudspeakers were positioned 2 m from the listener at ear height, with azimuth angles of 0°, ±30° and ±110° from the centre position in front. Four height-layer loudspeakers were positioned directly above the four ±30° and ±110° main-layer loudspeakers, at an elevation angle of +30° to the listening position (vertically spaced by 1.15 m). Auro-3D was chosen as the main-layer is identical to the 5.1 format (ITU-R, 2012), allowing for a straightforward comparison between 2D and 3D content. It is also a similar layout to the vertically-arranged loudspeaker conditions tested during the subjective experiments of Chapters 4 and 6, minus the centre height-channel loudspeaker.

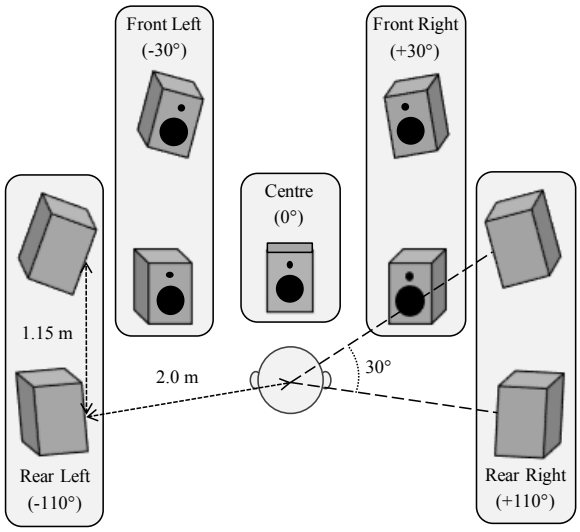

Figure 7.1 Physical loudspeaker setup used during testing (Auro-3D 9.1 (Auro Technologies, 2015a)). Five main-layer loudspeakers positioned 2 m from the listener at ear height with azimuth angles of 0°, ±30° and ±110°. Four height-layer loudspeakers positioned directly above the ±30° and ±110° main-layer loudspeakers at +30° elevation from the listening position.

Testing was conducted at the University of Huddersfield in a critical listening room that fulfils the specification of ITU-R BS.1116-3 (ITU-R, 2015a) (6.2m x 5.6m x 3.8m; RT = 0.25 s; NR

12). Time and level alignment were applied between the two loudspeaker layers, to compensate for interlayer differences at the listening position. An acoustically transparent curtain was used to obscure the loudspeakers from view, so as to avoid visual bias during testing.

7.2.2 Stimuli Creation

The present investigation focuses on 'upmixing' 2D 5.1 content to Auro-3D 9.1, by applying vertical interchannel decorrelation to ambient signals. Typically, ambient signals feature in the rear Left Surround (Ls) and Right Surround (Rs) channels of 5.1 content, which would make a simple scenario for practical upmixing; however, ambient signals can also be obtained through ambience extraction techniques (Avendano & Jot, 2002), or by using the raw ambient signals captured in surround sound recording. When recording audio for 5.1 surround sound, a common method is to divide the recording into the frontal stereo image and ambient parts (Rumsey, 2001). The frontal stereo image refers to the auditory image across the Left (L: -30°), Centre (C: 0°) and Right (R: +30°) loudspeaker channels – these three signals are often recorded by positioning three directional microphones to capture (mostly) direct sound from the source. In contrast, the ambient signals for the L, R, Ls (-110°) and Rs (+110°) channels are captured by facing directional microphones away from the source to capture the reflective energy. In the context of the current experiment, the inclusion of direct sound with each stimuli condition is important for providing a realistic upmixing situation, rather than just assessing the upmixed ambience alone.

7.2.2.1 Multichannel Room Impulse Responses (MRIRs)

To represent different 2D-to-3D upmixing scenarios, it is thought that using multichannel room impulse responses (MRIRs) of recognised surround sound recording techniques would be a versatile approach. Convolution of anechoic signals with concert hall MRIRs gives the ability to keep the direct and reverberant conditions consistent between stimuli, while also replicating the practical scenario of upmixing from a 5.1 surround sound recording. To generate the frontal array (C, L and R), MRIRs of the Optimised Cardioid Triangle (OCT) microphone technique

were convolved with anechoic samples (Thiele, 2001). The OCT method features a central cardioid microphone facing the source to capture the direct sound for the C channel; and two hypercardioid microphones are spaced either side of the centre microphone, angled off-centre from the source by -90° for L and +90° for R (i.e. facing the side walls), so that the direct sound is reduced by -9 dB and mostly lateral reflections are captured. This approach gives a focused direct image in the C channel, while providing a slight apparent source width (ASW) due to the lateral reflections.

For the ambient signals of the L, R, Ls and Rs channels, the source signals were convolved with MRIRs captured using the "Hamasaki-Square" (HS) microphone technique (Hamasaki, 2003) (Figure 7.2). This technique employs four side-facing figure-of-eight microphones in a 2 by 2 m square, positioned beyond the critical distance of the performance space (one microphone for each of the four channels). It is these four ambient signals that are decorrelated (upmixed) into their respective height-channels. Given that the HS technique captures ambience, all four impulses have a similar frequency response; therefore, Figure 7.2 only shows the waveform and FFT of the 'Ls' impulse. Both the direct and ambient MRIRs were obtained from a database of impulses captured in St. Paul's Concert Hall at the University of Huddersfield (RT = avg. 2.1s), using the HAART impulse response toolbox (Johnson et al., 2015) with an exponential sine-sweep (Farina, 2000). A summary of how these MRIRs translate to both the 5.1 and Auro-3D 9.1 formats in the stimuli creation can be seen in Table 7.1.

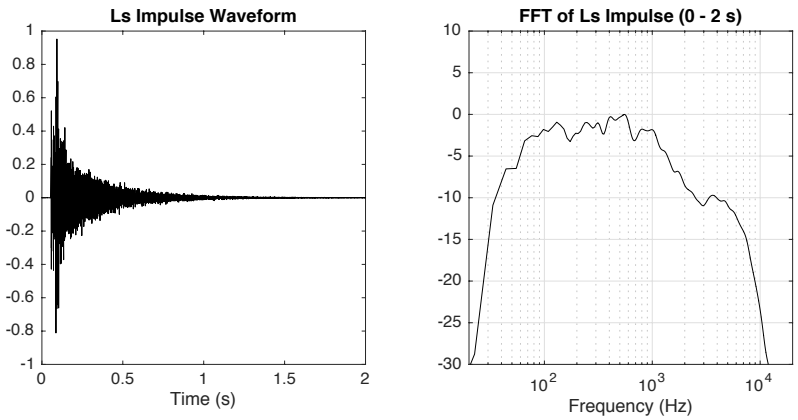

Figure 7.2 Waveform and impulse response of the 'Ls' impulse from the Hamasaki-Square array.

Table 7.1 Loudspeaker channel routing for multichannel room impulse response stimuli.

Format	Loudspeakers	OCT (Direct Signal)	Hamasaki-Square (Ambient Signal)
5.1 (2D)	Centre (C)	Centre	
9.1 (3D)	Left (L)	Left	Front Left
	Right (R)	Right	Front Right
	Left Surround (Ls)		Rear Left
	Right Surround (Rs)		Rear Right
9.1 (3D)	Left Height (LH)		Front Left (Decorrelated)
	Right Height (RH)		Front Right (Decorrelated)
	Left Surround Height (LsH)		Rear Left (Decorrelated)
	Right Surround Height (RsH)		Rear Right (Decorrelated)

7.2.2.2 MRIR Convolution with Anechoic Stimuli

The six anechoic sources used for the MRIR convolution were samples of a Cello, Drumkit, Acoustic Guitar, String Quartet, Male Speech and Trumpet, waveforms of which are presented in Figure 7.3 below. With the addition of the trumpet, these were the same complex anechoic samples used during the high-pass decorrelation experiment in Chapter 6 (for reasons discussed in Section 6.2.2). The Trumpet was included as a continuous condition, since pink noise would not be suitable for the assessment of 2D-to-3D upmixing under practical conditions.

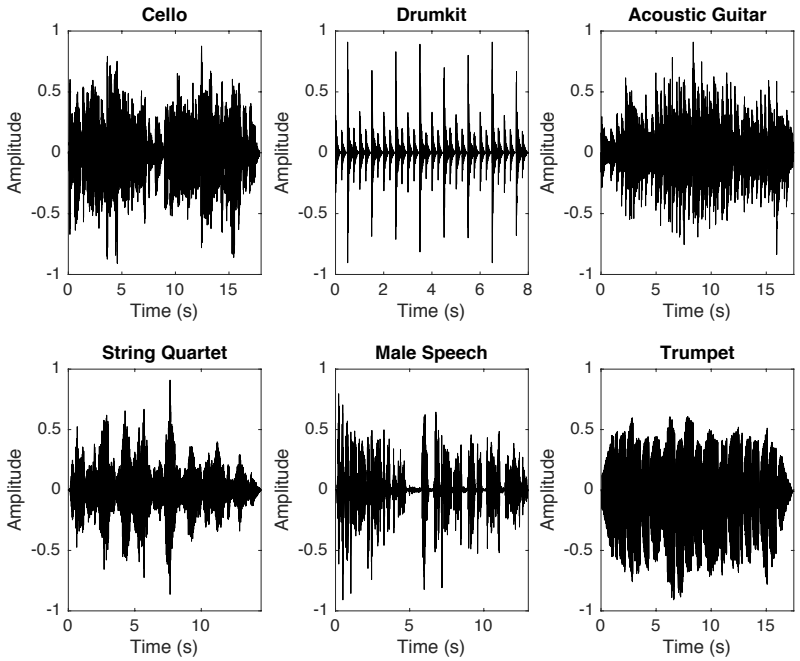

Figure 7.3 Waveforms of the raw anechoic stimuli prior to convolution.

There are inherent time-delays between each impulse response of the MRIRs, due to the different distances from the source to the microphone receivers in the concert hall. These were retained for all conditions during convolution, as would be the case in the mixing of surround recordings, which also helps to reduce the precedence effect (Litovsky, 1999) (i.e. a slight delay for the ambient signals from the direct signals (~14-23ms)).

Convolution was performed in MATLAB and long-term average FFTs of both the OCT-convolved direct signal (C channel) and a HS-convolved ambient signal (Ls Channel) are displayed for each source in Figure 7.4 – FFTs were calculated with 4096 FFT-points and a frame length of 4096 samples, with 50% overlapping windows and 1/6th-octave spectral smoothing. Comparing the direct and ambient FFT plots, it is generally seen that the ambient signals have less high frequency energy than the respective direct signals above around 1 kHz, and more low frequency energy than the direct signals below around 100 Hz. Chapters 4, 5 and 6 of the present thesis indicate that octave-bands of 500 Hz and above are most important to VIS perception by vertical decorrelation, particularly around the 8 kHz band. Given this, it is possible that the reduction of high frequency content seen in the ambient signals may have an impact on the effectiveness of vertical decorrelation; however, it is thought that the configuration used here would most accurately represent a typical upmixing scenario.

7.2.2.3 'Real-Life' Stimulus

Another source used during testing was an extract from a professional ensemble recording of The Debussy Trio, providing a 'real-life' example of how decorrelation might be used to upmix existing content. The Debussy Trio sample featured a pair of Left and Right direct signal audio tracks and a pair of Left and Right ambient signal tracks – waveforms and FFTs of which are displayed in Figure 7.5 below. In the FFT plots, a similar trend to the MRIR-convolved stimuli is seen, where the ambient signals have slightly less high frequency energy and slightly more low frequency energy than the direct signals.

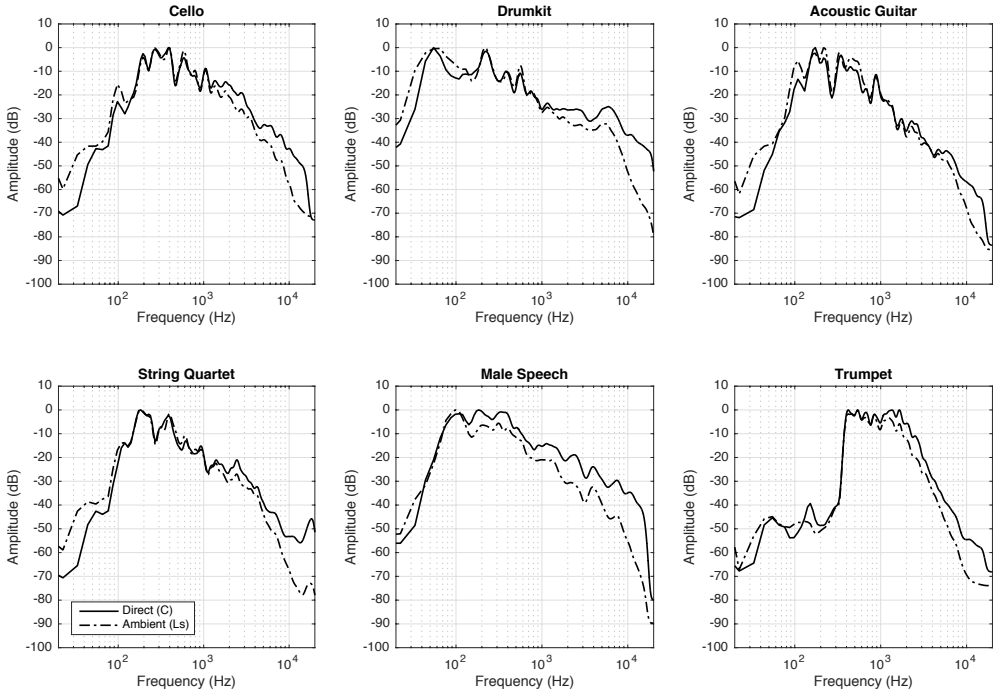

Figure 7.4 Long-term average FFT over time of the convolved direct (C) and ambient (Ls) stimuli signals. Calculated with 4096 FFT-points and a 4096 sample frame length, with 50% overlapping windows and 1/6th-octave spectral smoothing.

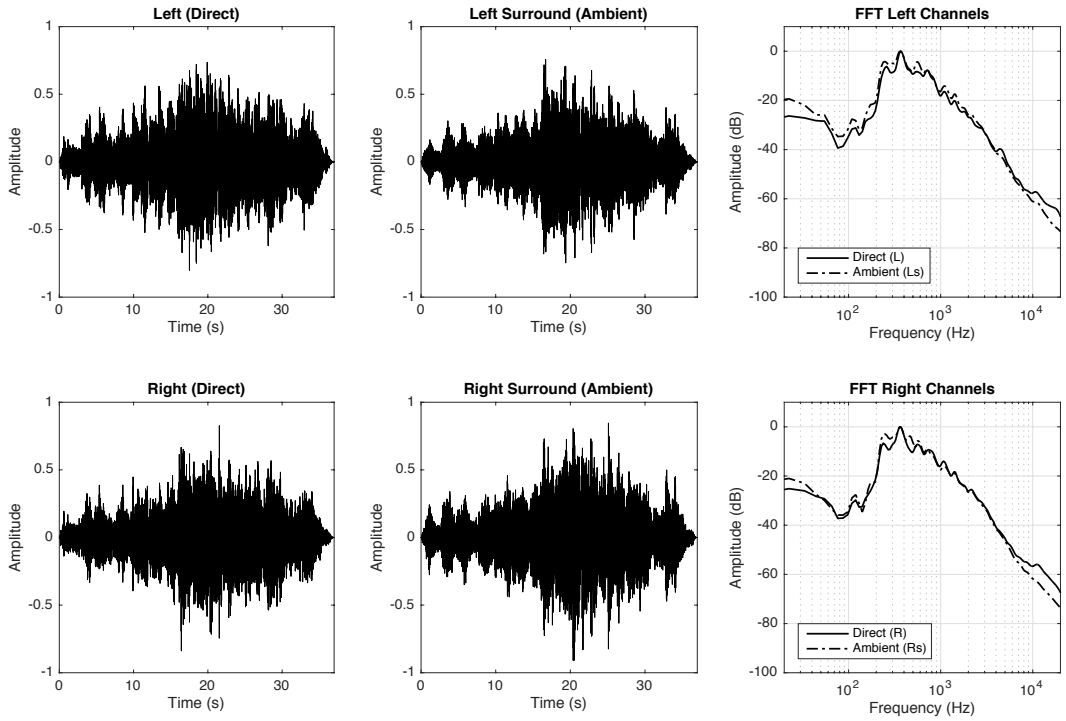

Figure 7.5 Waveforms and FFTs of the Debussy Trio sample. FFTs were calculated with 4096 FFT-points and a 4096 sample frame length (50% overlapping windows), with 1/6th-octave smoothing.

For the current test, the direct signals from The Debussy Trio sample were routed to the front Left (L) and Right (R) loudspeaker channels, and the ambient signals to the Left Surround (Ls) and Right Surround (Rs) (with vertical decorrelation / upmixing into the rear surround height-channels only) – the Centre (C) channel and front L & R height-channels were not used at all. Since vertical decorrelation was only performed at the rear, it provides a realistic upmixing example of where ambient signals are only available from the surround channels. Furthermore, as it was a real recording, the instruments are panned broadly across the frontal image between the L & R channels – whereas with the anechoic samples, the direct sound was focused at the C channel – this provides another interesting factor that may affect spatial perception.

7.2.2.4 Vertical Interchannel Decorrelation (Upmixing) Techniques

Combining the six MRIR-convolved samples (Section 7.2.1.3) with the 'real-life' Debussy Trio sample (Section 7.2.1.4), a total of seven complex sources have been assessed, consisting of transient, continuous, monophonic and polyphonic stimuli. The ambient signals of each source have been vertically decorrelated between each main-layer loudspeaker channel and its respective height-channel pair – in the case of the MRIR-convolved stimuli, this was the four Hamasaki-Square ambient signals (L, R, Ls and Rs), and for the Debussy Trio sample, it was the two rear surround ambient signals (Ls and Rs).

Three decorrelation methods have been implemented and compared in the study. The first method is the all-pass filter phase randomisation method proposed by Kendall (1995), as featured in Chapters 4 and 6 of this study. Full details of the process and implementation can be found in Section 4.2.1.3. As with the previous experiments, two random numbers sequences of length 30 ms were generated as the all-pass filter coefficients. Only two all-pass filters were used for each source: one for the main-layer loudspeakers and another for the height-layer loudspeakers. This was to preserve the original horizontal correlation between the ambient signals, so that only the effect of vertical decorrelation between the two layers is observed (as discussed in Section 7.2.1.4 below). From here on, this method shall be referred to as 'KP'.

The other two decorrelation techniques used in the current experiment were developed more recently by Zotter and Frank (2013). One of the approaches changes the phase of the input signal, as with the all-pass filter method. The other creates regularly opposing amplitude differences between the two channels, similar to that of the complementary comb-filtering method used in Chapters 3 and 4 (Breebaart & Faller, 2007). In the remainder of this chapter, the phase-based of the two methods is referred to as 'ZP', and the amplitude-based is referred to as 'ZA'. Both methods are based on a delay network, of which weighting coefficients are derived from Bessel functions of the first kind (Figure 7.6). The delay networks of 'ZP' and 'ZA' can be seen in Figures 7.7 and 7.8, respectively. In the delay networks, 'g_x' refers to the integer order of the Bessel function (represented graphically in Figure 7.6) and 'N' is the time-delay between the taps. Decorrelation is controlled through the modulation depth and tap-delay – modulation depth is along the x-axis in Figure 7.6, and the weightings for each tap are obtained from the values of each Bessel function order at the chosen modulation depth. The 'g_x' weighting at each tap is then multiplied by either +1 or -1, depending on whether the decorrelation is phase-based or amplitude-based (as per Figures 7.7 and 7.8).

