Future Spatial Audio: Subjective Evaluation of 3D Surround Systems

Paul James Power

Acoustics Research Centre

School of Computing, Science and Engineering

College of Science and Technology

University of Salford, Salford, UK

Submitted in Partial Fulfilment of The Requirements of The Degree of Doctor of Philosophy

Contents

List of Figures

List of Tables

Acknowledgements

Firstly i would like to thank my supervisors Bill Davies, Jos Hirst, Chris Dunn and Frank Melchior, also the BBC audio research partnership and the EPSRC for providing the funding and opportunity to embark on a PhD and my supervisors for their guidance and encouragement throughout the process. I would also like to thank my parents for providing support and encouragement and instilling belief in my capabilities and for their support throughout. Thanks also goes to all my colleagues and friends in the acoustic research centre especially Darius Satongar for his help with Matlab, Rob Oldfield and Henric Mattsson for lending equipment to make the experiments possible and Danny McCaul for providing access to test facilities. Thanks also to the post-docs and lecturers who gave up their time to make the descriptive analysis a success and also the subjective test participants. I would also like to thank Bruno Fazenda for helpful conversations on statistics and all audio aspects. Appreciation is also extended to Technicolor in Hanover Germany for the provision of ambisonic recordings and also the BBC R&D for the multichannel stimulus. Finally thanks to Aaron Heller and Eric Benjamin for use of their ambisonic decoder toolbox.

Abstract

Current surround systems are being developed to include height channels to provide the listener with a 3D listening experience. It is not well understood the impact the height channels will have on the listening experience and aspects associated with multichannel reproduction like localisation and envelopment or if there are any new subjective attributes concerned with 3D surround systems. Therefore in this research subjective factors like localisation and envelopment were investigated and then descriptive analysis was used. In terms of localisation it was found that for sources panned in the median plane localisation accuracy was not improved with higher order ambisonics. However for sources in the frontal plane higher order ambisonics improves localisation accuracy for elevated sound sources. It was also found that for a simulation of a number of 2D and 3D surround systems, using a decorrelated noise signal to simulate a diffuse soundfield, there was no improvement in envelopment with the addition of height. On the other hand height was found to improve the perception of envelopment with the use of 3D recorded sound scenes, although for an applause sample which had similar properties to that of the decorrelated noise sample there was no significant difference between 2D and 3D systems. Five attribute scales emerged from the descriptive analysis of which it was found that there were significant differences between 2D and 3D systems using the attribute scale *size* for both ambisonics and VBAP rendered systems. Also 3D higher order ambisonics significantly enhances the perception of *presence*. A final principal component analysis found that there were 2 factors which characterised the ambisonic rendered systems and 3 factors which characterised the VBAP rendered sound scenes. This suggests that the derived scales need to be used with a wider number of sound scenes in order to fully validate them.

1 Introduction

In this section a background to the research is given along with the areas to be investigated.

1.1 Perceptual Benefits of the Inclusion of Surround Channels

There has been a large amount of research into the perception of surround systems in the horizontal plane. The development of the stereo system to include rear channels creating the 5.1 system has been added for two purposes, the reproduction of lateral energy and rear sound effects (Rumsey, 2005). It is also stated by Stuart (1996) that the move from stereo to multichannel is significant, providing better sound source separation with lower masking thresholds. Further, it has been stated by Soulodre et al. (2003) that the addition of centre and surround channels allows sound to arrive from a number of locations, in turn allowing for a greater sense of realism. Therein the benefits of the addition of rear channels to the stereo system have been established.

There has been a move recently to extend horizontal surround systems to include height channels in a bid to develop surround systems and enhance the listening experience further. Recently layouts have been proposed which are additions to the 5.1 set up which are the Auro 3D layout which features four height channels and a total number of nine channels, whilst the NHK 22.2 system which has been developed by the Japanese broadcasting corporation which features nine height channels and a total of twenty four channels. There has not been a large amount of perceptual research into these systems to establish the perceptual benefits and these systems feature differing numbers of speaker channels, therefore in terms of wider consumer adoption it will be necessary to balance the number of additional channels with the perceived perceptual performance.

A statement from Davis (2003) also points out that the current 5.1 systems have been developed through years of intensive research and as a consequence provide a highly convincing reproduction. Further, he states:

"Any future system will have to provide a significantly better experience in an economical manner"

Proposing to include height channels in surround systems is not a new concept, since several decades ago Gerzon (1973) was proposing what was called 'Periphonic' systems which reproduced 3D surround using what was termed 1st order ambisonics.

The 1st order ambisonics provides limited resolution in the soundfield but has the ability to produce enveloping soundfields (Zacharov and Koivuniemi, 2001), (Guastavino et al., 2007), (Berge and Barrett, 2010) and (Cengarle, 2012). However now there is the availability of higher order ambisonics which extends the ambisonic reproduction beyond 1st order, this provides a greater resolution in the reproduced soundfield (Bertet et al., 2013). It was first proposed to extend 1st order to 2nd order by Bamford and Vanderkooy (1995) and can now be extended further to higher orders, the most widely used is 3rd order.

Even around the time when Gerzon was proposing to use ambisonics to reproduce surround with height, there were engineers who did not believe there were any perceptual advantages as the statement by Nisbett (1970) below communicates:

"I am not being totally facetious when i suggest that if God had meant us to take an interest in the vertical separation of sounds, we would have an ear on the top of our heads. Lacking such a rainwater collector, i don't see much need to feed directional information in this sense"

Therefore the research question arises: *Does the addition of height significantly enhance the listening experience?*

To answer the research question set out various aspects of the listening experience have been assessed using quantitative and qualitative methods. This required a number of different approaches utilising objective and subjective methods. From this an assessment of 3D surround systems from several different perspectives and an insight into how 3D surround systems impact on the listening experience have been gained.

1.1 Approach

In assessing reproduction systems it is necessary to identify aspects of which are important to the spatial audio listening experience, since asking listeners to provide a rating of overall sound quality does not tap into the underlying aspects of the reproduction, it is stated by Letowski (1989) that a global assessment is too general to answer any questions. Therefore the method used to assess surround systems with height broke the assessment of the reproduction into several different constituent parts. This allowed controlled listening tests and objective measurements to be performed in order to investigate the impact of the height channels on the listening experience.

Sound quality is considered to be composed of two broad categories by (Letowski, 1989) these are timbral and spatial. Further, Rumsey (2006) also states that spatial attributes account for over a third of all quality ratings in listening tests and therefore are important in determining the quality of a system. It has been stated by Rumsey (1998b) that it is important to establish the ability of a system to place a sound source with accuracy, referring to localisation. In addition, Berg and Rumsey (2000) describe that spatial impression is also an important attribute and more recently George et al. (2010) also states that the sub attribute of 'spatial impression' envelopment is one of the main attributes driving multichannel development.

Detailed in *Figure 1.1* it can be seen that the spatial audio quality broadly consists of several different components based on the statements of the previous authors, attributes of which include localisation, apparent source width and envelopment.

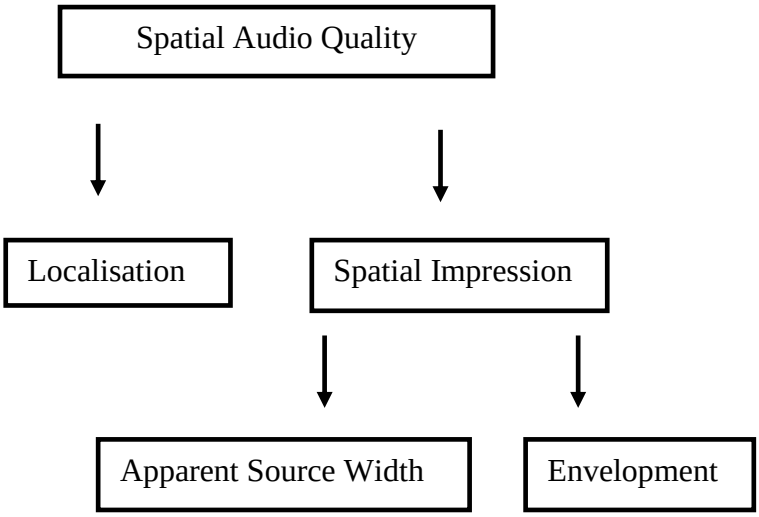


Figure 1.1 Hierarchy of Spatial Audio Quality

The previous description has helped define the direction of assessment of with-height systems. Since the aforementioned attributes are important in spatial audio reproduction. Although these attributes, especially envelopment, have been concerned mostly with surround in the horizontal therefore it is uncertain how the height channels impact on these attributes and whether the addition of height could generate new percepts.