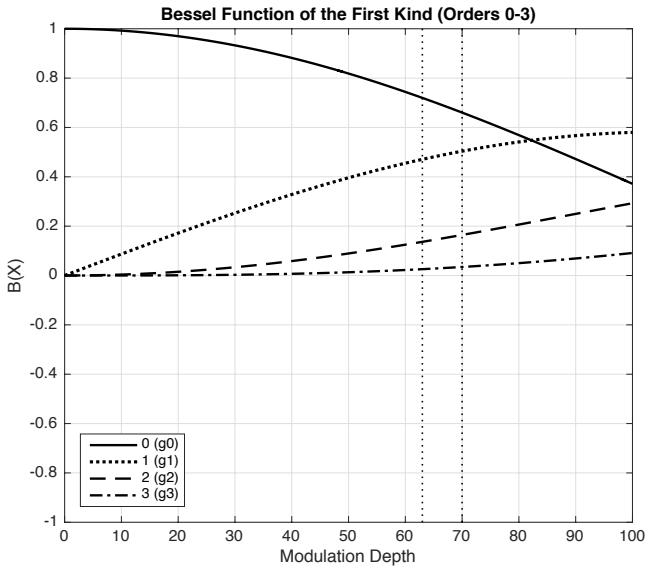

Figure 7.6 Bessel functions of the first kind with integer orders of 0-3. The vertical dotted lines at 63 and 70 are the modulation depths used for the amplitude- and phase-based decorrelation, respectively.

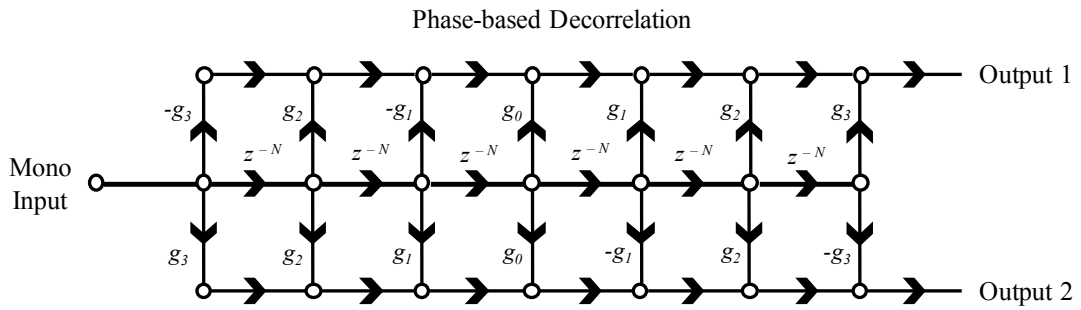

Figure 7.7 Phase-based decorrelation delay-network by Zotter and Frank (2013), where 'g_x' is the Bessel function order coefficient weighting and 'N' is the time-delay between taps.

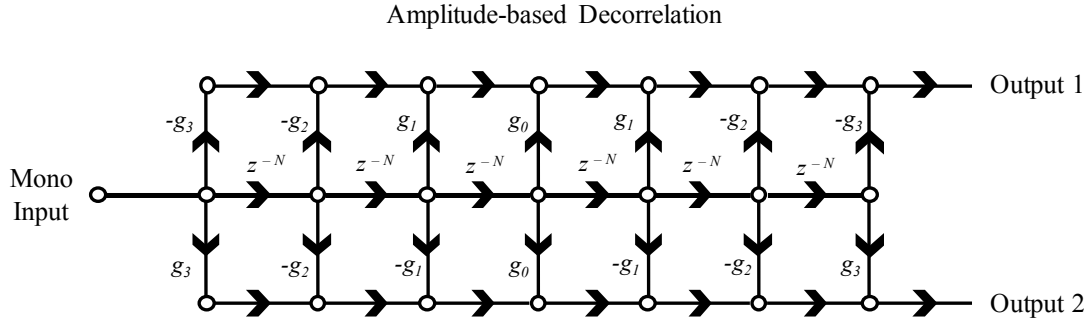

Figure 7.8 Amplitude-based decorrelation delay-network by Zotter and Frank (2013), where 'g_x' is the Bessel function order coefficient weighting and 'N' is the time-delay between taps.

The time-delay and modulation depths of both methods were selected through informal testing and ICCC calculation using the seven different sources. For the phase-based approach, a tap-delay of 2.5 ms and phase modulation depth of 70 was decided upon; and for the amplitude-based method, the parameters were set to a 2.5 ms tap-delay with an amplitude modulation depth of 63. Both of these were chosen as a compromise for achieving the maximum possible decorrelation across all seven sources, while maintaining a consistent time-delay and modulation depth throughout. The resulting four tap weightings of each method can be seen in Table 7.2 below, and the impulse responses of both the 'ZP' and 'ZA' filters are shown in Figure 7.9, alongside an example of the all-pass filter responses for the two channels of the 'KP' method. From observation of Figure 7.9, it is thought that the 'ZA' and 'ZP' approaches may improve the TQ of decorrelation in comparison to the KP method, particularly with regard to transient smearing (Laitinen et al., 2011).

Table 7.2 The coefficient weightings used during testing for each tap of the 'ZP' and 'ZA' delay networks, as derived from the Bessel functions in Figure 7.5 (based on the set modulation depth (MD)).

Bessel Function Order	Phase-based (ZP) (MD = 70)	Amplitude-based (ZA) (MD = 63)
0 (g_0)	0.6690	0.7280
1 (g_1)	0.4994	0.4656
2 (g_2)	0.1603	0.1326
3 (g_3)	0.0332	0.0245

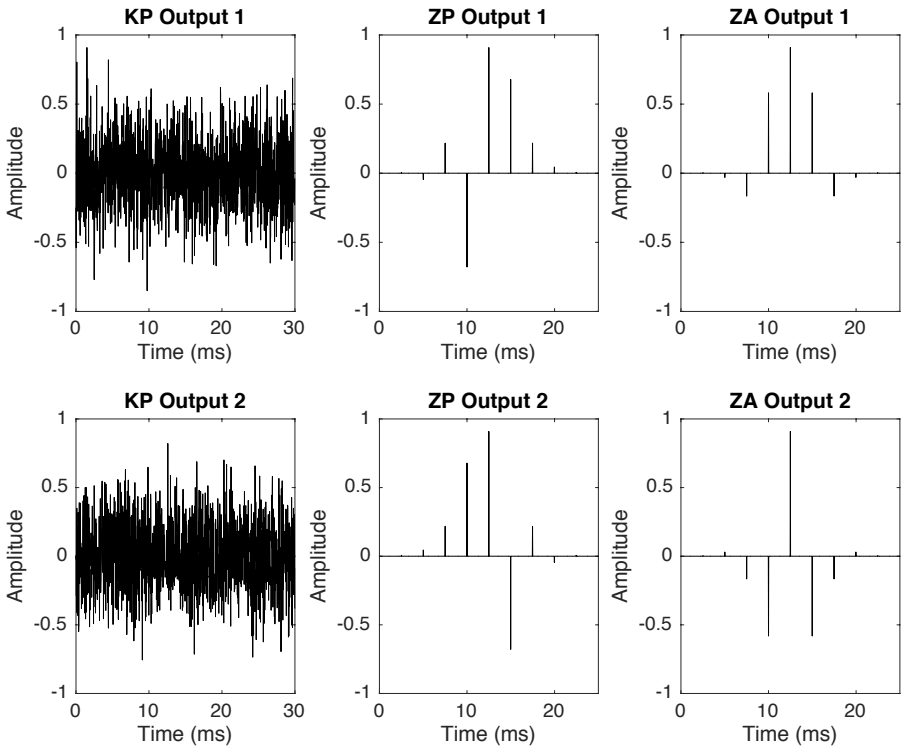

Figure 7.9 Normalised impulse responses of the decorrelation filters for the KP, ZP and ZA decorrelation methods, showing Output 1 on the top row and Output 2 on the bottom.

Chapters 4, 5, and 6 of the present thesis suggest that decorrelating lower frequencies (< 500 Hz octave-band) has little effect on vertical image spread (VIS). Therefore, in addition to broadband vertical decorrelation of the ambient signals, a high-pass decorrelation condition has also been included for each decorrelation method ('KP', 'ZP' and 'ZA'). With the high-pass condition, only frequencies of the 500 Hz octave-band and above have been decorrelated, in order to

observe whether decorrelation of low frequencies is necessary in 2D-to-3D upmixing. When creating the stimuli, the respective low-passed frequencies below the 500 Hz octave-band were routed to the lower main-layer loudspeaker only – that is, the 63 Hz, 125 Hz and 250 Hz octave-bands were reproduced monophonically at ear height and not decorrelated vertically. These bands were also amplified by +3 dB, in order to maintain the original spectral distribution when the decorrelated main- and height-layer signals are summed at the ear. From here on, the high-pass decorrelation conditions are referred to as 'KP500', 'ZP500' and 'ZA500', representing the three decorrelation methods (as described above).

Figure 7.10 below displays decorrelation of a broadband pink noise signal using the six approaches described above ('KP', 'KP500', 'ZP', 'ZP500', 'ZA' and 'ZA500'). The first two rows show the main- and height-layer decorrelated signals, respectively, and the bottom row is the sum of these decorrelated signals. Given the random nature of the Kendall all-pass filter method ('KP' and 'KP500'), the spectra displayed here are only an example of the distortions that can occur, as each generation of the all-pass filters causes random and unique frequency distortion. On the other hand, the 'ZP', 'ZP500', 'ZA' and 'ZA500' methods use the same filters each time (based on the delay-networks above), and consequently feature the same frequency distortion with each implementation (as displayed in Figure 7.10). Looking at the summed signals in Figure 7.10, it is seen that all decorrelation methods seem to suffer from greater distortion at lower frequencies, whereas for the high-pass conditions, this distortion is noticeably decreased. The spectra presented here suggest that the high-pass condition may improve the TQ for all decorrelation conditions, as was previously demonstrated in Chapter 6.

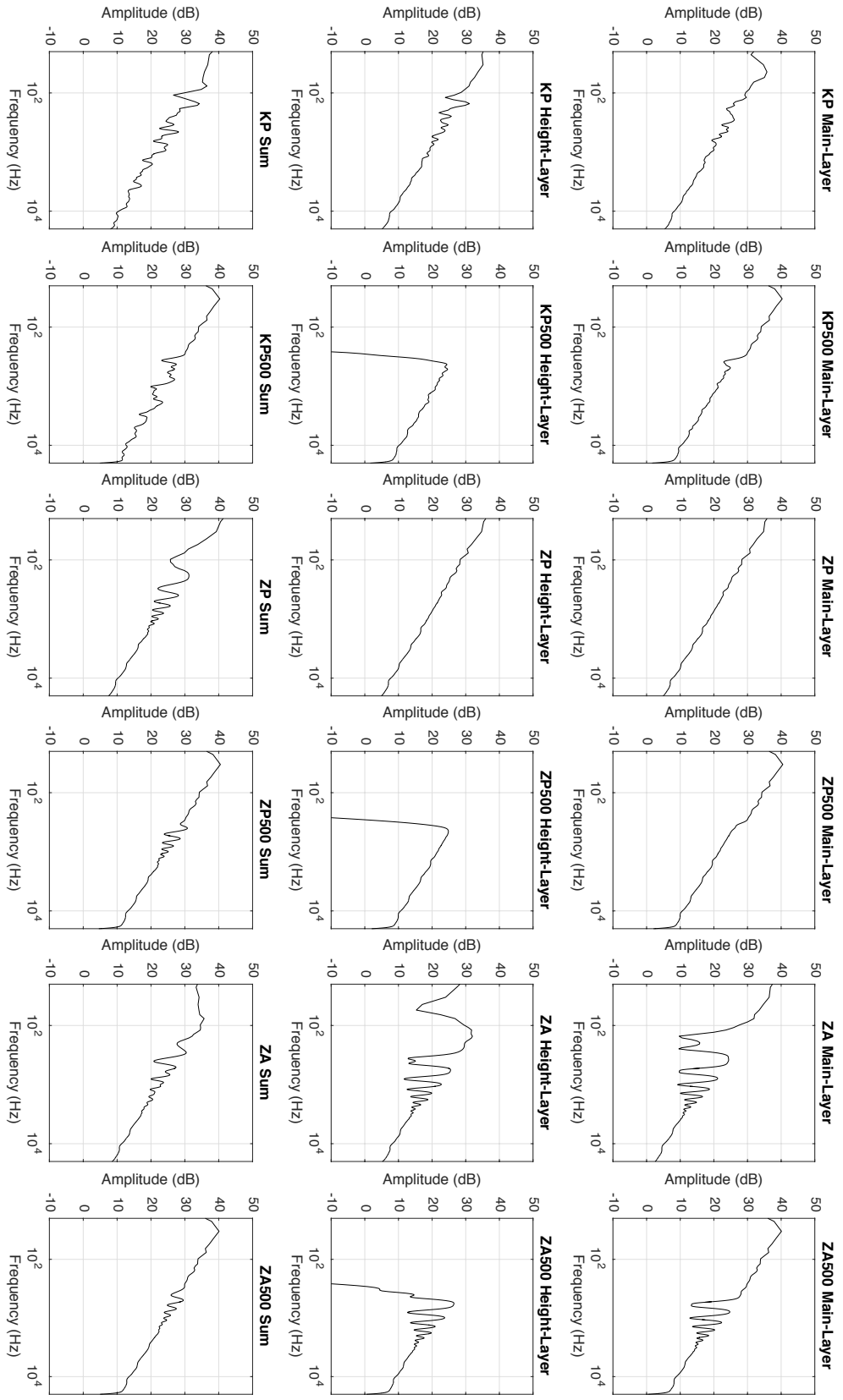

Figure 7.10 Long-term average FFTs of the output signals for each decorrelation condition, using broadband pink noise as the input. Each column features one decorrelation condition, where the top two rows are the two decorrelated outputs, and the bottom row is the sum of the two outputs.

7.2.1.5 Interchannel Cross-Correlation (ICC)

With all three decorrelation methods, the exact same decorrelation filters or settings were applied to each of the four loudspeaker pairs (L, R, Ls and Rs). This was in an attempt to preserve the original horizontal spatial image of the source signal, by only reducing the interchannel cross-correlation coefficient (ICCC) vertically between the main- and height-layer signals. For instance, if a unique all-pass filter were applied to every channel, it would result in horizontal decorrelation as well as vertical, likely increasing the horizontal image spread (HIS) (Zotter & Frank, 2013). The aim of this study is to observe the effect of vertical decorrelation on LEV and TQ; therefore, it is important to maintain the same ICCCs between the main-layer channels, as changes to HIS could potentially bias the perception of LEV – Berg and Rumsey (2001) report that width extension can sometimes be interpreted as an increase of LEV. The original horizontal ICCCs between the L, R, Ls and Rs convolved Hamasaki-square signals (i.e. the original ambient signals before decorrelation) and the two ambient Debussy Trio signals are presented in Table 7.3.

Table 7.3 Interchannel cross-correlation coefficients (ICCCs)

	L / R	**Ls / Rs**	**L / Ls**	**R / Rs**	**L / Rs**	**R / Ls**
Cello	0.45	0.36	0.27	0.26	0.27	0.26
Drumkit	0.58	0.48	0.24	0.28	0.27	0.26
Guitar	0.48	0.33	0.19	0.27	0.24	0.23
Quartet	0.42	0.36	0.24	0.24	0.29	0.23
Speech	0.55	0.46	0.18	0.17	0.18	0.18
Trumpet	0.42	0.32	0.35	0.38	0.34	0.34
Debussy	N/A	0.30	N/A	N/A	N/A	N/A

Table 7.3 displays the natural decorrelation between the raw ambient signals used during testing, where the ICCC between the front L & R channels is around 0.4-0.6 for the Hamasaki-Square signals, and around 0.3-0.5 between the rear Ls and Rs channels for all sources. It was thought that similar levels of ICCC should also be achieved when decorrelating between the vertical pairs of channels in the present experiment. The resulting vertical ICCCs of each source and condition from decorrelation are presented in Table 7.4 below. Since the same decorrelation filters and settings were applied to the four vertical pairs within a condition (L, R, Ls and Rs),

a similar ICCC value was calculated for each of them; therefore, the values shown in Table 7.4 are the average ICCC of the four pairs.

Table 7.4 Vertical interchannel cross-correlation coefficients (ICCCs) of the decorrelated ambient signals for the L, R, Ls and Rs vertical pairs (taken as the average of the four pairs), where 'BB' is broadband decorrelation, '500+ Only' is high-pass decorrelation and '500+ w/ Low' is high-pass decorrelation with the low frequencies routed to one of the channels (the high-pass test condition).

	Kendall Phase (KP)			**Zotter Phase (ZP)**			**Zotter Amplitude (ZA)**		
	BB	**500+ Only**	**500+ w/ Low**	**BB**	**500+ Only**	**500+ w/ Low**	**BB**	**500+ Only**	**500+ w/ Low**
Cello	0.26	0.24	0.18	0.32	0.38	0.30	0.34	0.33	0.24
Drumkit	0.23	0.25	0.07	0.21	0.19	0.07	0.20	0.12	0.04
Guitar	0.26	0.25	0.16	0.21	0.23	0.16	0.22	0.23	0.16
Quartet	0.23	0.22	0.12	0.33	0.35	0.26	0.25	0.26	0.19
Speech	0.25	0.22	0.09	0.22	0.22	0.13	0.36	0.16	0.08
Trumpet	0.26	0.22	0.21	0.38	0.31	0.30	0.45	0.36	0.34
Debussy	0.27	0.30	0.25	0.31	0.29	0.27	0.34	0.26	0.22

It was more difficult to achieve consistently low levels of ICCC for some sources than others. However, every source and condition achieved at least a vertical ICCC of 0.5, with the vast majority ranging between 0.2 and 0.4. The Trumpet source was particularly hard to decorrelate, which may relate to its continuous and relatively narrow-band nature (see Figures 7.3 and 7.4). In general, lower levels of correlation appear more achievable for signals with transient information (e.g. the Drumkit and Guitar), although it is known that decorrelation of such signals can have a detrimental effect on tonal quality through transient smearing (Laitinen et al., 2011).

7.2.2.6 SPL and Ambience to Direct Sound Ratio (A/D)

When combining the direct and ambient components, the decorrelated ambience for each source and condition were SPL level-matched to 72 dB LAeq, with the OCT direct sound set to 69 dB LAeq – this provided an ambience to direct sound ratio (A/D) of +3 dB. Given that The Debussy Trio excerpt only featured ambience from the rear (Ls and Rs), the direct sound was set at 72 dB LAeq, in order to balance the front-back SPL (D/A = 0 dB), while keeping the ambience level consistent with the other sources. Along with the six decorrelated stimuli (KP, KP500, ZP, ZP500, ZA and ZA500), three further control conditions were included in the

multiple comparison trials as follows (where the OCT direct sound is set to 69 dB LAeq and the Debussy direct sound is set to 72 dB LAeq):

1) "Lower" – the original ambient signals in the lower main-layer only, level-matched with the decorrelated ambient signals (72 dB LAeq; A/D = +3 dB for the HS-convolved stimuli and 0 dB for The Debussy Trio).
2) "Lower -3dB" – the original ambient signals in the lower main-layer only, attenuated by 3dB against the decorrelated ambient signals (69 dB LAeq), giving a A/D ratio of 0 dB for the convolved stimuli and -3 dB for The Debussy Trio.
3) "Upper" – the original ambient signals routed to the upper height-channels only, level-matched with the decorrelated stimuli (72 dB LAeq; A/D = +3 dB for the convolved stimuli and 0 dB for The Debussy Trio).

The above SPL LAeq values for each component and condition are summarised in Table 7.5 below, separated by the Hamasaki-Square (HS-convolved) stimuli and the stimuli of The Debussy Trio sample.

Table 7.5 Sound pressure level (SPL) (dB LAeq) of the direct and ambient sound components for both the Hamasaki-Square convolved stimuli and The Debussy Trio sample

	Direct Sound	Lower Only Ambience	Lower Only -3dB Ambience	Upper Only Ambience	Decorrelated Ambience
HS-Convolved	69 dB	72 dB	69 dB	72 dB	72 dB
Debussy Trio	72 dB	72 dB	69 dB	72 dB	72 dB

An A/D of +3 dB for the decorrelated conditions was chosen through informal listening – it was found that little-to-no change was perceived between any conditions when A/D was 0 dB, due to the dominance of the direct sound. A +3 dB A/D is also the equivalent of adding four decorrelated ambient height-channels to a 5.1 surround sound recording where the A/D is 0 dB (i.e. doubling the number of ambient signals) – this represents a possible upmixing scenario where ambience energy normalisation is not present. As one might expect, the higher the ambience level, the easier it was to perceive differences from decorrelation. However, the intention

was to keep the experiment as practical and realistic as possible, so a +3 dB increase of ambience was considered a suitable compromise.

7.2.3 Testing Procedure

Testing was conducted over two sittings of 20-25 minutes each, split by the two attributes under testing (Listener Envelopment (LEV) and Tonal Quality (TQ)). Each session consisted of 7 multiple-comparison trials, one for each of the seven sound sources (Table 7.6), based on an adaptation of MUSHRA (ITU-R, 2015b) (as discussed in Section 3.2.4). The testing interface was constructed using HULTI-GEN, a Max tool that was developed by the author (Gribben and Lee, 2016). In a single trial, subjects were asked to grade the perceived LEV or TQ on a bipolar scale of -30 to 30, comparing the nine different ambience conditions (described above and summarised in Table 7.6 below). Stimuli were compared relatively against each other and a reference at 0 in the centre of the scale (Figure 7.11). The reference stimulus was the 'Lower' 5.1 condition, where ambient signals were routed to the lower main-layer loudspeakers only and SPL level-matched with the decorrelated stimuli.

Similar to the procedure of the TQ experiment in Chapter 6, subjects were further instructed to report descriptive terms for the stimuli they perceived as having the best/greatest and worst/least TQ and LEV. A list of potential LEV and TQ terms and their definitions were provided to each listener, in order to give them an idea of attributes to listen for. For the LEV part, the suggested terms were 'Narrow / Wide', 'Vertically Spread', 'Enveloping / Flat', 'Dry / Reverberant', 'Deep / Shallow', 'Spacious' and 'Full / Thin'; and the potential terms for TQ were 'Clear / Muddy', 'Natural / Unnatural', 'Phasiness', 'Full / Thin', 'Hard / Soft', 'Bright / Dark', 'Loud / Quiet' and 'Distortion'. It was made clear that the descriptors provided were only a guide for listeners, and written responses were not limited to those listed. The aim of this was to give further insight into which aspects of LEV and TQ were typically influencing subjects' judgements, as well as to provide an informal understanding of how decorrelation and ambience level affect the perception of each attribute.

Table 7.6 Summary of the seven source signals for the seven trials of each attribute and a summary of the nine ambience stimuli conditions within each multiple comparison trial

Source Signals	Ambience Stimuli Conditions
Cello	Lower Only (Reference)
Drumkit	Lower Only -3dB
Acoustic Guitar	Upper Only
String Quartet	Kendall Phase (KP)
Male Speech	Kendall Phase 500+ (KP500)
Trumpet	Zotter Phase (ZP)
The Debussy Trio	Zotter Phase 500+ (ZP500)
	Zotter Amplitude (ZA)
	Zotter Amplitude 500+ (ZA500)

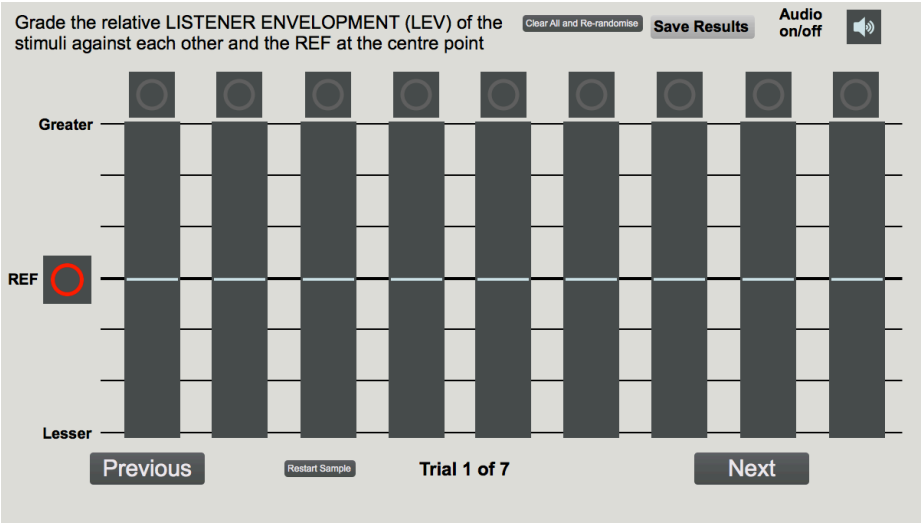

Figure 7.11 The listening test interface used during testing, comparing 9 stimuli on a bipolar scale.

7.2.4 Subjects

A total of 18 subjects participated in the listening tests, all of who reported normal hearing. The same 18 listeners took part in both the LEV and TQ testing sessions, consisting of staff and students from the University of Huddersfield that are experienced in critical spatial listening exercises. A sample size of 18 achieves a statistical power of 0.67 for this experiment, based on a two-tailed t-test with an effect size of 0.6 and α error probability of 0.05 (type I error i.e. false-positive), as calculated using G*Power 3.1 (Faul et al., 2007). This indicates that the probability of a type II statistical error (false-negative) is 0.33 (β). All subjects were instructed to remain still and forward-facing throughout the judgement and grading of stimuli, which was ensured by the use of a small headrest.

7.3 Experiment Part 1: Listener Envelopment (LEV) Results

Results from the Listener Envelopment (LEV) part are displayed in Figure 7.12 below, where the medians of the subjects' scores are presented with 95% confidence notch edge bars (McGill et al., 1978). All scores were normalised in accordance with ITU-R BS.1116-3 (ITU-R, 2015a) (as described in Section 3.3) and the subsequent analysis was performed in SPSS. Shapiro-Wilk normality tests indicated that the data for each condition was not always normally distributed; therefore, all groups of data have been compared using the non-parametric Wilcoxon signed-rank test with Bonferroni correction. In the graphs and following results, 'Lower' is ambience presented in the main-layer only (the level-matched 'Reference'), 'Lower-3' is ambience in the main-layer only attenuated by -3 dB, and 'Upper' is ambience in the height-layer channels only (level-matched). For the decorrelated stimuli, 'KP' refers to the Kendall all-pass filter method, and 'ZP' and 'ZA' refer to the Zotter phase and amplitude methods, respectively. 'KP500', 'ZP500' and 'ZA500' indicate the use of high-pass decorrelation (of the 500 Hz octave-band and above) with each method.

A tally of the written responses for stimuli with the greatest and least LEV are presented in the Tables 7.7 and 7.8, respectively. These responses are intended to provide a general indication of the reasoning behind judgements, rather than be a statistical assessment of semantic terms related to LEV. For the sample with the greatest LEV, it is seen that the most common term used was 'Reverberant' (32 occurrences), which was defined as the ratio of ambience energy to direct sound energy i.e. the level of ambience. Other relatively frequent terms included 'Wide', 'Full', 'Spacious', 'Deep' and 'Enveloping' (20-29 occurrences each). In contrast, 'Vertical Spread' (the assumed effect of vertical decorrelation) was only used to describe the greatest stimuli 12 times. These results suggest that the perception of LEV may relate most to the perceived level of ambience and low frequencies, along with other lateral / horizontal spatial attributes, rather than the perception of vertical spread. If vertical decorrelation does generate a greater vertical image spread in the application of 2D-to-3D upmixing, it appears that the

effect may be relatively weak and less influential in the grading of LEV, compared to the other descriptive terms mentioned here.

Table 7.7 Summary of the qualitative responses for the sample with the greatest listener envelopment (LEV)

	Cello	Drumkit	Acoustic Guitar	String Quartet	Male Speech	Trumpet	Debussy Trio	Total
Reverberant	4	9	6	3	7	2	1	**32**
Wide	6	3	7	4	3	5	1	**29**
Full	5	4	2	6	1	2	6	**26**
Spacious	4	5	1	4	2	4	2	**22**
Deep	3	5	2	3	6	1	2	**22**
Enveloping	4	2	1	3	3	2	5	**20**
Vertical Spread		1	2	2	2	2	3	**12**
Open		1	1		1	1		**4**
Diffuse	1	1						**2**
Spread	1							**1**
Distance		1						**1**
Bright		1						**1**
Resonant		1						**1**
Side Spread							1	**1**

Table 7.8 Summary of the qualitative responses for the sample with the least listener envelopment (LEV)

	Cello	Drumkit	Acoustic Guitar	String Quartet	Male Speech	Trumpet	Debussy Trio	Total
Narrow	11	8	10	8	10	7	9	**63**
Thin	8	4	5	4	3	3	8	**35**
Dry	5	6	7	2	7	5	1	**33**
Flat	2	3	1	2	2	2	1	**13**
Shallow	1		1	5		4	2	**13**
Quiet	1	1	1	1	2	2	1	**9**
Focused		1	1		2	1		**5**
Less Enveloping	1			2	1			**4**
Dense		1						**1**
Artificial						1		**1**

Looking at the responses for the sample with the least LEV in Table 7.8, the most frequent term reported was ‘Narrow’ (65 occurrences), which refers to the width of the auditory image and was defined as the opposite of ‘Wide’ (seen in Table 7.7). The second and third most common descriptors were ‘Thin’ (opposite of ‘Full’) and ‘Dry’ (opposite of ‘Reverberant’), with 35 and 33 occurrences, respectively. It is clear that a narrow horizontal width had the greatest impact on negative LEV perception – this could potentially relate to the level of ambience, where less ambient energy would have given more focus to the direct sound in the centre channel (thus relating ‘Narrow’ to ‘Dry’ and ‘Focused’ as well). Furthermore, as with the positive responses,

both 'Thin' and 'Dry' also suggest that LEV may be largely influenced by the level of low frequency energy and ambient energy.

For the subjective LEV results of the Cello in Figure 7.12 below, the Bonferroni-corrected Wilcoxon tests indicated no significant difference between the 2D ambience reference ('Lower') and any of the 3D upmixed stimuli, when both are level-matched ($p > 0.05$). However, 'Lower-3' had significantly less LEV than 'Lower' and some upmixed conditions ('KP', 'KP500', 'ZA' and 'ZA500') ($p < 0.02$). The most common term to describe 'Lower-3' was 'Narrow', as determined by looking at the individual scores for each subject. This supports the hypothesis that a lower level of ambience may cause the auditory image to become more focused towards the direct sound, resulting in less LEV, and also suggests the importance of ambience level on the perception of LEV. Considering differences between the decorrelation methods, 'ZA' and 'ZA500' both had significantly greater LEV than 'ZP500' ($p < 0.05$). Furthermore, 'KP500', 'ZA' and 'ZA500' also had significantly greater LEV than the 'Upper' condition ($p < 0.05$).

Looking at the Drumkit source, 'ZA' and 'ZA500' had significantly greater LEV than both the 2D reference ('Lower') and 'Upper' conditions ($p < 0.01$). All 3D upmixed stimuli and the 'Lower' reference also had significantly greater LEV than the 'Lower-3' condition ($p < 0.05$). Looking at the grades for each subject, 'Lower-3' was mostly referred to as 'Narrow' and 'Dry', again suggesting the impact ambience level can have on LEV. In terms of differences between decorrelation methods, 'ZA' and 'ZA500' were both significantly greater than 'KP', 'KP500' and 'ZP500' ($p < 0.04$), while 'ZA500' was also significantly greater than 'ZP' ($p < 0.01$) – demonstrating that the perception of decorrelation can differ between techniques. Table 7.8 indicates that the most frequent term to describe the greatest LEV for the Drumkit sample was 'Reverberant', of which all nine occurrences were for either the 'ZA' or 'ZA500' condition (when looking at the individual scores for each listener).

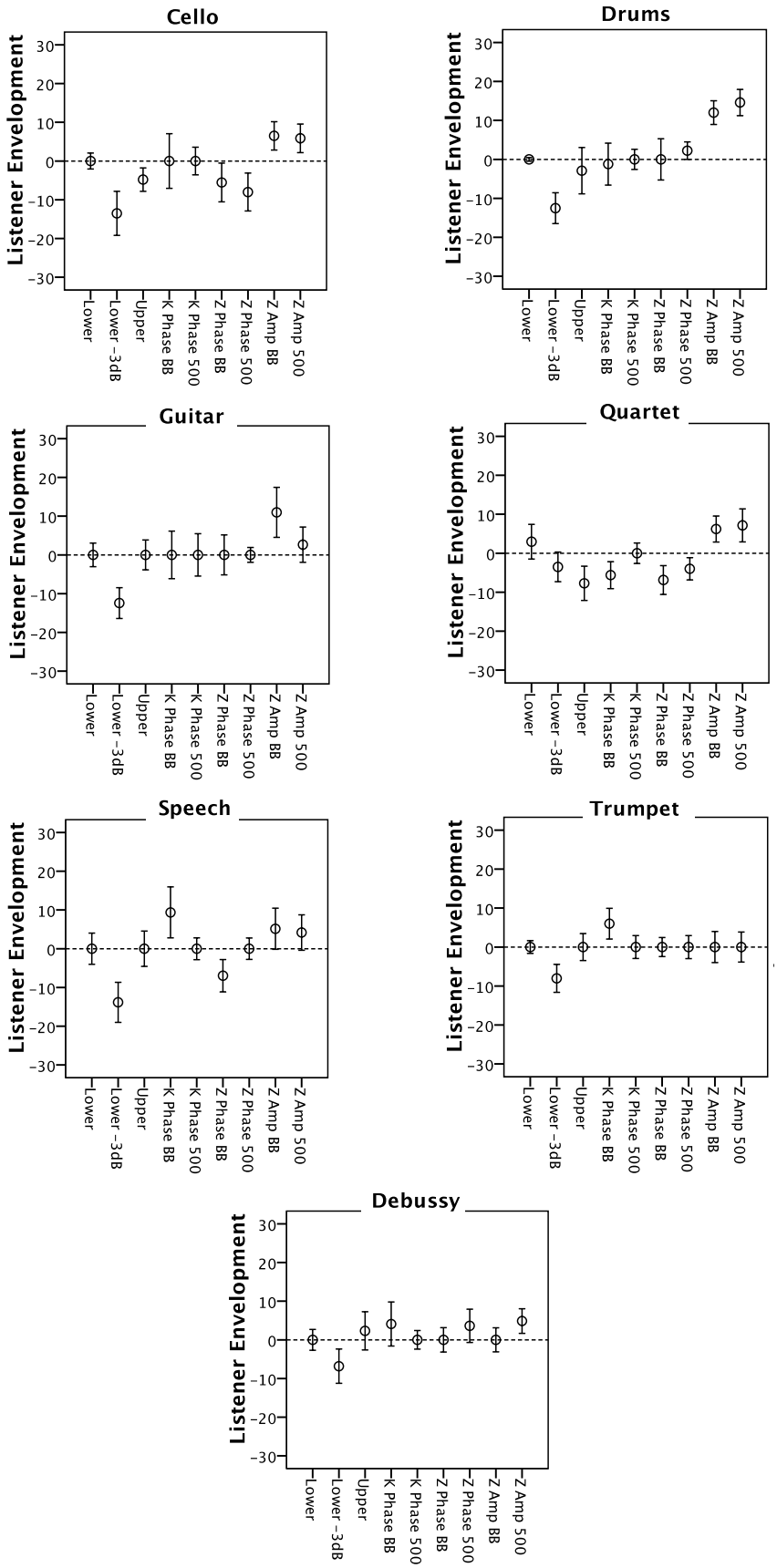

Figure 7.12 Listener Envelopment (LEV) median scores with 95% confidence notch edge bars.

For the Acoustic Guitar, Trumpet and The Debussy Trio sources, the Wilcoxon tests indicated no significant increase or changes to LEV for any of the 3D upmixed conditions, when compared against each other and the 2D 'Lower' reference ($p > 0.05$). However, as seen above, the perception of LEV for 'Lower-3' was significantly less than the 'Lower' reference (as well as most upmixed conditions) for all three sources ($p < 0.04$). Again, the 'Lower-3' condition was most often described as 'Narrow' and 'Dry' in each case, when looking at the subjects' individual scores. Observing the Acoustic Guitar results in Figure 7.12, a slight increase of LEV is also seen for the 'ZA' condition, though not to a significant level ($p > 0.05$). Similarly, in the Trumpet results, the 'KP' LEV score is slightly greater than other methods, but not significantly so ($p > 0.05$). This variation of results for different methods suggests that vertical interchannel decorrelation perception is source-dependent, and one single method may not be suitable for all applications.

For the Male Speech sample in Figure 7.12, the 'Lower' reference had significantly greater LEV than 'Lower-3' ($p < 0.04$) – this further supports the idea that the perception of LEV is largely level dependent, with comments mostly referring to the 'Lower-3' condition as 'Narrow' (when looking into the results for each listener). There was also no significant difference between the 2D 'Lower' reference and all 3D upmixed stimuli ($p > 0.05$), with the only significant difference between upmixed methods seen for 'ZA500', which was significantly greater than 'ZP' ($p < 0.04$).

Unlike other sources, the String Quartet 2D 'Lower' reference was actually perceived as having significantly greater LEV than the 3D 'ZP' upmixed condition ($p < 0.04$), with no significant difference between the reference and other upmixed conditions ($p > 0.05$). Table 7.7 shows that the most common term for the String Quartet samples with the greatest LEV was 'Full', which suggests that stimuli with lesser LEV may have lacked low frequency energy, possibly from phase cancellation. Between decorrelation methods, 'ZA' had significantly greater LEV than

both 'ZP' and 'ZP500' ($p < 0.05$). Furthermore, the 'Upper' condition also had significantly less LEV than the 'Lower', 'ZA' and 'ZA500' conditions ($p < 0.04$).

7.3.1 Discussion of the LEV Results

In general, upmixing by vertical decorrelation seems to have little effect on the perception of listener envelopment (LEV), particularly between level-matched conditions. The only upmixed 3D stimuli to have a significant increase of LEV over the 2D level-matched reference were both the Zotter amplitude-based conditions ('ZA' and 'ZA500') for the Drumkit sample. The results and objective analysis in Chapters 4 and 5 suggest that the perception of vertical image spread (VIS) by vertical decorrelation is largely influenced by higher frequencies, particularly around the 8 kHz octave-band. Table 7.9 below presents the RMS distribution of octave-band energy for an ambient signal of each source (Left Surround (Ls) channel), normalised to 0 dB at the octave-band with the greatest energy. It can be seen that the Drumkit sample has the greatest relative energy in the 8 kHz and 16 kHz bands, which may have improved the perception of vertical decorrelation, resulting in a greater sense of LEV. The Drumkit sample also has the greatest amount of energy in the 63 Hz octave-band, which could contribute further to an increase of LEV – it was seen in the absolute grading of VIS in Chapter 4 that low frequencies have an inherently broad VIS. In contrast, the Trumpet was the only source that had no significant change of LEV with the 'ZA' method – Table 7.9 clearly shows that the Trumpet has the least amount of energy in both low (63-250 Hz) and high (8-16 kHz) octave-bands. These results suggest that the effectiveness of vertical decorrelation may be dependent on the level of low and high frequency energy in the source content.

Table 7.9 Octave-band RMS levels of the convolved ambient stimuli (Left Surround (Ls) channel), normalised to 0dB at the octave-band with the greatest energy for comparison.

	63 Hz	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	16 kHz
Cello	-20.3	-11.1	-0.3	0	-2.7	-5.2	-9.6	-15.6	-26.8
Drumkit	-1.9	-5.4	0	-2.2	-6.9	-8.9	-8.6	-10	-20.9
Guitar	-15.3	-2.9	-1.3	0	-3.5	-8.7	-11.7	-15.3	-28.5
Quartet	-17.6	-3.9	-0.1	0	-4.4	-6.2	-8.2	-17.6	-27.7
Speech	-2.4	-0.9	0	-0.6	-4.2	-7	-7.6	-11.6	-22.3
Trumpet	-29.2	-28.3	-12.2	-0.5	0	-1.7	-10.7	-23.2	-32.2
Debussy	-18.8	-12.4	0	-0.3	-2.6	-6.1	-12.3	-19.8	-25.6

Comparing the three decorrelation methods, the Zotter amplitude-based approach generally appears to be most effective for increasing listener envelopment (LEV). In certain cases, the Cello, Drumkit, String Quartet and Male Speech samples all display a significant increase of LEV with the Zotter amplitude-based method over the other two phase-based techniques, both with broadband and high-pass decorrelation. Although this approach appears to be effective, the differences in amplitude between signals only reconstruct to give a flat spectrum in the central listening position, where the signals are time and level aligned. If the listener were positioned outside of this 'sweet-spot', the reconstruction would be biased toward one of the channels, potentially causing a comb-filtering effect and destruction of the signal at the ears – this is something to consider when using such a method. In terms of the phase-based methods, there is little change to LEV between the 3D upmixed stimuli and the 2D 'Lower' reference. As hypothesised at the beginning of the chapter, it may be that when many similar signals are reproduced simultaneously, differences in phase between two specific channels are difficult or impossible for the hearing system to detect. Considering this, it might be suggested that an amplitude-based method is most suitable when decorrelating signals for high-order 3D multichannel systems. Further investigation is required to determine the effectiveness of phase-based decorrelation methods, in particular, it would be of interest to observe changes from varying the number of decorrelated signals arriving at the ear.