1.2 Research Objectives

Overall the objective of the research was to investigate the impact of the addition of height to surround systems. This has included the factors outlined earlier in terms of localisation of virtual sources in with height systems, an investigation into the number and positioning of

height channels to enhance envelopment and finally the use of a more exploratory technique to discover if there are any unique attributes concerned with 3D surround systems. More specifically these objectives were:

1. To investigate if higher order ambisonics can increase the localisation accuracy of virtual sources in the median plane and in the frontal plane.

2. To investigate if there are any perceptual advantages in using higher order ambisonics

3. To investigate the effect of the positioning and number of height channels on the perception of envelopment.

4. To discover if there are unique attributes connected to the perception of 3D surround systems.

1.3 Organisation of the Thesis

In Chapter 2 a review is included which examines previous work carried out in several different areas concerning surround sound with height. This starts with an historical account of the development of reproduction systems to the present state of the art. The rendering methods available for creation of surround with height are also explored and the perceptual benefits of each method. The review then moves onto localisation in human hearing with a particular focus on the localisation of sources in the vertical plane and the cues available in terms of the direction of arrival and in addition the localisation accuracy in terms of the frequency content of the sound source. The review then moves onto spatial attributes which have been identified as being important in the reproduction of surround sound in the horizontal plane and then moving to the research which has been carried out into the perception of 3D surround systems. Particular focus has been on highlighting the fact that the perceptual qualities of spatial audio is multidimensional and that research methods used have asked more general questions of listeners therefore not revealing in which way 3D surround systems are impacting on the listening experience.

Chapter 3 documents the development of surround systems starting with mono and stereo reproduction systems and then moving to surround systems in the horizontal plane and then to the current 3D systems. Rendering methods are also detailed which allow the creation of

3D reproduction over loudspeakers and the benefits of each method. Details are then given of the rendering method used in this thesis to create 3D soundfields, after which an account of the human hearing mechanisms used to perceive these soundfields are detailed with a particular focus on elevated sources. Finally an account is given of the perceptual benefits of surround sound systems and the possible benefits of extending surround reproduction with the use of height channels.

The first area of investigation involved assessing localisation in 3D systems using ambisonics as the rendering method. This broke the localisation experiments into two tests. In Chapter 4 this allows sources to be evaluated in the median plane using one stimulus source and varying the virtual source position. Three different ambisonic orders are evaluated and the localisation error for each order is detailed and a discussion is presented and the validity and limitations of the results are considered.

In Chapter 5 a second localisation test is carried out using two stimulus sources. Particular focus is given to sources elevated in the frontal plane. Results are presented which compares the perceived localisation of pink noise versus a female speech sample. Further, virtual sources are compared to real speaker positions. Results are presented which detail the use of higher order ambisonics in terms of localisation accuracy for virtually elevated sources and the limitations of the experiment are considered.

In Chapter 6 a number of different with height speaker layouts are investigated focussing on the attribute envelopment. This involved an objective and subjective evaluation of a number of different 2D and 3D speaker layouts. A computer simulation of different 2D and 3D layouts is carried out and the resultant output of each system is binaurally measured using IACC. The following stage involved a controlled subjective test with the same layouts and further measurements of the real soundfields again using the IACC. The results of the objective and subjective evaluation are presented and discussed, the limitations and validity of the investigation are also considered.

In Chapter 7 a height system is chosen based on the objective and subjective results from Chapter 6, a further subjective test is executed again focussing on envelopment but in this case utilising three different types of recorded material. The results of the objective and subjective test are presented and the impact of the height channels on the perception of

envelopment are discussed along with effects of different sound scenes on the results. Finally the limitations of the experiment are also considered.

In Chapter 8 a descriptive analysis is initiated by carrying out an elicitation utilising 2D and 3D surround systems. A large number of descriptors are collected based on spatial and timbral attributes. Three separate focus group meetings reduced 384 descriptors into five scales based on spatial attributes of the reproduced sound. The final scales could then be used in a subjective evaluation.

Chapter 9 details a subjective test utilising the unique attribute scales derived from the previous elicitation in Chapter 8. 2D and 3D systems utilised in the elicitation stage were used in the subjective test and a number of additional systems were also included and also additional sound scenes. Objective measures were also carried out on the different systems. The subjective and objective results are presented along with further analysis of the subjective scale usage using PCA (principal components analysis) are also presented.

In Chapter 10 a summary and conclusion is detailed with reference to each of the chapters included in the thesis. The avenues for further work are also detailed.