Comparing the perception of LEV between the broadband and high-pass decorrelation conditions, there is no significant difference between the two for any source or decorrelation method. These results agree with those of Chapter 6, suggesting that there is no significant benefit to decorrelating lower frequencies vertically, in terms of increasing spatial perception. Decreasing the number of signals containing lower frequencies could also have a positive impact on maintaining the low-end of a sample, reducing the risk of phase cancellation and loss of frequency – as demonstrated in Figure 7.10, decorrelation appears to cause greater distortion at lower frequencies than higher frequencies.

An interesting finding from this part of the experiment is the influence of ambience level on the perception of LEV. For every source except the String Quartet, the 'Lower-3' condition (ambience from the main-layer only with -3 dB attenuation) had significantly less LEV than the level-matched 'Lower' reference. For the MRIR-convolved stimuli, 'Lower-3' gives an ambience to direct sound ratio (A/D) of 0 dB i.e. equal energy for both components. In this case, if the ambient channel-count were simply doubled by introducing four decorrelated height-channels with no level adjustment / normalisation, the overall ambience level would increase to +3 dB greater than the direct sound (A/D = +3 dB), giving the impression of greater LEV. However, the results here show that when the 2D ambience is amplified to match the ratio of A/D = +3 dB (i.e. the 'Lower' reference), the perceived LEV is also increased to match that of the 3D upmixed stimuli. Given this similarity, it suggests there might not be a reason to upmix ambience into height-channels at all, if simply adjusting the ratio of reverberation to direct sound in the main-layer can control LEV as effectively. In concert hall acoustics, although overhead ceiling reflections can contribute to LEV perception, it is widely accepted that the perception of LEV remains largely dictated by lateral reflection energy (Furuya et al., 1995; Furuya et al., 2001; Furuya et al., 2007). It seems there may still be some benefit of using height-channels in terms of 3D localisation and increasing the audio coverage of large spaces. However, if the additional height-channels are not required for ambience enhancement in smaller listening environments, discarding them may improve comb-filtering and phase cancellation issues, potentially increasing the 'sweet-spot' of the listening area. Of course, the greater the energy of ambience emanating from the height-channel loudspeakers, the greater the influence it will have on LEV and immersion – though this is unlikely to be a realistic representation of an actual concert hall environment.

7.4 Experiment Part 2: Tonal Quality (TQ) Results and Discussion

The Tonal Quality (TQ) results can be seen in Figure 7.13 below, displaying the median scores with 95% confidence notch edge bars (McGill et al., 1978) (normalised as per ITU-R BS.1116-3 (ITU-R, 2015a) described in Section 3.3). All conditions throughout the results and discussion are named the same as in the listener envelopment (LEV) results section. As with the LEV results, the TQ data for each condition was not normally distributed; given this, non-parametric Wilcoxon signed-rank tests were performed between the groups of data for each condition. Results of the Bonferroni-corrected Wilcoxon tests indicate that there are no significant differences in TQ between any of the conditions tested ($p > 0.05$). This suggests that changes of ambience may have an insignificant impact on overall quality when direct sound is present. Having said that, observations of the 95% confidence bars in Figure 7.13 can also suggest some general trends, despite statistical insignificance. For the Cello, String Quartet, Trumpet and The Debussy Trio sources, such observations still demonstrate little-to-no apparent change of TQ across all conditions.

A summary of the qualitative responses for the samples with the best and worst TQ are presented in Tables 7.10 and 7.11, respectively. As with the tallied written responses for LEV, the TQ terms presented here are intended to give a general insight into the differences of TQ that were perceived, rather than provide a statistical analysis of semantic terms for elicitation purposes. The most common TQ descriptor used for the 'best' sample(s) was 'Clear' with 49 occurrences, followed by 'Bright', 'Full' and 'Natural' (22-29 occurrences each). For the 'worst' samples, the most frequent term was 'Muddy' with 38 occurrences, while 'Thin' also featured relatively often (28 occurrences). These results clearly indicate that the interpretation of TQ is largely focused on the clarity of the source. Reference to 'Full' and 'Thin' also suggest potential differences of low frequency energy between conditions, possibly relating to low frequency distortion from decorrelation. Comparing these comments against those for the TQ results in Chapter 6 – where ambient signals were vertically decorrelated between single loudspeaker

pairs only – the term 'Phasey' has notably less occurrences in the present experiment. This could be due to a masking effect from the direct sound signal, or there may be a potential reduction of the 'Phasey' effect at the ear when more signals are introduced around the listener.

Table 7.10 Summary of the qualitative responses for the sample with the best tonal quality (TQ)

	Cello	Drumkit	Acoustic Guitar	String Quartet	Male Speech	Trumpet	Debussy Trio	Total
Clear	9	8	5	7	10	3	7	**49**
Bright	6	4	1	5	3	6	4	**29**
Full	2	4	5	4	2	4	3	**24**
Natural	4	4	4	6		3	2	**23**
Loud	2	1	1	3	2	2		**11**
Balanced	1	1	1	1	1		1	**6**
Focused	1		1		1			**3**
Less Phasey	1			1	1			**3**
Rich		1		1	1			**3**
Warm			2			1		**3**
Defined							2	**2**
More Attack		1						**1**
Hard		1						**1**
More Body		1						**1**
Reverberant		1						**1**
Controlled				1				**1**
Open				1				**1**
Soft							1	**1**
Dry							1	**1**
Bassy							1	**1**

Table 7.11 Summary of the qualitative responses for the sample with the worst tonal quality (TQ)

	Cello	Drumkit	Acoustic Guitar	String Quartet	Male Speech	Trumpet	Debussy Trio	Total
Muddy	4	5	8	6	8	2	5	**38**
Thin	4	4	3	4	2	7	4	**28**
Quiet	4	2	2	2	1	1	2	**14**
Phasey	2	4	1	4	1			**12**
Dark	2	3	1	3	2	1		**12**
Unnatural	1	3		1	3			**8**
Dull	3	1		1		2		**7**
Boomy			2	1	2		1	**6**
Distorted	1	1	2		1			**5**
Soft	1			1		1	1	**4**
Hard	1		1			1		**3**
Too Narrow	1			1		1		**3**
Loud			1		1		1	**3**
Too Wide			1			1	1	**3**
Too Reverberant			2					**2**
Muffled			1		1			**2**
Shallow	1							**1**
Resonant		1						**1**
Slurred					1			**1**
Too Bright						1		**1**
Harsh						1		**1**

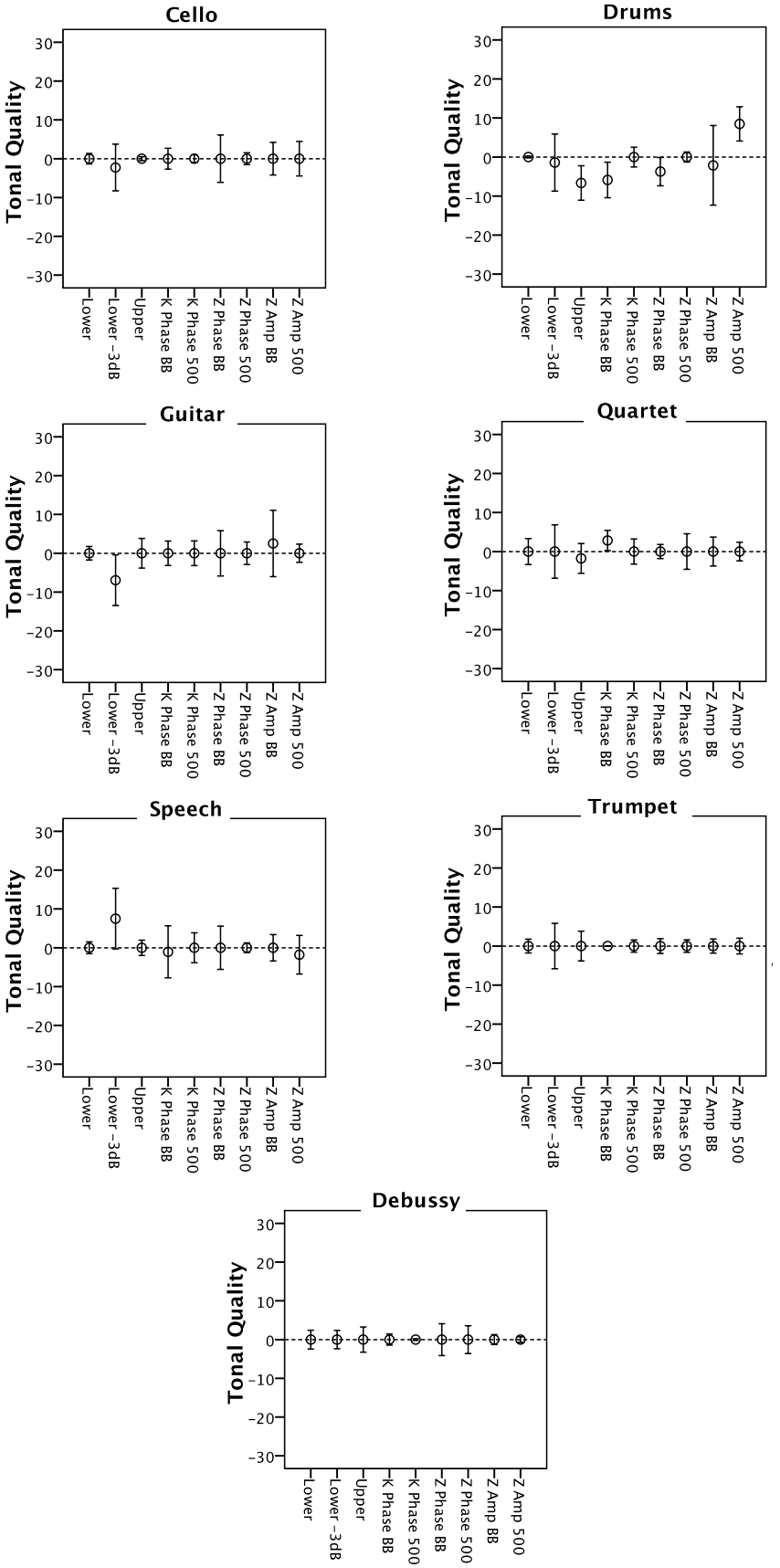
Cello
Drums
Guitar
Quartet
Speech
Trumpet
Debussy
Tonal Quality
30
20
10
0
-10
-20
-30
Lower
Lower -3dB
Upper
K Phase BB
K Phase 500
Z Phase BB
Z Phase 500
Z Amp BB
Z Amp 500

Figure 7.13 Tonal Quality (TQ) median scores with 95% confidence notch edge bars.

Looking at the subjective TQ results for the Drumkit sample in Figure 7.13 above, there appears to be a significant decrease of TQ when using phase-based broadband decorrelation. However, TQ appears to improve slightly when decorrelating just the 500 Hz octave-band and above, resulting in a similar level of quality to the unprocessed reference. Comments for the 'KP' and 'ZP' conditions were both 'Phasey' and 'Thin' (as found when looking at subjects' individual responses), possibly indicating the low frequency distortion when the signals are summed at the ear. The Zotter amplitude-based method also appears to have an improvement of TQ when high-passing the decorrelation, which actually results in a considerably greater TQ than the reference for 'ZA500'. Going back to the octave-band RMS levels in Table 7.9 (Section 7.3.1), it is seen that the Drumkit has the greatest energy in the 63 Hz octave-band compared to the other sources (due to the kick drum), as well as relatively great energy in the 250 Hz octave-band from the snare. Given the slight improvement of TQ seen for the high-pass decorrelation, it suggests that the low frequency distortion seen in Figure 7.10 may have a negative effect on TQ. It is possible that broadband decorrelation noticeably distorts the low frequencies from the kick and snare elements, potentially smearing the transients of the hits, whereas the high-pass condition maintains the low frequency energy and transient attack – this may have also influenced the increase of perceived LEV (Figure 7.12).

The only other two sources with any noticeable change of TQ are the Acoustic Guitar and Male Speech samples, although these appear to be related to the level of ambience rather than the processing of decorrelation. With the Acoustic Guitar, the 'Lower-3' condition has been graded with less TQ compared to all other conditions (which were graded similarly), comments of which were 'Thin' and 'Muddy' (when looking at individual responses). These results suggest that an increase of ambience might be preferable for this particular sample, potentially from a boost of lower frequency energy. For the Male Speech source, the opposite occurs, where the lower ambience of the 'Lower-3' condition actually improves perceived TQ. The most common

written response for this sample was ‘Clear’, suggesting that a decrease of ambience improves the intelligibility of the direct speech signal, as might be expected.

7.5 Objective Analysis of the Stimuli Signals

Both the LEV and TQ results appear to be somewhat related to the frequency-dependent level within the source signals. In order to investigate this further, the stimuli signals have been binauralised and objectively assessed over the following sections. As with Chapter 5, head-related impulse responses (HRIRs) from the MIT KEMAR database (Gardner & Martin, 1994) have been convolved with the stimuli signals for the nine loudspeaker positions. These were chosen to maintain consistency with the objective analysis in Chapter 5 of the present thesis. The nine HRIR-convolved signals of each ear were then summed together independently to create binaural signals of the stimuli. Assuming relative symmetry between the two ears, the observations discussed below are for the left ear signal only.

The left ear RMS levels (dB) for each stimulus condition can be seen in Table 7.12, normalised to 0 dB at the 'Lower' reference for ease of comparison. Here it is seen that 'Lower-3' generally has the least amount of energy for each source, as would be expected due to the reduced ambience level. However, for the String Quartet source, the RMS levels for the 'Upper', 'KP', 'ZP' and 'ZP500' conditions are also relatively low compared to the 'Lower' reference and 'ZA/ZA500' decorrelated stimuli – it seems that this slight reduction in level might relate to the grading of less LEV for these samples in the subjective testing (Figure 7.12). Furthermore, the 'ZA' and 'ZA500' conditions have the greatest RMS level for each source, and it was these samples that were generally graded as having the most LEV throughout testing.

Table 7.12 Broadband RMS levels (dB) of the HRIR-convolved stimuli (Left Ear), normalised to 0 dB at the 'Lower' reference for comparison between conditions and sources.

	Lower	Lower-3	Upper	KP	KP500	ZP	ZP500	ZA	ZA500
Cello	0.0	-0.7	-0.4	-0.1	0.0	-0.4	-0.4	0.3	0.2
Drumkit	0.0	-0.6	-0.1	0.1	-0.2	0.2	0.1	0.2	0.3
Guitar	0.0	-0.8	-0.3	0.0	-0.3	-0.2	-0.3	0.1	0.1
Quartet	0.0	-0.7	-0.8	-0.8	-0.5	-0.7	-0.7	0.2	0.3
Speech	0.0	-0.7	-0.2	0.1	-0.1	-0.3	-0.2	0.3	0.1
Trumpet	0.0	-0.7	0.0	0.1	-0.2	-0.2	-0.2	0.2	0.2
Debussy	0.0	-0.5	0.5	-0.3	0.1	0.0	0.2	0.7	0.5

Given the subjective LEV results for the 'Lower-3' condition (where it was perceived as having significantly less LEV than the reference for all sources), it has been suggested that the perception of LEV may largely be influenced by the ratio of ambience energy to direct sound energy (A/D). This can be related to concert hall acoustics, where Lee (2013) demonstrates that the strength factor (G) of late reflections has a strong correlation with the perception of LEV at different source-listener distances. Although the differences of RMS observed in Table 7.12 are relatively small (~1 dB), the present results also suggest that ambient sound strength (or sound strength in general) may contribute most to LEV perception, rather than any perceptual effect from vertical interchannel decorrelation. Moreover, the cause of these changes in level is not clear, and there is no indication of the frequencies at which the changes occur. To investigate this further, the following section looks at energy differences within octave-bands, in order to provide insights into how different frequencies affect the perception of LEV (and TQ).

7.5.1 Octave-Band RMS Levels

Octave-band RMS levels have been calculated from the left ear signals of the HRIR-convolved stimuli. The relative differences of octave-band RMS for the stimuli test signals from the 'Lower' reference levels are presented in Figure 7.14 (i.e. all loudspeaker signals combined at the ear). The 'Lower' reference octave-band RMS levels are displayed in Table 7.13 below, normalised to 0dB for the octave-band with the greatest energy – the relative differences of octave-band RMS in Figure 7.14 are derived from these values.

Table 7.13 Octave-band RMS levels (dB) of the 'Lower' reference condition for each source, taken from the Left Ear of the HRIR-convolved stimuli, normalised to 0dB at the octave-band with the greatest energy.

	63 Hz	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	8 kHz	16 kHz
Cello	-26.4	-14.4	-3.7	-2.9	-4.1	0	-5	-13.9	-19.9
Drumkit	-8.5	-6.6	-1.4	-2.1	-5.2	0	-2	-5.2	-9.2
Guitar	-18.5	-2.1	-1.1	-0.4	-1.4	0	-4.7	-10.5	-17.4
Quartet	-23.6	-6.3	-3.5	-3.1	-6	0	-3.7	-15.6	-21
Speech	-8.3	-4.3	-2.3	-2.4	-4.1	0	-2.5	-8.5	-12.5
Trumpet	-40.8	-34.6	-17.7	-6.3	-4.2	0	-8.4	-24.1	-30.9
Debussy	-26.5	-14	-0.9	-2	-2.6	0	-5.8	-16.2	-20.8

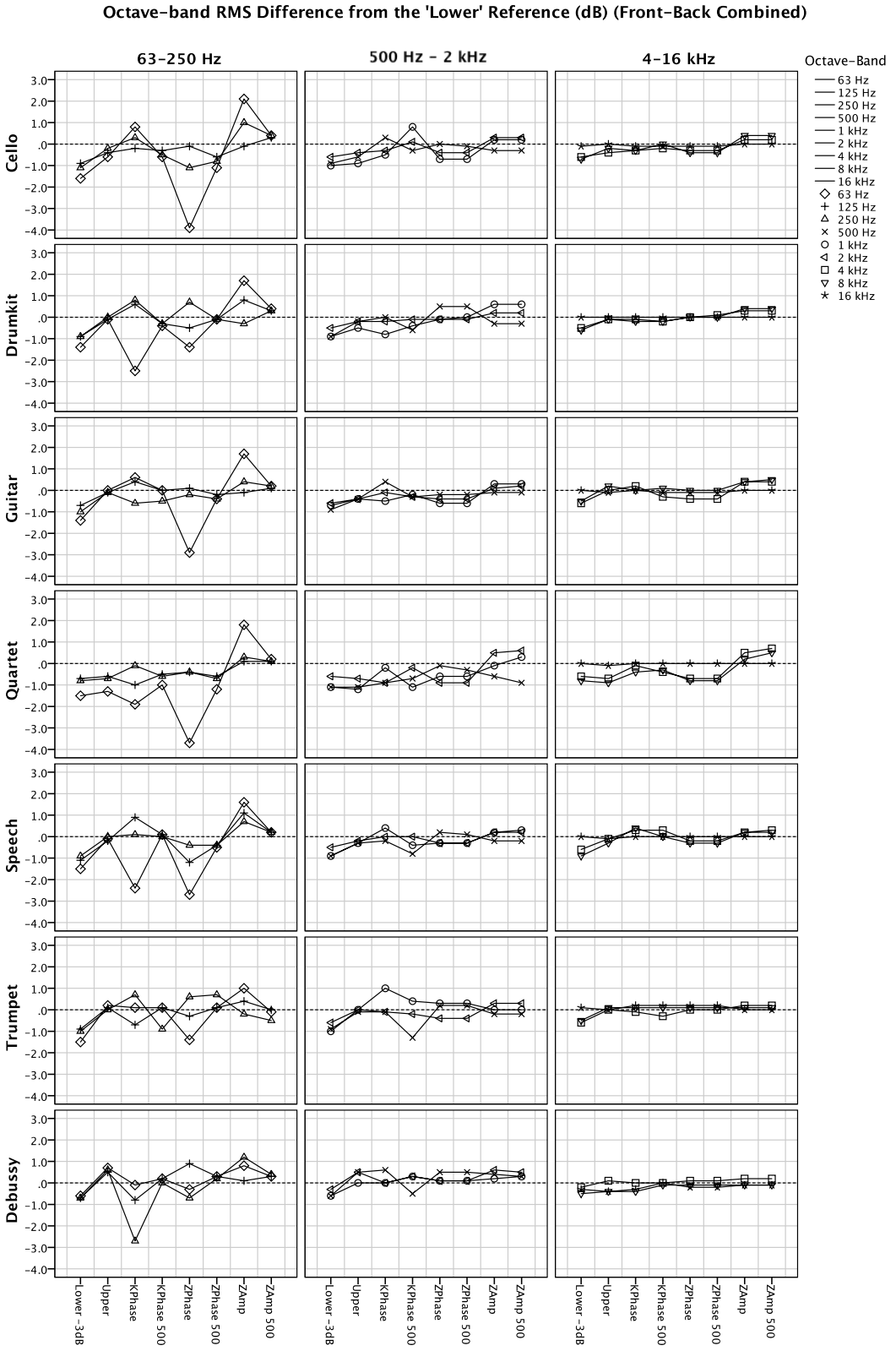

Figure 7.14 Difference of octave-band RMS (dB) from the 'Lower' reference for each test condition – calculated from the left ear of the HRIR-convolved stimuli. The 'Lower' RMS is represented by the dotted line at 0 on the y-axis, where above the dotted line is more octave-band energy for a particular condition and below the dotted line is less energy.