1.4 Contributions to the Field

Overall the research carried out in this thesis is novel, in the fact that it sets out to contribute to the understanding of the addition of height channels to the surround listening experience through several separate subjective evaluations.

Contributions have been made to the improvement of localisation of sources panned in a 3D surround system by using higher order ambisonics (HOA).

Contributions have also been made to develop an understanding of the addition of height channels to the perception of envelopment, where it has been shown that 3D surround systems can be used to enhance the perception of envelopment.

Experimentation has shown that for certain sound scene compositions the addition of height does not significantly enhance the listening experience over current reproduction formats for example stereo and 5.1.

This work has also illustrated some of the differences between 2D and 3D surround using unique attribute scales derived from a descriptive analysis. From this descriptive analysis it was found that the use of HOA significantly enhances the perception of presence and realism

using 3D surround. Finally utilising 3D surround with HOA and another rendering method called vector based amplitude panning using different sound scenes showed comparable performance, in that the addition of height in both cases enhances the ability to perceive the size of the reproduced sound scene.

2 Literature Review

In this section a review of previous research carried out is given, particular focus is paid to surround systems with height which includes methods of rendering 3D sound scenes over 3D speaker arrays. The review then moves onto the research carried out into the ability of the human hearing system to localise elevated sound sources and the advantages and disadvantages of sources originating in the vertical plane. An overview is given to the methods previously used to quantify elevated sound sources and the advantages and disadvantages of each method. Finally an account of important attributes in surround reproduction is reviewed and consideration is given as to the implication of these attributes in with height surround systems.

The review starts with an historical account of the development of reproduction systems from mono to multichannel.

2.1 History of Surround Systems

The development of multichannel systems was started by Blumlein in 1931 who developed no less than 128 patents, which included the stereo format allowing the recording, transmission and reproduction of sound using two channels of audio, however this system was not put to practical use until the 1950's (Torick, 1998). Around the same time Harvey Fletcher of Bell labs was experimenting with a wall of microphones which picked up the source at one side and reproduced the audio on the reverse side with a wall of speakers, which was later reduced to a three channel system (Streicher and Everest, 1998).

The previously mentioned stereo system utilised only two channels of audio in front of the listener. Although a conductor named Leopold Stokowski was working with Disney and RCA on a commercial theatre system called Fantasound in 1940. This provided three frontal channels and additional speakers placed around the audience for extra effects (Davis, 2003).

The availability of surround systems in the home did not come until 1968 when the quadraphonic system was proposed by Peter Scheiber. Although this system never took off for a number of reasons which included, perceptual factors and the requirement of consumers to buy additional equipment (Torick, 1998).

The beginnings of the current 5.1 system in use today came in 1975 with the creation of Dolby stereo, this system included three frontal channels, a stereo configuration with an additional centre channel at the front and a single rear channel. Around the same time another researcher Michael Gerzon was proposing ambisonics (Gerzon, 1973), which was a complete recording, transmission and reproduction system with the ability to reproduce 3D surround, however this was around the time when quadraphonics was also vying for commercial success so ambisonics never realised commercial adoption. Another researcher called Mantel was also proposing surround systems with height to provide better reproduction of the recording space, with a paper entitled 'optimum multiphony' (Mantel, 1975), although similar to Gerzon this never attained widespread adoption. Yet, the Dolby stereo system won through and became a commercial success and was the subject of much development which then led to a standardised system in 1991 known as 5.1 defined in ITU-R Rec. B.S775-1. , which included five discrete channels, three frontal channels two rear surrounds and a low frequency channel.

2.2 Surround With Height

The circumstances in which the previous 3D surround systems had been introduced and ultimately their failure may have been because of the disillusionment caused by the quadraphonic system. Although this was not the end of the foray into 3D surround systems because in 1999 a 10.2 system with two front height channels and a ceiling speaker, dubbed 'the voice of god channel' was proposed by Thomlinson Holman (Willis, 1999). This then led to another system proposed by the Japanese broadcaster NHK in 2006 which features height; the 22.2 system which has been in development for several years now featuring nine height channels (Hamasaki et al., 2006). In addition the I.T.U (International Telecommunications Union) have produced a document detailing numerous spatial audio systems which include a 9.1 system which consists of the original I.T.U 5.1 system with the addition of four height channels and which also details the 22.2 system (ITU, 2011).

A system which utilises two front height channels similar to Thomlinson's system that did make it into commercial success and people's homes is the Dolby Pro Logic IIz system (Tsingos et al., 2011), which utilises the common 5.1 system but includes two front height channels measured at 45° inclination above the two stereo channels.