The plots in Figure 7.14 display the difference of octave-band RMS from the 'Lower' reference – that is, the octave-band RMS levels for the 'Lower' reference (Table 7.13) subtracted from the octave-band RMS levels of each source condition. The x-axis is split by condition, with the markers indicating the different octave-bands for each particular stimulus – the three columns divide the octave-bands into tri-octave groups for clarity. Above the reference line at 0 dB indicates an increase of RMS level for that octave-band in comparison to the 'Lower' reference, and below shows a decrease of RMS level.

In general, little RMS deviation from the 'Lower' reference is seen for the higher frequency bands in Figure 7.14. However, looking at lower frequency bands, the low frequency distortion that was seen in previous experiments is clearly present for the broadband decorrelated signals ('KP', 'ZP' and 'ZA'), particularly within the 63 Hz octave-band (±1-3 dB). This variation is shown to decrease for the high-pass decorrelation conditions, where the low frequency energy (63 – 250 Hz octave-bands) is relatively similar to that of the 'Lower' condition, as might be expected. Related to this, the subjective TQ results for the Drumkit in Figure 7.13 show a slight trend of TQ improving with high-pass decorrelation, in comparison to the broadband decorrelated conditions for each method. It is also seen in the ambience RMS octave-band levels of each source signal in Table 7.9 (Section 7.3.1) that the Drumkit sample has the greatest amount of energy in the 63 Hz octave-band. Considering these points, it is likely that the observed low frequency decorrelation distortion would have been more noticeable in the ambience for the Drumkit than other samples, with high-pass decorrelation reducing the negative effect.

With regard to the other subjective results, there appears to be some further association between octave-band RMS levels at the ear and the perception of LEV. Of the decorrelated Cello conditions, 'ZA' was graded as the sample with greatest LEV, while 'ZP' was the least enveloping. Observing the plots in Figure 7.14, the 'ZA' condition has considerably more energy around the 63 Hz band than 'ZP' (~6 dB). Similarly, the Acoustic Guitar 'ZA' condition had greater LEV than the other conditions – Figure 7.14 shows that this sample also has an increase of

energy in the 63 Hz octave-band, when compared against the other stimuli (~2-5 dB). With the String Quartet source, the 'ZP' condition was graded as having significantly less LEV than the 'Lower' reference. It appears that this result may also relate to the 63Hz band, which sees a noticeable reduction of energy compared to the other conditions (~2-6 dB).

7.5.2 Front-Back Energy Ratio (F/B)

The RMS octave-band levels in Section 7.5.1 above appear to show some relationship between the energy of stimuli signals at the ears and the perception of LEV. In concert halls, it is thought that LEV is typically dictated by the energy and direction of late reflections (Furuya et al., 2007). Studies have demonstrated that comparing the energy of reflections from in front of the listener, with the energy of those from behind, can indicate the degree of LEV – this energy ratio has been shown to be applicable with both early reflections (where direct sound is present) and late reflections (Morimoto & Iida, 1993). The described objective measure is named the Front-Back energy ratio and can be calculated using Equation 7.1 below.

$$F/B\ Energy\ Ratio = 10\log_{10}\left(\frac{\int_0^\infty f^2(t)dt}{\int_0^\infty b^2(t)dt}\right) \qquad (7.1)$$

Since a relationship between F/B energy and LEV has previously been established in concert halls, it is of interest to calculate the F/B energy ratio for the stimuli of the current experiment. Given that the F/B ratio of both early reflections (with direct sound) and late reflections seem to be associated with LEV, the direct signal and ambience parts have both been preserved when assessing the F/B energy ratio below. Further binauralised stimuli were created for the front loudspeaker array and rear loudspeaker array independently, using the MIT KEMAR HRIR database (Gardner & Martin, 1994). For the 'Front' part, the signals from the frontal loudspeakers include both direct and ambient sound from the Centre, Left, Right, Left Height and Right Height channels; whereas the 'Back' signals feature ambience only from the Left Surround, Right Surround, Left Surround Height and Right Surround Height. In effect, the ratio of energy is calculated between the frontal loudspeaker signals (C, L, R, LH and RH) and the rear

loudspeaker signals (Ls, Rs, LsH and RsH) at the ear position. Calculation of the energy ratio was performed on binauralised signals, so that the spectral HRTF filtering can be included in the calculation i.e. representing the F/B ratio that is actually experienced by the listener.

Given that the direct sound from the front was at a consistent level for all conditions, it is thought that presenting the Back-Front energy ratio (B/F) would more relevant to this study. B/F can be calculated using Equation 7.2 below, where the 'Back' energy is the overall energy from the HRIR-convolved rear loudspeaker array and 'Front' is the overall energy from the HRIR-convolved frontal array (as calculated at the ear from the binaural signals). The B/F values for each HRIR-convolved source condition are displayed in Table 7.14 below, where the dB values indicate the amount of energy from the back array in relation to the front.

$$B/F\ Energy\ Ratio = 10\log_{10}\left(\frac{\int_0^\infty b^2(t)dt}{\int_0^\infty f^2(t)dt}\right) \quad (7.2)$$

Table 7.14 Back-Front energy ratio (B/F) (dB) between the HRIR-convolved stimuli for the back loudspeaker signals only and front loudspeaker signals only (Left Ear).

	Lower	Lower -3 dB	Upper	KP	KP500	ZP	ZP500	ZA	ZA500
Cello	-5.1	-7.1	-5.6	-5.5	-4.9	-5.8	-5.9	-4.3	-4.4
Drumkit	-5.8	-7.9	-6.0	-5.7	-6.2	-5.3	-5.6	-5.5	-5.1
Guitar	-5.2	-6.9	-5.4	-4.6	-5.5	-6.0	-5.9	-4.6	-4.7
Quartet	-5.8	-7.8	-7.4	-7.7	-6.4	-6.7	-7.2	-5.3	-5.1
Speech	-5.4	-7.3	-5.5	-4.7	-5.6	-6.0	-5.8	-4.6	-5.0
Trumpet	-6.1	-8.2	-5.2	-5.6	-5.7	-6.5	-6.5	-5.2	-5.2
Debussy	-1.6	-4.6	-0.4	-2.0	-1.0	-1.0	-0.5	1.0	0.2

In these results, it is seen that the B/F energy ratio does not vary greatly between conditions. The 'ZA' and 'ZA500' conditions generally show the greatest energy from behind for each source, which corresponds with a slight increase of LEV for the same conditions in the subjective testing (Figure 7.12). This is likely due to a decrease of phase cancellation at the ear in comparison to the phase-based methods (ZP and KP), however, it is not known whether such a small change of B/F would have a noticeable effect. Further research is required to investigate whether B/F energy contributes to the perception of LEV within multichannel surround sound systems, in addition to determining the just noticeable difference of B/F energy ratio.

7.6 Practical Implications

For the conditions of the current experiment, vertical decorrelation of ambience appears to have little benefit in terms of increasing LEV. The subjective results and objective analysis seem to suggest that the perception of LEV in this experiment was mostly related to the ear-input level of the stimuli, both overall and potentially between the front and back loudspeaker arrays (i.e. the Front-Back energy ratio). Furthermore, the 2D condition with -3 dB less ambience was consistently graded as having significantly less LEV than all other 2D and 3D samples. This also supports the notion that the level of ambience plays a more important role in LEV perception than the vertical decorrelation of ambient signals.

Of the three decorrelation approaches assessed in the current experiment, the amplitude-based 'ZA' appears to be slightly more effective than the phase-based methods ('KP' and 'ZP'). It is hypothesised that cues from phase-based decorrelation may break down with the inclusion of multiple similar signals. On the other hand, vertical interchannel amplitude differences effectively result in interlayer amplitude differences, which may have been easier to perceive. For decorrelation to have any significant effect, it is thought that the source signal must have sufficient relative high frequency energy, as previously demonstrated in Chapter 6 of the present thesis. The 'ZA' method results in the current experiment support this with comparative LEV results between the broadband and high-pass decorrelation conditions. Moreover, broadband and high-pass decorrelating the Drumkit sample with the 'ZA' method were the only conditions to show a significant increase of LEV, which is likely due to the Drumkit being the source with the greatest amount of energy in the 8 kHz and 16 kHz octave-bands.

Additional study is required to investigate the above points, in order to fully determine the effectiveness of vertical decorrelation for 2D-to-3D upmixing applications. In particular, it would be interesting to experiment with the number and position of loudspeakers, as well as assess a broader range of source material and decorrelation techniques.

7.7 Conclusion

Two listening tests have been conducted to assess the perceptual effect of vertical interchannel decorrelation in a 2D-to-3D upmixing scenario. The first test looked at the perception of Listener Envelopment (LEV) by decorrelation, and the second at the impact of decorrelation on Tonal Quality (TQ). Three decorrelation techniques were assessed: 1) Kendall's all-pass filter approach, 2) Zotter and Frank's phase-based approach, and 3) Zotter and Frank's amplitude-based approach. These decorrelation methods were applied to ambient signals only between the main-layer and height-layer loudspeakers of the Auro-3D 9.1 format. Both broadband decorrelation and high-pass decorrelation of the 500 Hz octave-band and above were assessed for each of the methods. Level-matched ambience in the main-layer only, -3 dB ambience in the main-layer only and level-matched ambience in the height-layer only were also compared against the decorrelated stimuli during testing. Furthermore, direct sound was present for all conditions in the front left, right and centre loudspeakers, to provide a realistic situation. For both attributes under testing, LEV and TQ, a multiple-comparison test was performed on a bipolar scale, with the level-matched ambience in the main-layer only as a reference for the centre of the scale i.e. the 2D condition. A variety of complex sources were assessed, consisting of a Cello, Drumkit, Acoustic Guitar, String Quartet, Male Speech, Trumpet and an ensemble recording.

The key findings from the listening tests are as follows:

- Upmixing of ambience by decorrelation to 3D had little significant effect on LEV, in comparison to the level-matched ambience from the 2D main-layer only reference.
- The relative level of ambience to direct sound appeared to be most influential on the perception of LEV– this was shown by the significant decrease of LEV for the condition where the ambience was attenuated by -3 dB compared to the other conditions.
- There was no perceptual difference between broadband and high-pass decorrelation, suggesting that the decorrelation of low frequencies is unnecessary for enhancing spatial perception, as was suggested by the results in Chapters 4 and 6.

- The amplitude-based decorrelation method appeared to be marginally more effective than the phase-based methods at increasing LEV. This was particularly the case for sources with greater energy in the 4 kHz to 16 kHz octave-bands.
- In contrast, both phase-based methods demonstrated little change to LEV across the majority of sources – it has been suggested that phase changes from decorrelation are difficult or impossible to detect when multiple similar signals are present.
- Some changes to LEV appear to be influenced by changes of energy in low frequency bands (63 Hz to 250 Hz octave-bands), potentially from a phase cancellation effect.
- The decorrelation of ambience had very little impact on TQ in the listening position, possibly due to the direct sound masking any spectral distortions.
- The only negative effect on TQ by decorrelation was with the broadband phase-based methods on the Drums sample, which may be related to low frequency decorrelation distortion or the smearing of the kick and snare drums transients. As a result, a slight improvement of TQ was seen for the high-pass decorrelation conditions.
- There was some source-dependency of ambience level on TQ – for the Male Speech, less ambience resulted in better TQ due to clarity and intelligibility, whereas for the Acoustic Guitar, less ambience seemed to decrease the TQ.
- For sources with sufficient high frequency content, high-pass decorrelation may be a useful approach for achieving similar levels of LEV with little-to-no impact on TQ.

8 SUMMARY AND CONCLUSIONS

This chapter summarises the findings of the experiments and subsequent analysis that are presented in the current thesis. The first section details the experiments that have been undertaken, along with the key findings during each phase. Following this, overall conclusions are drawn, featuring discussions that regard the practical implications of the findings. Lastly, further work is considered based on the experimental results described in this thesis.

8.1 Summary of Chapters

8.1.1 Chapter 0 (Introduction)

The first chapter of the present thesis (Chapter 0) introduces the background of the study and the initial research questions that were proposed. The basic understanding of interchannel decorrelation in the horizontal plane is described – that is, as signal correlation between a pair of left and right loudspeakers decreases, the horizontal extent of the phantom auditory image increases. However, it is determined that little knowledge is known about the perception of interchannel decorrelation between vertically-arranged loudspeakers. This leads to the main research question being: "What is the perceived effect of vertical interchannel decorrelation?"

A focus of the study is placed on 2D-to-3D upmixing as a potential application of vertical interchannel decorrelation. In this regard, it is stated that upmixing from 5.1 Surround to Auro-3D 9.1 (Auro Technologies, 2015a) shall provide the basis for the proceeding experiments. As a result, the effects of vertical interchannel decorrelation have been observed at three azimuth angles throughout the current thesis: ±30° and ±110°, along with 0° to assess the median plane. Furthermore, all presentations of vertical interchannel decorrelation in the thesis are between a main-layer loudspeaker at ear height, and a height-layer loudspeaker elevated directly above by +30°, as specified for the Auro-3D 9.1 format.

8.1.2 Chapter 1 (The Spatial Perception of Audio)

Chapter 1 reviews the auditory mechanisms for localising and perceiving sound within space. The chapter begins with the fundamental cues of horizontal (lateral) localisation (Section 1.1.1) – these are a combination of interaural level difference (ILD) at high frequencies (> 1.5 kHz) and interaural time difference (ITD) at low frequencies (< 1.5 kHz), known as the 'duplex theory'. The effect of ILD in 3D surround sound reproduction is also considered, with regard to a reduction of the head-shadowing effect (less ILD) when sources are presented from height. Following this, the auditory mechanisms for localisation in the vertical domain are explored (Section 1.1.2). These primarily come in the form of spectral cues at higher frequencies (particularly around 8 kHz from the front); however, it is thought torso and shoulder reflections can also generate elevation notch cues as low as 700 Hz. Phenomena that see the inherent distribution of frequencies in the vertical domain known as the 'pitch-height' effect and the 'directional bands' effect are also discussed, suggesting an association with high frequency pinna filtering.

The second section of Chapter 1 (Section 1.2) examines literature regarding the inherent spread of sound sources, both horizontally and vertically, where it is seen that lower frequencies are generally perceived as broader than higher frequencies. Furthermore, it is shown that loudness and signal duration can also have an impact on spread perception, in addition to frequency. The following section (Section 1.3) looks at spatial impression (the perception of sound within an enclosed space – e.g. a concert hall), of which the two main components are apparent source width (ASW) and listener envelopment (LEV). It is shown that both attributes are strongly dictated by lateral reflections, where early reflections (< 80 ms) contribute to ASW and late reflections (> 80 ms) contribute to LEV, with the impact of reflections from above (the ceiling) also considered in this section. A discussion on the role of the interaural cross-correlation coefficient (IACC) within ASW perception follows, which relates directly to the perception of horizontal decorrelation. The final section of the chapter (Section 1.4) looks at objective measures for quantifying and predicting spatial attributes (i.e. ASW and LEV).

8.1.3 Chapter 2 (The Spatial Control of Audio in Surround Sound)

In Chapter 2, the perception of sound by loudspeaker reproduction is considered, particularly with regard to commercial multichannel surround sound systems. The first section (Section 2.1) describes the loudspeaker formats that are widely in use today – from two-channel stereophony (left and right) up to more recently developed 3D surround sound systems (such as, Dolby Atmos, Auro-3D and DTS:X). The following section (Section 2.2) considers the control of a point source within loudspeaker reproduction, mostly by way of amplitude panning. It is found that the perception of panning between a pair of vertically spaced loudspeakers is largely inaccurate. However, the studies demonstrate that a phantom auditory image can still be perceived in the vertical domain. There is also some suggestion that high frequencies contribute to the perception of vertical panning, which corresponds with the cues for vertical localisation.

Section 2.3 discusses interchannel decorrelation and its perceived effects. It is found that decorrelation techniques can broadly be split into phase-based and amplitude-based approaches. Phase-based methods alter the phase component of frequencies, typically by implementing all-pass filters or random time-delays of critical bands. On the other hand, amplitude-based approaches tend to manipulate the spectral-amplitude of the input. This can be achieved both randomly or with fixed filters, and is often complementary between a pair of signals i.e. opposing differences. The potential effects of vertical interchannel decorrelation are also considered, however, little literature exists on the subject at present. The final section of the chapter (Section 2.4) describes proposed two-to-five channel upmixing methods, most of which feature ambience extraction and decorrelation to generate signals for reproduction in surround channels. Lastly, recent proposals of potential 2D-to-3D upmixing algorithms are discussed.

8.1.4 Chapter 3 (Horizontal and Vertical Decorrelation)

Chapter 3 describes a two-part experiment comparing horizontal and vertical interchannel decorrelation using the same stimuli. Three frequency bands were assessed: ‘Low’ (octave-bands with centre frequencies of 63 Hz, 125 Hz and 250 Hz), ‘Middle’ (centre frequencies of

500 Hz, 1 kHz and 2 kHz) and 'High' (centre frequencies of 4 kHz, 8 kHz and 16 kHz). These frequency bands were decorrelated using the complementary comb-filtering method with time-delays of 1 ms, 5 ms, 10 ms and 20 ms and gain factors between 0.0 and 1.0 (at increments of 0.2, which varies the ICC). In the horizontal domain, stimuli were presented between a left and right loudspeaker pair with a base angle of 60°; and in the vertical domain, stimuli were presented through a vertically spaced loudspeaker pair in the median plane, with one position at 0° elevation (ear height) and the other at +30° to the listening position.

The key findings from the experiment are as follows:

- Interchannel decorrelation contributes to both horizontal and vertical image spread.
- The vertical decorrelation effect is noticeably weaker than horizontal decorrelation.
- Horizontal decorrelation was effective for all frequency bands, presumably due to the relationship with the interaural cross-correlation (IAC).
- Vertical decorrelation appears to be slightly more effective at lower frequencies ('Low' and 'Middle'), though this may be due to a strong floor reflection in the listening room.
- Vertical decorrelation of the 'High' band is associated with spectral notches.
- The lack of VIS change for the 'High' frequency band could be related to an inherent vertical spread of frequencies from the 'pitch-height' effect.
- A time-delay of 1 ms with the complementary comb-filter method is unsuitable for low frequency decorrelation, due to an uneven distribution of frequencies.