It can be seen that there have been ideas based around adding height, however perceptual research has been sparse regarding these systems, recent work by Silzle et al. (2011) where

29 industry professionals were interviewed about applications for 3D surround systems indicated that news broadcast on radio or TV as expected would not require 3D sound, but cinema, home DVD, and TV, would benefit from 3D surround. With a view to developing surround systems with height whichever way is proposed to do this it will need to be backward compatible since it will also have to cater for current legacy systems like stereo, 5.1, and future developments and also the current commercial studios be it post production, music, or indeed broadcast facilities.

There are a choice of spatial rendering methods however not all of them provide the facility to create surround with height. An overview of these methods will be given.

2.3 Spatial Rendering Methods in 3D Audio

Within the bounds of surround reproduction there are a number of techniques to facilitate the creation of spatial sound-fields. The first and foremost is the amplitude panning method used in stereo and surround systems today, invented in 1931 by Alan Blumlein. This method is designed to allow sources to be placed at different angular offsets to allow instruments to be placed in different positions to build up a sound scene utilising the sine/cosine panning law (Tomlinson, 2000), or in more advanced cases in surround systems to place sources around a listener in the horizontal plane.

A more sophisticated method of surround reproduction was developed at Delft University by Berkhout, de Vries and Vogel (Berkhout, 1992) this system worked on the theory that any point of a wave-front can be considered as a secondary source. The basis of wavefield synthesis (WFS) is in the mathematical principles stated by the Huygens theory. WFS systems require a large number of planar arrays of speakers which are used to reproduce virtual sources. The benefits of WFS are in the fact that the reproduction of the source is not only valid at one point in space but at any point within the whole area, which is only delimited by the speaker array, drawbacks of WFS are in the fact that spatial aliasing occurs above a certain frequency caused by the physical distance between the transducers, thus correct frequency reproduction can only be guaranteed up to a frequency limit (Daniel et al., 2003), Traditionally WFS is confined to the horizontal only reproduction, although recently there has been a small amount of research into the inclusion of height in WFS systems (Rohr et al., 2013).

Another more recent addition to the family of spatial rendering techniques is the VBAP (Vector Based Amplitude Panning) this allows sources to be panned utilising horizontal

surround systems, using two speakers at any one time or arrangements of three speakers in triangles for 3D systems. This system was conceptualised by Pulkki (2001) and was a tool used for spatial audio panning and also to study human hearing, within his paper he detailed the fact that with elevated sources in the median plane, not one panning law could be used to characterise human hearing. VBAP is not a complete system since it does not allow the capture of acoustical events, only the panning of recorded or synthesized sounds.

Ambisonics is a technology that was proposed back in 1973 which would allow the reproduction of surround systems with height. A paper on the topic of 3D systems which used the term 'periphonics' to refer to systems with height was presented at the AES (Audio Engineering Society) entitled 'Periphony with height sound reproduction' by Gerzon (1973). This paper encompassed a whole theory from recording to reproduction and also grounded the technology in human hearing mechanisms by taking into account the optimisation of frequencies relating to the ITD (inter-aural time difference) and ILD (inter-aural level difference). However, at the time it was not received with much enthusiasm due to failings of the quadraphonic system. Although now with the development of spatial audio systems to include height, ambisonics could provide a means for creation and reproduction of 3D sound.

Some of the current surround systems with height feature large numbers of channels which would be a limiting factor when proposing systems for wider adoption in consumers' homes for example or in terms of required bandwidth for broadcast. Although it may be that the number of channels in the vertical could be reduced compared with the horizontal and still provide an enhanced listening experience.

Psychoacoustic research has been carried out into the capabilities of human hearing with sources in the vertical plane. The next section will illuminate some of the shortcomings of human hearing with sources in the vertical plane.

2.4 Perceptual Cues

Since this work considers surround systems which can deliver sounds from the vertical dimension in addition to the horizontal, it is necessary to consider how this information might be processed by the human hearing system. For sound sources in the horizontal dimension the duplex theory is well known (Rayleigh, 1875), This is where the human hearing uses two different mechanisms to localise a sound, one being inter aural time differences for low frequencies <700Hz and inter aural level differences for high frequencies >1200Hz. Knowledge of the duplex theory can be used to manipulate sound sources so as to appear in a

specified direction using amplitude or time differences which are found in common stereo reproduction. However sources placed above a listener or at a specified angular height will utilise spectral filtering caused by a combination of head, pinna, and shoulders. These cues in addition to the duplex theory are said to provide the auditory system with the necessary information to interpret where the sound is coming from. Other factors involved include the spectral content of the actual sound source which also helps in determining the location of elevated sound sources, but can also fool the auditory system causing inaccurate localisation.