8.1.5 Chapter 4 (Octave-Band Decorrelation: Subjective Testing)

Two experiments are described in Chapter 4, concerning the vertical interchannel decorrelation of octave-band stimuli from different azimuth angles. The first experiment looks at the relative perception of vertical image spread for octave-bands with centre frequencies from 63 Hz to 16 kHz and broadband pink noise. Eight conditions were tested for each of the nine frequency bands: a monophonic condition (signal from the lower main-layer loudspeaker only), a

correlated condition (coherent signals from both loudspeakers where the interchannel cross-correlation coefficient (ICCC) is 1.0) and six decorrelated conditions. The decorrelated conditions consisted of two decorrelation methods (phase randomisation by all-pass filters (PR) and complementary comb-filtering (CF)), each with three degrees of correlation (ICCCs of 0.1, 0.4 and 0.7). These eight stimuli were presented as a multiple comparison trial from three azimuth angles of vertically-arranged loudspeaker pairs (based on Auro-3D 9.1 with an additional centre height): 0°, ±30° and ±110° azimuth, where the height channels were elevated by +30°.

The key findings from this first experiment are as follows:

- A linear association between ICC and VIS appears to occur around the 500 Hz octave-band and above. That is, as correlation decreases, the perceived VIS increases.
- The strongest relationship was seen for the 8 kHz octave-band at ±30°.
- The strength of the trend appears to be direction-dependent for each band:
    - 500 Hz and 1 kHz produced the strongest results in the median plane at 0° azimuth (where energy is equal in both ears).
    - 2 kHz, 4 kHz and 16 kHz produced the strongest results at ±110° (where head-shadowing is greatest, increasing ILD from the main-layer channel).
    - 8 kHz and Broadband pink noise produced the strongest results at ±30° (where interaural differences and frontal spectral filtering from the pinna are present).
- PR and CF performed similarly, with PR producing slightly greater VIS in some cases.
- The monophonic condition was perceived as having a similar VIS to other conditions, which may have been related to the 'pitch-height' effect and/or room reflections.

The second experiment assessed the absolute VIS of extreme stimuli conditions from the first experiment. These consisted of the monophonic condition, the coherent condition and one decorrelated condition (PR with an ICCC of 0.1) for each frequency band. Azimuth angles of 0° and ±30° were tested, and a light emitting diode (LED) strip by the 0° azimuth position was

used to help capture the responses of absolute VIS. Controlling the LEDs, subjects were asked to visually define the upper and lower boundary of the VIS for each stimulus independently.

The key findings from this second experiment are as follows:

- Lower frequencies tended to have a greater inherent VIS than high frequencies.
- Significant changes to boundaries were seen for octave-bands above 500 Hz.
- Large deviation around the boundary positions indicates a difficulty defining them.
- Evidence of the 'pitch-height' effect is observed with the 4 kHz and 8 kHz bands.
- 'Directional bands' may have influenced the 1 kHz, 4 kHz and 8 kHz band results.
- The 8 kHz band was strongly biased towards the height-channel loudspeaker.

8.1.6 Chapter 5 (Octave-Band Decorrelation: Objective Analysis)

Chapter 5 presents objective analysis of the stimuli signals from the subjective experiments in Chapter 4, as well as analysis of the binaural room impulse responses (BRIRs) captured in the listening room. Stimuli signals were binauralised with two sets of impulse responses – the first were anechoic head-related impulse responses (HRIRs) from the MIT KEMAR database, and the second were the BRIRs from the listening room. In the first analysis section, the spectra of the HRIR-convolved stimuli were inspected. The next section looks at the interaural cross-correlation (IAC) of the both the HRIR-convolved and BRIR-convolved stimuli. Lastly, the early reflection to direct sound ratio is calculated from the BRIRs of the listening room.

The key findings from the objective analysis are as follows:

- Spectral changes associated with decorrelation occur in the 8-16 kHz octave-bands at 0° azimuth, 4-16 kHz bands at ±30° azimuth and 2-16 kHz bands at ±110° azimuth.
- There is a strong association between vertical decorrelation and spectral cues in the 8 kHz band at all angles, where decorrelation 'fills in' vertical localisation notch cues.

- For the 16 kHz band, the spectra of the decorrelated and monophonic conditions are mostly similar. However, the correlated condition results in phase cancellation at the ipsilateral ear – this cancellation is then reduced as the signal correlation decreases.
- Perception of VIS for the 2 kHz and 4 kHz bands appear to be mostly affected by a head-shadowing effect and changes to the spectrum in the contralateral ear.
- At wide azimuths (±110°) vertical decorrelation relates to IACC, due to the head-shadowing of the main-channel signal and decorrelation of the height-channel signal.
- VIS for 500 Hz and 1 kHz may have been affected by decorrelated early room reflections (IACC), hence the strong median plane result (where reflective energy is equal).
- 'The Main-Layer Effect' is observed, where a monophonic main-layer signal has greater early reflective energy than a vertical pair of loudspeaker signals (when both are level-matched), which in turn decreases IACC. This is particularly the case for the 500 Hz and 1 kHz bands, where the monophonic conditions had greater perceived VIS.

8.1.7 Chapter 6 (High-Pass Decorrelation of Complex Stimuli)

In Chapter 6, a two-part experiment is described that observes the effect of decorrelating high frequencies only in ambient complex signals. The first part assessed the relative vertical image spread (VIS) of stimuli, and the second part looked at tonal quality (TQ). Six sources were presented for each part: Male Speech, Cello, Drumkit, Acoustic Guitar, a String Quartet and Broadband Pink Noise. The cut-off frequencies for the high-pass decorrelation were defined by the lower limit of octave-bands between 63 Hz and 8 kHz, resulting in high-pass cut-offs of 44 Hz (Broadband), 88 Hz, 177 Hz, 355 Hz, 710 Hz, 1420 Hz, 2840 Hz and 5680 Hz. Each of the eight cut-off conditions and a monophonic reference (lower main-layer loudspeaker only) were compared in multiple comparison trials (9 stimuli). All six sources were presented from three azimuth angles of vertically-arranged loudspeaker pairs (0°, ±30° and ±110°, where the height-channel is elevated by +30°) – this resulted in 18 trials for both the VIS and TQ parts of testing.

The key findings of the experiment are as follows:

- High-pass vertical decorrelation with cut-off frequencies of 355 Hz and below (i.e. decorrelating the 500 Hz band and above) had a similar VIS to broadband vertical decorrelation for all sources and azimuth angles (0°, ±30° and ±110°).
- Vertical decorrelation causes a summing distortion, notably at low frequencies.
- For sources with a relatively static response at low frequencies (e.g. Pink Noise, Speech and Drumkit), TQ begins to decrease as the cut-off frequency is decreased.
- For dynamic musical sources (e.g. Cello, Acoustic Guitar and the String Quartet), decorrelation of lower frequencies can cause a slight *increase* of TQ in some cases.
- Decorrelating only the 8 kHz band and above can result in a significant increase of VIS over the monophonic reference, if the source signal has sufficient high frequency energy (e.g. the Drumkit and Pink Noise).
- For all the sources except Pink Noise from 0°, decorrelating the 500 Hz octave-band and above had no significant difference in TQ from the monophonic reference.
- It is concluded that high-pass decorrelation of the 500 Hz band and above (355 Hz cut-off frequency) can significantly increase VIS, with little impact on TQ for the vast majority of conditions (by avoiding the summing distortion at low frequencies).

8.1.8 Chapter 7 (2D-to-3D Upmixing)

Chapter 7 describes a two-part experiment where 2D-to-3D upmixed stimuli are assessed for both listener envelopment (LEV) and tonal quality (TQ). All 3D upmixed stimuli were created using vertical interchannel decorrelation of ambient signals between a main-layer and a height-layer of loudspeakers – this was based on upmixing from 5.1 Surround to Auro-3D 9.1. Three decorrelation methods were used for upmixing the ambience: two phase-based (Kendall, 1995; Zotter & Frank, 2013) and one amplitude-based (Zotter & Frank, 2013). For each decorrelation method, two decorrelation conditions were implemented: broadband decorrelation and high-pass decorrelation of the 500 Hz octave-band and above only (based on the results from Chapter

6). Three control conditions were also included (ambience in the main-layer only (the reference), ambience in the main-layer only attenuated by -3 dB and ambience in the height-layer only). This resulted in a total of nine conditions being compared in multiple comparison trials for both TQ and LEV. With all nine conditions, direct sound was also included in the front left, right and centre channels, providing a practical context for the experiment. A total of seven sources were assessed: Cello, Drumkit, Acoustic Guitar, String Quartet, Male Speech, Trumpet and an ensemble recording.

The key findings from the experiment are as follows:

- Upmixing by vertical interchannel decorrelation had little significant effect on LEV.
- The strength of ambience had a significant effect on LEV, where the -3 dB attenuated condition had significantly less LEV than the main-layer reference for all sources.
- There was no significant difference between broadband and high-pass decorrelation.
- The amplitude-based method had slightly greater LEV than the phased-based ones.
- The greatest increase of LEV was seen for sources with more high frequency energy. For example, the Drumkit sample had the greatest energy in the 8 and 16 kHz bands and was the only source to show a significant increase of LEV.
- For sources with less high frequency energy, LEV appears to be dictated by changes in low frequency energy (from the summing distortion seen in Chapter 6).
- Changes in the ratio of front-back energy (F/B) also seem to correspond somewhat with perceived LEV, however, the variation between conditions is slight.
- Upmixing by vertical decorrelation had very little impact on perceived TQ.
- Only the Drumkit sample showed a slight decrease of TQ with broadband decorrelation using the phase-based methods – this was improved with high-pass decorrelation.
- Ambience level had some impact on TQ – for example, the -3 dB condition for the Male Speech improved TQ, due to an increase of clarity and intelligibility.

8.2 Conclusions

In Section 8.1 above, a summary of the experimental findings from the present thesis are broken down by chapter. This section aims to consolidate some of the key findings and briefly discuss the practical implications that they bring. Firstly, it was seen in Chapters 4 and 5 that the perception of vertical image spread (VIS) was strongly associated with the 8 kHz octave-band (particularly from the front). This relates directly to vertical localisation notches that occur within this region (Roffler & Butler, 1968a; Hebrank & Wright, 1974). The spectral analysis in Chapter 5 demonstrated how vertical decorrelation 'filled in' these directivity notches, most notably around 10-11 kHz. In the case of binaural rendering, it may be found that artificially 'filling in' localisation notches of HRTFs can invoke a virtual sense of VIS – alternatively, vertically decorrelating between virtual loudspeakers is likely to have a similar effect. Chapter 5 also suggests that vertical interchannel decorrelation at wide azimuth angles (±110°) could be related to changes in ILD and IACC, due the head-shadowing of the main-channel signal. It may be found that these cues can also influence the perception of VIS around the head in binaural rendering, as well as provide a general improvement of localisation accuracy to the side.

As is clearly demonstrated in the subjective results of Chapter 6, similarly significant increases of VIS can be generated by vertically decorrelating high frequencies only (when compared to vertical decorrelation of a broadband signal) i.e. 'high-pass decorrelation'. This is of particular use for avoiding the low frequency summing distortion that has been observed in Chapters 5, 6 and 7. It was determined that a potential cut-off frequency for high-pass decorrelation might be 355 Hz (the lower limit of the 500 Hz octave-band), which produced a significant increase of VIS with little impact on the tonal quality for all complex sources. Having said that, the results in Chapter 6 also demonstrate that when the source signal has sufficient high frequency energy, a significant increase of VIS can be achieved by vertically decorrelating just the 8 kHz and 16 kHz bands in a broadband signal. This supports the case that the 8 kHz octave-band is particularly important for VIS perception. In object-based surround sound, where direct sound objects

are likely to contain more high frequency energy than ambient signals, it may be found that simply decorrelating the 8 kHz octave-band between spaced points is able to provide greater control over the object's extent.

An interesting effect that has been called the 'main-layer effect' is discussed in Chapter 5. It was observed that a monophonic signal from the main-layer loudspeaker had greater early reflection energy than a vertical stereophonic pair of signals, when both conditions were level-matched. This is due to reflections from the two vertically-arranged loudspeaker signals arriving at the ears at different times, causing a distribution of reflection energy, whereas the two direct signals are summed together in alignment. When the monophonic signal is amplified to match the summed direct signals, the reflection energy is amplified along with it. It appears that this increase in early reflection energy may have caused an increase of VIS for monophonic conditions in the subjective testing (particularly for the 500 Hz and 1 kHz octave-bands). An implication of this might be that, if the presentation room is important to spatial perception, use of height-channel loudspeakers may be unnecessary for enhancing spatial impression. That is, a greater sense of LEV might be generated by utilising early reflections within the listening room. This hypothesis is somewhat shown in the upmixing experiments of Chapter 7, where there was very little change in LEV between the 2D main-layer reference and 3D upmixed stimuli, when both conditions are level-matched.

One of the research aims proposed in the Introduction (Chapter 0) was to determine whether interchannel decorrelation is suitable for 2D-to-3D upmixing. The results in Chapter 7 suggest that there may be some slight increase of LEV when using an amplitude-based decorrelation method. However, it seems to rely on the source signal having sufficient high frequency energy, which is often lacking in ambient signals. To counter this, the high frequency energy of the source could be boosted, though this would not preserve the energy balance of frequencies within the original signal and may sound unnatural. Furthermore, vertical decorrelation could potentially be used in conjunction with other upmixing methods, for example, the perceptual

band allocation (PBA) method proposed by Lee (2016b). PBA routes different octave-bands to either main-layer or height-layer loudspeakers, based on their inherent vertical location due to the 'pitch-height' phenomena. It is shown to be an effective method for 2D-to-3D upmixing, however, it may also lead to specific frequency bands being perceived as points rather than a vertical spread of sound. Decorrelation could potentially be used in this instance, by generating a greater spread for frequencies that are inherently quite focused (e.g. 8 kHz).

8.3 Further Work

Based on the conclusions drawn in Section 8.2, further experiments are proposed to lead on from the work presented in the current thesis. Firstly, it appears that the effect of vertical inter-channel decorrelation for upmixing may be weak with four decorrelated pairs (Chapter 7). As a result, it is proposed that a following experiment should be conducted with less vertically decorrelated pairs. In other words, comparing vertical decorrelation from the front left and right loudspeakers against decorrelation from rear left and right vertical pairs alone, in order to observe whether the sense of LEV can be increased. This is to try and preserve the spectral localisation cues that appear to be associated with vertical interchannel decorrelation, as displayed in Chapter 5 for discrete pairs of vertically-spaced loudspeakers. It was also shown in Chapter 7 that changes in LEV broadly correspond with changes in energy from behind; therefore, it might be assumed that vertical decorrelation of the rear channels only (±110°) will be most effective, particularly when direct sound is present from in front.

A second experiment proposed is related to the potential binaural rendering of VIS. In Chapter 5, it was seen that vertical decorrelation had a clear influence on interaural differences, particularly at wider azimuth angles (±110°). It is thought that these are from the head-shadowing of the main-layer loudspeaker signal, resulting in vertical decorrelation contributing to a decrease of IACC from the decorrelated height-channel signal. It would be interesting to investigate whether simply applying interaural level difference (ILD), interaural time difference (ITD) and interaural decorrelation (decreasing IACC) can control the perception of VIS over headphones, based on the octave-band observations made in Chapter 5. To achieve this, vertically-decorrelated conditions would be binauralised as reference signals, much like the process described in Chapter 5 – these reference signals would then be octave-band analysed for ILD, ITD and IACC, with the results applied to an unprocessed sample. Comparing the artificially generated 'VIS' signals against the binaural references could potentially provide further insight into how we perceive VIS and vertical interchannel decorrelation.

Lastly, the vertical decorrelation experiments described in the present thesis have been limited to a small number of decorrelation methods: complementary comb-filtering (Lauridsen, 1954; Schroeder 1958), phase randomisation by all-pass filtering (Kendall, 1995), and deterministic delay networks for both amplitude and phase decorrelation (Zotter & Frank, 2013). Given the similarity between the results for the two methods assessed in Chapter 4, it suggests that VIS may be controlled by the interchannel cross-correlation (ICC) between a single pair of vertically-arranged loudspeakers, rather than the specific decorrelation method that is used to control ICC. Further testing is required to explore this hypothesis; therefore, it is proposed that other methods of decorrelation should also be assessed for vertical decorrelation. These could include random critical band delays (Bouéri & Kyirakakis, 2004) and random amplitude differences as a function of frequency (Faller & Baumgarte, 2003; Fink et al., 2015). Furthermore, the absolute grading in Chapter 4 was only observed for one method of decorrelation (phase randomisation by all-pass filtering) – it would also be useful to observe whether the vertical phantom image is perceived differently for different decorrelation techniques.

APPENDIX A: HULTI-GEN[6]

A.0 Abstract

This appendix describes HULTI-GEN (Huddersfield Universal Listening Test Interface Generator), a tool based in Cycling '74's Max. HULTI-GEN is a user-customisable environment, which takes user-defined parameters (e.g. the number of trials, stimuli and scale settings) and automatically constructs an interface for comparing auditory stimuli, whilst also randomising the stimuli and trial order. To assist the user, templates based on ITU-R recommended methods have been included. As the recommended methods are often adjusted for different test requirements, HULTI-GEN also supports flexible editing of these presets. Furthermore, some existing techniques have been summarised within this appendix, including their restrictions and how they might be altered through using HULTI-GEN. HULTI-GEN is freely available online at: http://www.hud.ac.uk/research/researchcentres/mtprg/projects/apl/

A.1 Introduction

Within the audio industry, efficient and reliable means of assessing auditory attributes are vital to helping us understand our perception of sound. Subjective listening tests on a computer are often used to carry out such assessments, where a listener is presented with a graphical user interface and asked to grade auditory stimuli one way or another. There is no 'one method fits all' approach to listening tests, and there are cases where the formats of existing recommendations and methods (as discussed briefly in Section A.2 below) require adjustment to test for novel attributes. When designing a listening test for these attributes, a robust and repeatable testing method is often thought to be the most important consideration. A few key features that require thought during test design are: the scale on which the user is grading, whether there are

[6] Gribben, C. & Lee, H. (2015). Towards the Development of a Universal Listening Test Interface Generator in Max. Presented at the *138th Convention of the Audio Engineering Society*.

audible reference or anchor points on the scale, and also how a large number of stimuli might be split into separate trials, to make testing more manageable for the listener.

The content of this appendix describes HULTI-GEN (Huddersfield Universal Listening Test Interface Generator), a Cycling '74 Max-based tool that generates a listening test interface from user-defined parameters. The main aim of HULTI-GEN is to address the key features of test design mentioned above, by delivering a step-by-step process to customise and build a listening test interface. This has resulted in an adaptable tool that can be used for a broad range of listening test scenarios, as well as being a platform for novel test development.

Initially developed for final year students at the University of Huddersfield, HULTI-GEN can be useful to both experienced and inexperienced audio researchers alike. In particular, those who want to pilot various approaches for comparing auditory stimuli, as well as develop new techniques, can benefit from the user-friendly flexibility of the software. Templates based on commonly used listening test methods have also been included, along with the ability to alter these presets.

A.2 Listening Test Methods

This section provides a brief summary and discussion on common listening test methods.

As defined in ITU-R Recommendation BS.1116-3 (ITU-R, 2015a), the double-blind triple-stimulus test method features a set of three auditory stimuli per trial. Two of the stimuli are graded for impairments against a third reference signal; one of those being graded is a hidden case of the reference signal. These judgments are made on a continuous five-grade scale, with descriptive anchors ranging from 'Imperceptible' (5.0) to 'Very annoying' (1.0). The method is commonly used to compare and detect small impairments of audio quality between high quality audio samples. It is recommended that 'Basic audio quality' be set as the single, global attribute during testing, although an experimenter may also choose to define and evaluate other

attributes. Potential attributes in the document, include, 'Stereophonic image quality', 'Timbral quality', and 'Localisation quality'. Although these attributes also relate to audio quality, the grading scale might benefit from alternative labelling, which is something to consider during test design. The five-grade scale implemented here is also similar to the Mean Opinion Score (MOS) standard, which has been used for assessing the transmission quality of audio (ITU-T, 1996).

In contrast to the double-blind triple-stimulus, a format named the "MUltiple Stimulus test with Hidden Reference and Anchor (MUSHRA)" concerns medium to large impairments of intermediate audio quality, as described in ITU-R Recommendation BS.1534-2 (ITU-R, 2015b). While MUSHRA also assesses perceived audio quality, the test features a multi-comparison layout instead, and was designed to test for the impairments of audio codec processing. These judgments are made on a continuous scale of 0-100, with five grading regions from Bad (0-20) through to Excellent (80-100). For each trial, multiple stimuli are compared against a high quality, unprocessed reference. Amongst the stimuli, three anchor samples derived from the reference are included – two with low-pass filters at 3.5 and 7 kHz (low and intermediate anchors), and the third is the unprocessed reference (hidden reference i.e. high anchor at 100).

Variations of both the main test methods described above are often used for assessing auditory attributes other than audio quality (e.g. spatial characteristics, such as, apparent source width and listener envelopment). For example, if there were large spatial differences between stimuli, a multi-comparison method based on MUSHRA might be used. In this instance, the choice of reference is subjective and requires rational consideration. It would be invalid for an investigator to use the stimulus they perceive to be most spacious as the high anchor reference (i.e. 100 on a scale of 0-100), as a test subject may perceive another stimulus to be more so. Therefore, movement of the reference from 100 would reduce bias and give room for the listener to grade higher. One solution for this has been to use a continuous bipolar scale (e.g. -50 to 50) where the reference is at 0, similar to the seven-grade comparison scale in BS.1284 (ITU-R, 2003).

This type of scale could also feature a semantic differential grading system, where two opposing adjectives are at either end of the scale (i.e. louder and quieter).

All of the examples discussed so far use a continuous scale. A potential issue with this scale-type is a lack of control over the way a subject grades it, in terms of the different spread of scores between listeners – this is usually addressed by normalisation of the results (ITU-R, 2015b). Alternatively, a practical method to guide the use of a scale has been to introduce additional audible anchors, helping to audibly define the scale limits and support the labelling (George et al., 2010). However, this technique may bring unwanted bias and affect the scale's continuous nature, removing the option to normalise the data. An ABX test is a simple alternative for detecting slight perceptual difference between two samples, without the need for a scale. It is the same triple-stimulus format as (ITU-R, 2015a), but instead of grading, the listener is forced to identify which of the two stimuli is the hidden reference. Likewise, a pairwise comparison test is also used for small impairments, but does not feature a reference (De Man & Reiss, 2013).

As there are many factors contributing to the design of a listening test, no single test method is correct for all situations, and new formats may need to be developed as novel auditory attributes are proposed. When contemplating test features, for instance, the inclusion of audible reference anchors or suitable labelling during test design, a listening test interface tool that allows the user to easily alter them would be of great use.

A.3 Software

A.3.1 Existing Software

Considering the thoughts in Section A.2 above, a number of adaptable listening test interfaces are already in existence (Ciba, Wlodarski & Maempel, 2009; Giner, 2013; De Man & Reiss, 2014); however, they all have their limitations in terms of customisability, with some only available commercially. A large proportion of the open-source listening test programs are based

in MATLAB, due to its existing functions for handling audio and creating graphical user interfaces. Despite this functionality, it has been found that programming knowledge can be advantageous when preparing and designing a test in MATLAB-developed software, which can limit the adaptation and flexibility for a less experienced user. There have also been occasional instances of incompatibility with different versions of MATLAB, as well as some software being limited to certain operating systems.

A.3.2 Development Software: Cycling '74 Max

Given these restrictions, the tool described in this brief (HULTI-GEN) is a patch that has been developed using the software Max (also referred to as Max/MSP, of which MSP is the Max Signal Processing module). Max is a cross-platform visual programming language from Cycling '74, specifically designed for developing music and multimedia applications. Some useful features of Max, which have contributed in part to the initial development of HULTI-GEN, include: real-time manipulation of digital audio signals including multi-channel playback, an object/modular-based environment that encourages rapid prototyping, the ability to generate new objects (i.e. sliders) for a user without background editing, the capacity to export developed software as a standalone application or collective file, and the availability of graphical objects that allow the user to easily input and store data.

Many listening tests have already been conducted using reliable interfaces developed in Max. However, these patches are often restricted to a particular testing method, making them an inflexible tool to the layman who has no prior knowledge of programming in Max. As far as the authors are aware, no such universal listening test generator, similar to those mentioned in Section A.3.1, exists in the Max environment at the time of writing. Therefore, HULTI-GEN is considered to be a useful tool for both professionals and students conducting auditory research, no matter their level of experience in listening test design.

A.4 HULTI-GEN Overview

HULTI-GEN is a flexible listening test tool that generates a graphical user interface (GUI) for audio comparison tests. It has a focus on simplifying the user-experience for both the test designer and test subject, and provides a customisable foundation on which to build novel testing formats. At present, HULTI-GEN has templates based on the ITU-R Recommendations 1534-2 (MUSHRA) and 1116-2, as described above.

No previous knowledge of the running software (Max) is needed to use HULTI-GEN, as all end-users interact solely with the GUI. The tool features a simple drag-and-drop system to import and store the audio filenames of the stimuli for testing, as well as a step-by-step guide for constructing and editing a listening test. This allows the test designer to efficiently prototype different test methods during pilot experimentation, through the swift adjustment of various test parameters and instantaneous generation of a new interface.

In an attempt to address some of the listening test limitations discussed in Section A.2, key features of HULTI-GEN include the following:

- Full randomisation of the trial order and of stimuli within trials for each test conducted.
- Listening test templates of established methods are included, which can also be altered.
- A drag-and-drop function to quickly import the stimuli filenames when preparing a test.
- Definition of the scale limits and resolution, as well as the starting position of the slider.
- Flexible customisation of the scale labelling.
- The option to include an audible reference/ anchor at varying positions on the scale.

A basic flow-chart detailing the listening test design process of HULTI-GEN's GUI can be seen in Figure A.1. It demonstrates the ability to easily navigate from the Main Menu and edit many parameters, for instance, the distribution of stimuli, use and position of a reference, alterations to the grading scale, and importing new stimuli. There are also options to load the saved settings

or to create a completely new interface, which guides the user through the process and considerations to make.

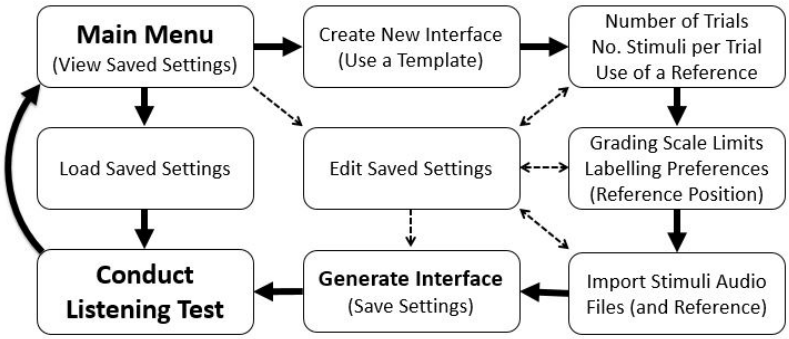

Figure A.1 Flow-chart of the process in HULTI-GEN.

An example of a listening test interface that has been generated in a prototype of HULTI-GEN can be seen in Figure A.2 – it features a multi-comparison format similar to MUSHRA, but with a bipolar scale, a reference signal at 0 and customised labelling.

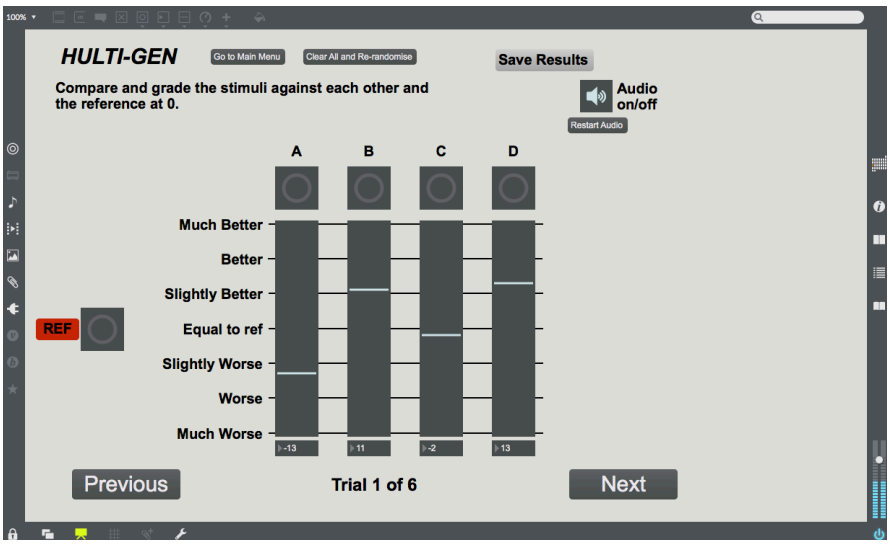

Figure A.2 An example HULTI-GEN interface in Max 7.

A.5 Conclusion and Future Work

HULTI-GEN is a Max-based tool for generating customisable listening test interfaces. It is thought the tool will be of interest (and use) to all audio researchers, particularly those interested in developing new testing methods. HULTI-GEN is freely available from:

http://www.hud.ac.uk/research/researchcentres/mtprg/projects/apl/

In future, the tool would benefit from the incorporation of additional features, such as a control for the subject to adjust the loop size and the capability of multi-channel playback, giving an increased compliance with the recommended documents. There is also the potential to include other test methods in further development, for example, Pairwise Comparison, ABX, Mean Opinion Score and a way to help elicit novel auditory attributes.

REFERENCES

Adami, A., Brand, L., & Herre, J. (2017). Investigations Towards Plausible Blind Upmixing of Applause Signals. Presented at the *142nd Convention of the Audio Engineering Society*. Paper Number 9750.

Algazi, V. R., Avendano, C., & Duda, R. O. (2001). Elevation Localization and Head-Related Transfer Function Analysis at Low Frequencies. *The Journal of the Acoustical Society of America, 109*(3), 1110-1122. doi:10.1121/1.1349185

Auro Technologies (2015a). Auro-3D Home Theater Setup: Installation Guidelines. Retrieved from https://www.auro-3d.com/wp-content/uploads/documents/Auro-3D-Home-Theater-Setup-Guidelines_lores.pdf (Accessed May 2018).

Auro Technologies (2015b). AUROMAX: Next Generation Immersive Sound System. Retrieved from https://www.auro-3d.com/wp-content/uploads/documents/AuroMax_White_Paper_24112015.pdf (Accessed May 2018).

Avendano, C., & Jot, J-M. (2002). Ambience Extraction and Synthesis from Stereo Signals for Multi-Channel Audio Up-mix. Paper presented at the *2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi:10.1109/ICASSP.2002.5745013

Avendano, C., & Jot, J-M. (2004). A Frequency-Domain Approach to Multichannel Upmix. *Journal of the Audio Engineering Society, 52*(7/8), 740-749.

Barbour, J. L. (2003). Elevation Perception: Phantom Images in the Vertical Hemi-sphere. Presented at the *Audio Engineering Society 24th International Conference on Multichannel Audio: The New Reality*, Conference Paper 14.

Bauer, B. B. (1961). Phasor Analysis of Some Stereophonic Phenomena. *The Journal of the Acoustical Society of America, 33*(11), 1536-1539. doi:10.1109/TAU.1962.1161613

Barron, M. (1971). The Subjective Effects of First Reflections in Concert Halls – The Need for Lateral Reflections. *The Journal of Sound and Vibration, 15*(4), 475-494. doi:10.1016/0022-460X(71)90406-8

Barron, M., & Marshall, A. H. (1981). Spatial Impression due to Early Lateral Reflections in Concert Halls: The Derivation of a Physical Measure. *The Journal of Sound and Vibration, 77*(2), 211-232. doi:10.1016/S0022-460X(81)80020-X

Bech, S., & Zacharov, N. (2006). *Perceptual Audio Evaluation – Theory, Method and Application.* Chichester, England: John Wiley & Sons. p. 113.

Bennett, J. C., Barker, K., & Edeko, F. O. (1985). A New Approach to the Assessment of Stereophonic Sound System Performance. *Journal of the Audio Engineering Society, 33*(5), 314-321.

Berg, J., & Rumsey, F. (2006). Identification of Quality Attributes of Spatial Audio by Repertory Grid Technique. *Journal of the Audio Engineering Society, 54*(5), 365-379.

Bernstein, L. R., & Trahiotis, C. (2002). Enhancing Sensitivity to Interaural Delays at High Frequencies by using "Transposed Stimuli". *The Journal of the Acoustical Society of America, 112*(3), 1026-1036. doi:10.1121/1.1497620

Blau, M. (2002). Difference Limens for Measures of Apparent Source Width. Presented at *Forum Acousticum Sevilla 2002: 3rd European Congress on Acoustics*.

Blauert, J. (1969/70). Sound Localization in the Median Plane. *Acustica, 22*(4), 205–213.

Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA & London, England: MIT Press. pp. 1-5, 63-77, 116-155 & 203-271.

Blauert, J., & Lindemann, W. (1986a). Auditory Spaciousness: Some Further Psychoacoustic Analyses. *The Journal of the Acoustical Society of America, 80*(2), 533-542. doi:10.1121/1.394048

Blauert, J., & Lindemann, W. (1986b). Spatial Mapping of Intracranial Auditory Events for Various Degrees of Interaural Coherence. *The Journal of the Acoustical Society of America, 79*(3), 806-813. doi:10.1121/1.393471

Bouéri, M., & Kyirakakis, C. (2004). Audio Signal Decorrelation Based on a Critical Band Approach. Presented at the *117th Convention of the Audio Engineering Society*. Paper Number 6291.

Bradley, J. S., & Soulodre, G. A. (1995). Objective Measures of Listener Envelopment. *The Journal of the Acoustical Society of America, 98*(5), 2590-2597. doi:10.1121/1.413225

Breebaart, J., & Faller, C. (2007). *Spatial Audio Processing: MPEG Surround and Other Applications.* Chichester, England: John Wiley & Sons. pp. 82-84.

British Standards Institution. (2009). *ISO 3382-1:2009: Acoustics – Measurement of room acoustic parameters: Part 1: Performance spaces*. London, England: BSI.

Brungart, D. S., & Rabinowitz, W. M. (1999). Auditory Localization of Nearby Sources. Head-Related Transfer Functions. *The Journal of the Acoustical Society of America, 106*(3), 1465-1479. doi:10.1121/1.427180

Cabrera, D., & Tilley, S. (2003). Vertical Localization and Image Size Effects in Loudspeaker Reproduction. Presented at the *Audio Engineering Society 24th International Conference: Multichannel Audio—The New Reality*. Conference Paper 46.

Chun, C. J., Kim, H. K., Choi, S. H., Jang, S-J., & Lee, S-P. (2011). Sound Source Elevation Using Spectral Notch Filtering and Directional Band Boosting in Stereo Loudspeaker Reproduction. *IEEE Transactions on Consumer Electronics, 57*(4), 1915-1920. doi:10.1109/TCE.2011.6131171

Ciba, S., Wlodarski, A., & Maempel, H-J. (2009). WhisPER – A New Tool for Performing Listening Tests. Presented at the *126th Convention of the Audio Engineering Society*. Paper Number 7749.

Damaske, P. (1971). Head-Related Two-Channel Stereophony with Loudspeaker Reproduction. *Journal of the Acoustical Society of America, 50*(4B), 1109-1115. doi:10.1121/1.1912742

De Man, B., & Reiss, J. D. (2013). A Pairwise and Multiple Stimuli Approach to Perceptual Evaluation of Microphone Types. Presented at the *134th Convention of the Audio Engineering Society*. Paper Number 8837.

De Man, B., & Reiss, J. D. (2014). APE: Audio Perceptual Evaluation Toolbox for MATLAB. Presented at the *136th Convention of the Audio Engineering Society*. Engineering Brief 151.

Dolby Laboratories. (2014). Dolby Atmos Next-Generation Audio for Cinema (Issue 3). Retrieved from https://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-next-generation-audio-for-cinema-white-paper.pdf (Accessed May 2018).

Dolby Laboratories. (2017). Dolby Atmos Home Theater Installation Guidelines. Retrieved from https://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-home-theater-installation-guidelines.pdf (Accessed May 2018).

Engdegård, J., Purnhagen, H., Rödén, J., & Liljeryd, L. (2004). Synthetic Ambience in Parametric Stereo Coding. Presented at the *116th Convention of the Audio Engineering Society*. Paper Number 6074.

Evjen, P., Bradley, J. S., & Norcross, S. G. (2001). The Effect of Late Reflections from Above and Behind on Listener Envelopment. *Applied Acoustics, 62*(2), 137-153. doi:10.1016/S0003-682X(00)00053-0

Faller, C. (2006). Parametric Multichannel Audio Coding: Synthesis of Coherence Cues. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(1), 299-310. doi:10.1109/TSA.2005.854105

Faller, C., & Baumgarte, F. (2003). Binaural Cue Coding – Part II: Schemes and Applications. *IEEE Transactions on Speech and Audio Processing, 11*(6), 520-531. doi:10.1109/TSA.2003.818108

Farina, A. (2000). Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique. Presented at the *110th Convention of the Audio Engineering Society*. Paper Number 5093.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods, 39*(2), 175-191. doi:10.3758/BF03193146

Feddersen, W. E., Sandel, T. T., Teas, D. C., & Jeffress, L. A. (1958). Localization of High-Frequency Tones. *The Journal of the Acoustical Society of America, 29*(9), 988-991. doi:10.1121/1.1909356

Fink, M., Kraft, S., & Zölzer, U. (2015). Downmix-Compatible Conversion from Mono to Stereo in Time- and Frequency-Domain. Presented at the *18th International Conference on Digital Audio Effects (DAFx-15)*.

Ferguson, S., & Cabrera, D. (2005). Vertical Localization of Sound from Multiway Loudspeakers. *Journal of the Audio Engineering Society, 52*(3), 163-173.

Furuya, H., Fujimoto, K., Ji, C. Y., & Higa, N. (2001). Arrival Direction of Late Sound and Listener Envelopment. *Applied Acoustics, 62*(2), 125-136.

Furuya, H., Fujimoto, K., Takeshima, Y., & Nakamura, H. (1995). Effect of Early Reflections from Upside on Auditory Envelopment. *Journal of the Acoustical Society of Japan (E), 16*(2), 97-104.

Furuya, H., Fujimoto, K., Wakuda, A., & Nakano, Y. (2005). The Influence of Total and Directional Energy of Late Sound on Listener Envelopment. *Acoustical Science and Technology, 26*(2), 208-211. doi:10.1250/ast.26.208

Furuya, H., Fujimoto, K., & Wakuda, A. (2008). Psychological Experiments on Listener Envelopment when both the Early-to-Late Sound Level and Directional Late Energy Ratios are Varied, and Consideration of Calculated LEV in Actual Halls. *Applied Acoustics, 69*(11), 1085-1095. doi:10.1016/j.apacoust.2007.06.006

Gardner, M. B., & Gardner, R. S. (1973). Problem of Localization in the Median Plane: Effect of Pinnae Cavity Occlusion. *The Journal of the Acoustical Society of America, 53*(2), 400-408. doi:10.1121/1.1913336

Gardner, B., & Martin, K. (1994). HRTF Measurements of a Kemar Dummy-Head Microphone. *MIT Media Lab Perceptual Computing*, Technical Report #280.

George, S., Zielinski, S., Rumsey, F., Jackson, P., Conetta, R., Dewhirst, M., Meares, D., & Bech, S. (2010). Development and Validation of an Unintrusive Model for Predicting the Sensation of Envelopment Arising from Surround Sound Recordings. *Journal of the Audio Engineering Society, 58*(12), 1013-1031.

Giner, A. V. (2013). Scale – A Software Tool for Listening Experiments. Presented at the *AIA/DAGA Conference on Acoustics*.

Grantham, D. W. (1984). Interaural Intensity Discrimination: Insensitivity at 1000 Hz. *The Journal of the Acoustical Society of America, 75*(4), 1191-1194. doi:10.1121/1.390769

Gribben, C., & Lee, H. (2014). The Perceptual Effects of Horizontal and Vertical Interchannel Decorrelation Using the Lauridsen Decorrelator. Presented at the *136th Convention of the Audio Engineering Society*. Paper Number 9027.

Gribben, C., & Lee, H. (2015). Towards the Development of a Universal Listening Test Interface Generator in Max. Presented at the *138th Convention of the Audio Engineering Society*. Engineering Brief 187.

Gribben, C., & Lee, H. (2017a). The Perceptual Effect of Vertical Interchannel Decorrelation on Vertical Image Spread at Different Azimuth Positions. Presented at the *142nd Convention of the Audio Engineering Society*. Paper Number 9747.

Gribben, C., & Lee, H. (2017b). A Comparison between Horizontal and Vertical Interchannel Decorrelation. *Journal of Applied Sciences, 7*(11), 1202. doi:10.3390/app7111202

Gribben, C., & Lee, H. (2018). The Frequency and Loudspeaker-Azimuth Dependencies of Vertical Interchannel Decorrelation on Vertical Image Spread. *Journal of the Audio Engineering Society,* (accepted May 2018).

Griesinger, D. (1999). Objective Measures of Spaciousness and Envelopment. Presented at the *16th International Conference of the Audio Engineering Society: Spatial Sound Reproduction*. Paper Number 16-003.

Haas, H. (1972). The Influence of a Single Echo on the Audibility of Speech. *Journal of the Audio Engineering Society, 20*(2), 146-159.

Hamasaki, K. (2003). Multichannel Recording Techniques for Reproducing Adequate Spatial Impression. Presented at the *Audio Engineering Society 24th International Conference: Multichannel Audio, The New Reality*. Conference Paper 27.

Hawksford, M. O. J., & Harris, N. (2002). Diffuse Signal Processing and Acoustic Source Characterization for Applications in Synthetic Loudspeaker Arrays. Presented at the *112th Convention of the Audio Engineering Society*. Paper Number 5612.

Hebrank, J., & Wright, D. (1974). Spectral Cues Used in the Localization of Sound Sources on the Median Plane. *The Journal of the Acoustical Society of America, 56*(6), 1829-1834. doi:10.1121/1.1903520

Herre, J., Hilpert, J., Kuntz, A., & Plogsties, J. (2014). MPEG-H – The New Standard for Universal Spatial/3D Audio Coding. *Journal of the Audio Engineering Society, 62*(12), 821-830.

Herre, J., Hilpert, J., Kuntz, A., & Plogsties, J. (2015). MPEG-H 3D Audio – The New Standard for Coding Immersive Spatial Audio. *IEEE Journal of Selected Topics in Signal Processing, 9*(5), 770-779.

Herre, J., Kjörling, K., Breebaart, J., Faller, C., Disch, S., Purnhagen, H., Koppens, J., Hilpert, J., Rödén, J., Oomen, W., Linzmeier, K., & Chong, K. S. (2008). MPEG Surround – The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding. *Journal of the Audio Engineering Society, 56*(11), 932-955.

Hidaka, T., Beranek, L. L., & Okano, T. (1995). Interaural Cross-Correlation (IACC), Lateral Fraction (LF), and Low- and High-Frequency Sound Levels (G) as Measures of Acoustical Quality in Concert Halls. *The Journal of the Acoustical Society of America, 97*(5), 3319-3319. doi:10.1121/1.412847

Howard, D. M., & Angus, J. (2017). *Acoustics and Psychoacoustics* (5th ed.). Abingdon, England: Routledge. pp. 108-118 & 378-380.

Irwan, R., & Aarts, R. M. (2002). Two-to-Five Channel Sound Processing. *Journal of the Audio Engineering Society, 50*(11), 914-926.

ITU-R. (2003). Recommendation ITU-R BS.1284-1. General Methods for the Subjective Assessment of Sound Quality. *International Telecommunications Union.*

ITU-R. (2012). Recommendation ITU-R BS.775-3: Multichannel Stereophonic Sound System With and Without Accompanying Picture. *International Telecommunications Union*.

ITU-R. (2015a). Recommendation ITU-R BS.1116-3: Methods for the Subjective Assessment of Small Impairments in Audio Systems. *International Telecommunications Union.*

ITU-R. (2015b). Recommendation ITU-R BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems. *International Telecommunications Union*.

ITU-R. (2015c). Recommendation ITU-R BS.1770-4: Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level. *International Telecommunications Union*.

ITU-T. (1996). Recommendation ITU-T P.800: Methods for Subjective Determination of Transmission Quality. *International Telecommunications Union.*

Johnson, D., Harker, A., & Lee, H. (2015). HAART: A New Impulse Response Toolbox for Spatial Audio Research. Presented at the *138th Convention of the Audio Engineering Society*. Engineering Brief 190.

Jot, J-M., & Avendano, C. (2003). Spatial Enhancement of Audio Recordings. Presented at the *Audio Engineering Society 23rd International Conference: Signal Processing in Audio Recording and Reproduction*. Conference Paper 15.

Keet, W. de V. (1968). The Influence of Early Reflections on Spatial Impression. Presented at the *6th International Congress on Acoustics*.

Kendall, G. S. (1995). The Decorrelation of Audio Signals and its Impact on Spatial Imagery. *Computer Music Journal, 19*(4), 71-87. doi:10.2307/3680992

Kietz, H. (1953). Das Räumliche Hören. *Acustica, 3*(2), 73-86.

Kraft, S., & Zölzer, U. (2015). Stereo Signal Separation and Upmixing by Mid-Side Decomposition in the Frequency-Domain. Presented at the *18th International Conference on Digital Audio Effects (DAFx-15)*.

Kraft, S., & Zölzer, U. (2016). Low-Complexity Stereo Signal Decomposition and Source Separation for Application in Stereo to 3D Upmixing. Presented at the *140th Convention of the Audio Engineering Society*. Convention Paper 9586.

Kuntz, A. Disch, S, Bäckström, T., Robilliard, J., & Uhle, C. (2011). The Transient Steering Decorrelator Tool in the upcoming MPEG Unified Speech and Audio Coding Standard. Presented at the *131st Convention of the Audio Engineering Society*. Paper Number 8533.

Lachenmayr, W., Haapaniemi, A., & Lokki, T. (2016). Direction of Late Reverberation and Envelopment in Two Reproduced Berlin Concert Halls. Presented at the *140th Convention of the Audio Engineering Society*. Paper Number 9503.

Laitinen, M-V., Kuech, F., Disch, S., & Pulkki, V. (2011). Reproducing Applause-Type Signals with Directional Audio Coding. *Journal of the Audio Engineering Society, 59*(1/2), 29-43.

Lauridsen, H. (1954). Nogle Forsøg Reed Forskellige Former Rum Akustik Gengivelse. *Ingeniøren, 47*, 906.

Lee, H. (2012). Apparent Source Width and Listener Envelopment in Relation to Source-Listener Distance. Presented at the *52nd International Conference of the Audio Engineering Society: Sound Field Control – Engineering and Perception*. Paper Number 3-1.

Lee, H. (2016a). Perceptually Motivated 3D Diffuse Field Upmixing. Presented at the *Audio Engineering Society Conference on Sound Field Control*. Paper Number 3-2.

Lee, H. (2016b). 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), 1003-1013. doi:10.17743/jaes.2016.0052

Lee, H. (2017a). Sound Source and Loudspeaker Base Angle Dependency of the Phantom Image Elevation Effect. *Journal of the Audio Engineering Society, 65*(9), 733-748. doi:10.17743/jaes.2017.0028

Lee, H. (2017b). Perceptually Motivated Amplitude Panning (PMAP) for Accurate Phantom Image Localisation. Presented at the *142nd Convention of the Audio Engineering Society*. Paper Number 9770.

Lee, H., & Gribben, C. (2014). Effect of Vertical Microphone Layer Spacing for a 3D Microphone Array. *Journal of the Audio Engineering Society, 62*(12), 870-884. doi:10.17743/jaes.2014.0045

Lee, H., Gribben, C., & Wallis, R. (2014). Psychoacoustic Considerations in Surround Sound with Height. Presented at the *28th Tonmeistertagung – VDT International Convention*.

Lee, H., & Rumsey, F. (2013). Level and Time Panning Images for Musical Sources. *Journal of the Audio Engineering Society, 61*(12), 978-988.

Li, Y., & Driessen, P. F. (2005). An Unsupervised Adaptive Filtering Approach of 2-To-5 Channel Upmix. Presented at the *119th Convention of the Audio Engineering Society*. Paper Number 6611.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., & Guzman, S. J. (1999). The Precedence Effect. *The Journal of the Acoustical Society of America, 106*(4), 1633-1654. doi:10.1121/1.427914

Martens, W. L., Braasch, J., & Woszczyk, W. (2004). Identification and Discrimination of Listener Envelopment Percepts Associated with Multiple Low-Frequency Signals in Multichannel Reproduction. Presented at the *117th Convention of the Audio Engineering Society*. Paper Number 6229.

Mason, R., Brookes, T., & Rumsey, F. (2003). Creation and Verification of a Controlled Experimental Stimulus for Investigating Selected Perceived Spatial Attributes. Presented at the *114th Convention of the Audio Engineering Society*. Paper Number 5771.

Mason, R., Brookes, T., & Rumsey, F. (2005). Frequency Dependency of the Relationship between Perceived Auditory Source Width and the Interaural Cross-Correlation Coefficient for Time-Invariant Stimuli. *The Journal of the Acoustical Society of America, 117*(3), 1337-1350. doi:10.1121/1.1853113

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of Box Plots. *The American Statistician, 32*(1), 12-16. doi:10.1080/00031305.1978.10479236

Mills, A. W. (1958). On the Minimum Audible Angle. *The Journal of the Acoustical Society of America, 30*(4), 237-246. doi:10.1121/1.1909553

Mironovs, M., & Lee, H. (2017). The Influence of Source Spectrum and Loudspeaker Azimuth on Vertical Amplitude Panning. Presented at the *142nd Convention of the Audio Engineering Society*. Convention Paper 9782.

Mohamed, M. I. J., & Cabrera, D. (2008). Human Sensitivity to Interaural Phase Difference for Very Low Frequency Sound. Presented at *Acoustics 2008 – Acoustics and Sustainability*.

Moore, B. C. J. (2012). *An Introduction to the Psychology of Hearing* (6th ed.). Bingley, England: Emerald. pp. 247-252.

Morimoto, M. (1997). The Role of Rear Loudspeakers in Spatial Impression. Presented at the *103rd Convention of the Audio Engineering Society.* Paper Number 4554.

Morimoto, M. (2002). The Relation Between Spatial Impression and the Precedence Effect. In *Proceedings of the 8th International Conference on Auditory Display (ICAD2002)*.

Morimoto, M., & Iida, K. (1998). Effects of Front / Back Energy Ratios of Early and Late Reflections on Listener Envelopment. *The Journal of the Acoustical Society of America, 103*(5), 7-8. doi:10.1121/1.422799

Morimoto, M., Iida, K., & Sakagami, K. (2001). The Role of Reflections from Behind the Listener in Spatial Impression. *Applied Acoustics, 62*(2), 109-124. doi:10.1016/S0003-682X(00)00051-7

Morimoto, M., Yairi, M., Iida, K., & Itoh, M. (2003). The Role of Low Frequency Components in Median Plane Localization. *Acoustical Science and Technology, 24*(2), 76-82. doi:10.1250/ast.24.76

Murtaza, A., Herre, J., Paulus, J., Terentiv, L., Fuchs, H., & Disch, S. (2015). ISO/MPEG-H 3D Audio: SAOC-3D decoding and rendering. Presented at the *138th Convention of the Audio Engineering Society*. Paper Number 9434.

Nielsen, S. H. (1993). Auditory Distance Perception in Different Rooms. *Journal of the Audio Engineering Society, 41*(10), 755-770.

Okano, T., Hidaka, T., & Beranek, L. (1994). Psychoacoustic Experiment to Determine the Influence of Inter-aural Cross-Correlation Coefficient and Sound Pressure Level of Low Frequencies on the Apparent Source Width in Concert Halls. Presented at the *128th Meeting of the Acoustical Society of America*.

Penniman, R. (2014). A General-Purpose Decorrelation Algorithm with Transient Fidelity. Presented at the *137th Convention of the Audio Engineering Society*. Paper Number 9170.

Perrott, D. R., Musicant, A., & Schwethelm, B. (1980). The Expanding-Image Effect: The Concept of Tonal Volume Revisited. *Journal of Auditory Research, 20*(1), 43-56.

Perrott, D. R., & Buell, T. N. (1982). Judgements of Sound Volume: Effects of Signal Duration, Level, and Interaural Characteristics on the Perceived Extensity of Broadband Noise. *The Journal of the Acoustical Society of America, 72*(5), 1413-1417. doi:10.1121/1.388447

Pihlajamäki, T., Santala, O., & Pulkki, V. (2014). Synthesis of Spatially Extended Virtual Sources with Time-Frequency Decomposition of Mono Signals. *Journal of the Audio Engineering Society, 62*(7/8), 467-484. doi:10.17743/jaes.2014.0031

Potard, G., & Burnett, I. (2004). Decorrelation Techniques for the Rendering of Apparent Sound Source Width in 3D Audio Displays. Presented at the *7th International Conference on Digital Audio Effects (DAFx'04)*.

Pratt, C. C. (1930). The Spatial Character of High and Low Tones. *Journal of Experimental Psychology, 13*(3), 278-285. doi:10.1037/h0072651

Pulkki, V. (1997). Virtual Sound Source Positioning Using Vector Base Amplitude Panning. *Journal of the Audio Engineering Society, 45*(6), 456-466.

Pulkki, V. (2001). Localization of Amplitude-Panned Virtual Sources II: Two- and Three-Dimensional Panning. *Journal of the Audio Engineering Society, 49*(9), 753-767.

Pulkki, V. (2007). Spatial Sound Reproduction with Directional Audio Coding. *Journal of the Audio Engineering Society, 55*(6), 503-516.

Pulkki, V., & Karjalainen, M. (2001). Localization of Amplitude-Panned Virtual Sources I: Stereophonic Panning. *Journal of the Audio Engineering Society, 49*(9), 739-752.

Rayleigh, Lord. (1907) On our Perception of Sound Direction. *Philosophical Magazine, 13*(74), 214-232. doi:10.1080/14786440709463595

Robotham, T., Stephenson, M., & Lee, H. (2016). The Effect of a Vertical Reflection on the Relationship between Preference and Perceived Change in Timbral and Spatial Attributes. Presented at the *140th Convention of the Audio Engineering Society*, Paper Number 9547.

Roffler, S. K., & Butler, R. A. (1968a). Factors that Influence the Localization of Sound in the Vertical Plane. *The Journal of the Acoustical Society of America, 43*(6), 1255-1259. doi:10.1121/1.1910976

Roffler, S. K., & Butler, R. A. (1968b). Localization of Tonal Stimuli in the Vertical Plane. *The Journal of the Acoustical Society of America, 43*(6), 1260-1266. doi:10.1121/1.1910977

Rosenzweig, M., & Rosenblith, W. (1950). Some Electrophysiological Correlates of the Perception of Successive Clicks. *Journal of the Acoustical Society of America, 22*(6), 878-880. doi:10.1121/1.1906709

Rumsey, F. (2001). *Spatial Audio.* Oxford, England: Focal Press. pp. 35-38, 53 & 196-201.

Sayers, B. M. (1964). Acoustic-Image Lateralization Judgments with Binaural Tones. *The Journal of the Acoustical Society of America, 36*(5), 923. doi:10.1121/1.1919121

Schroeder, M. R. (1958). An Artificial Stereophonic Effect Obtained from a Single Audio Source. *Journal of the Audio Engineering Society, 6*(2), 74-79.

Schroeder, M. R., Gottlob, D., & Siebrasse, K. F. (1974). Comparative Study of European Concert Halls: Correlation of Subjective Preference with Geometric and Acoustic Parameters. *The Journal of the Acoustical Society of America, 56*(4), 1195-1201. doi:10.1121/1.1903408

Searle, C. L., Braida, L. D., Cuddy, D. R., & Davis, M. F. (1975). Binaural Pinna Disparity: Another Auditory Localization Cue. *The Journal of the Acoustical Society of America, 57*(2), 448-455. doi:10.1121/1.380442

Soulodre, G. A., Lavoie, M. C., & Norcross, S. G. (2002). Investigation of Listener Envelopment in Multichannel Surround Systems. Presented at the *113th Convention of the Audio Engineering Society*. Paper Number 5676.

Soulodre, G. A., Lavoie, M. C., & Norcross, S. G. (2003). Objective Measures of Listener Envelopment in Multichannel Surround Systems. *Journal of the Audio Engineering Society, 51*(9), 826-840.

Terrace, H. S., & Stevens, S. S. (1962). The Quantification of Tonal Volume. *The American Journal of Psychology, 75*(4), 596-604. doi:10.2307/1420282

Theile, G. (2001). Natural 5.1 Music Recording Based on Psychoacoustic Principals. Presented at the *Audio Engineering Society 19th International Conference: Surround Sound – Techniques, Technology, and Perception*. Paper Number 1904.

Theile, G., & Plenge, G. (1977). Localization of Lateral Phantom Sources. *Journal of the Audio Engineering Society, 25*(4), 196-200.

Thomas, G. J. (1952). Volume and Loudness of Noise. *The American Journal of Psychology, 65*(4), 588-593. doi:10.2307/1418039

Thurlow, W., & Parks, T. (1961). Precedence Suppression Effects for Two-Click Sources. *Perceptual and Motor Skills, 13*(1), 7-12. doi:10.2466/pms.1961.13.1.7

Toole, F. E., & Sayers, B. McA. (1965). Lateralization Judgements and the Nature of Binaural Acoustics Images. *The Journal of the Acoustical Society of America, 37*(2), 319-324. doi:10.1121/1.1909329

Trimble, O. C. (1934). Localization of Sound in the Anterior-Posterior and Vertical Dimensions of "Auditory" Space. *British Journal of Psychology, 24*(3), 320-334. doi:10.1111/j.2044-8295.1934.tb00706.x

Vilkamo, J., Lokki, T., & Pulkki, V. (2009). Directional Audio Coding: Virtual Microphone-Based Synthesis and Subjective Evaluation. *Journal of the Audio Engineering Society, 57*(9), 709-724.

Wakuda, A., Furuya, H., Fujimoto, K., & Isogai, K. (2003). Effects of Arrival Direction of Late Sound on Listener Envelopment. *Acoustical Science and Technology, 24*(4), 179-185. doi:10.1250/ast.24.179

Wallach, H., Newman, E. B., & Rosenzweig, M. R. (1949). A Precedence Effect in Sound Localization. *The American Journal of Psychology, 62*(3), 315-336. doi:10.2307/1418275

Wallis, R., & Lee, H. (2015a). Directional Bands Revisited. Presented at the *138th Convention of the Audio Engineering Society.* Conference Paper 9278.

Wallis, R., & Lee, H. (2015b). The Effect of Interchannel Time Difference on Localization in Vertical Stereophony. *Journal of the Audio Engineering Society, 63*(10), 767-776. doi:10.17743/jaes.2015.0069

Wallis, R., & Lee, H. (2016). Vertical Stereophonic Localization in the Presence of Interchannel Crosstalk: The Analysis of Frequency-Dependent Localization Thresholds. *Journal of the Audio Engineering Society, 64*(10), 762-770. doi:10.17743/jaes.2016.0039

Wightman, F. L., & Kistler, D. J. (1992). The Dominant Role of Low-Frequency Interaural Time Differences in Sound Localization. *The Journal of the Acoustical Society of America, 91*(3), 1648-1661. doi:10.1121/1.402445

Zotter, F., Frank, M., Marentakis, G., & Sontacchi, A. (2011). Phantom Source Widening with Deterministic Frequency Dependent Time Delays. Presented at the *14th International Conference on Digital Audio Effects (DAFx-11)*.

Zotter, F., & Frank, M. (2013). Efficient Phantom Source Widening. *Archives of Acoustics, 38*(1), 27-37. doi:10.2478/aoa-2013-0004