Early studies by Roffler and Butler (1968) into the spatial location of different high and low tones using a bank of speakers displaced vertically with respect to a seated listener, found that sound sources were judged in terms of their respective pitch. This was demonstrated by asking the participants to call out which loudspeaker was producing the tone using numbered speakers which were hidden from view. They found that the sources were judged as coming from a higher position than that of the speaker which was reproducing the sound and that the sounds respective position was based on its pitch rather than speaker orientation. It was also concluded that in order for a sound to be correctly placed by a listener in the vertical, the sound must be complex, and include frequencies above 7 kHz, in order that the wavelengths of the sound source are small enough to interact with the human pinna. This is also reinforced by Pulkki (2001) in that in his tests it was found that high passed noise was localised significantly easier than low passed. Roffler and Butler also found that participants with occluded pinna were not able to localise noise containing only frequencies above 8kHz, contrary to this they were able to localize the sound when it was placed in the horizontal plane with occluded pinna, probably due to the available ITD (inter aural time differences) and ILD (inter aural level differences) cues in the horizontal.

Further studies were carried out by Hebrank and Wright (1974b) into the localisation of elevated sound sources in the median plane. The author’s manipulated white noise using several different notch filter types, in order to observe where their participants localised the sound in terms of elevation to predict the filtering effects of the external ear. They found that median plane localisation of white noise is based on spectral cues due to the filtering of the external ear, which concurs with the previous author, and that these cues were primarily based between 4-16kHz. They also found that there were three other general directions represented by other spectral cues which were for the front, a one octave notch with a cut off frequency of between 4-8kHz, above which had a quarter octave peak between 7-9kHz, and

behind which had a small peak between 10-12kHz with a decrease in energy above and below the peak. This was similar to the findings of Blauert (1996) in which he determined that the location of narrowband noise in the median plane was governed by the frequency of the sound source not the actual position. This phenomenon was also demonstrated in an experiment by Bloom (1977) who asked subjects to move the centre frequency of filters until a wideband source matched that of the fixed source. He found that the position subjects matched the virtual source to the real source corresponded with that of the notch frequencies of the external pinna. Hebrank and Wright summarise their findings by reporting that the relationship between the notch frequency and elevation is responsible for what they term as 'auditory highness'.

As previously illustrated sources in the median plane will provide equal cues to each ear, this leads to the question of whether localising a source in the median plane is a binaural or monaural process. Two experiments were carried out by Hebrank and Wright (1974a) in which the first experiment the participants had to localise a flat spectrum white noise which was chosen to be a familiar sound, white noise was also processed using delays and filtering to create an unfamiliar sound. This process encompassed the first experiment in which listeners localised the sources binaurally in the median plane. A second test was then carried out using the same sound sources which listeners then had to localise the sources in the median plane with one ear occluded. However in the second test listeners were provided with feedback once they had given their answer so that they could learn monaurally to localise the sources. The authors found that the confusion created by the processed white noise was a feature of the monaural and binaural condition, demonstrating that certainly in their situation two ears are not used for median plane localisation. Further to this the training provided in their second test showed that participants can localise as well monaurally in the median plane.

Outlined in the introduction localising in the horizontal plane is primarily a feature of the duplex theory, however what is the role when sounds are displaced vertically and off the median plane? This is studied by Butler and Humanski (1992) by utilising low and high passed noise bursts at 3kHz, they hypothesized that low passed filtered noise would not be localised accurately in the lateral vertical plane, diminishing the theory that the binaural difference cues are used with elevated sound sources. Several conditions were set up where low and high pass noise was utilised in the median and lateral vertical plane. It was observed

from the results that binaural localisation, comparing high passed noise with low passed noise in the median and lateral vertical plane that the proficiency for localisation of the high passed noise surpassed that of low passed noise, this agrees with the previous research of Hebrank and Wright. It was also found that monaural localisation in the lateral vertical plane was improved when high passed noise was used over low passed, this concurs with the theory that the filtering effects of the pinna helps localise elevated sound sources. Further, it is stated that when listening monaurally there is a need for high frequencies to locate a sound in the lateral vertical plane, however with the use of binaural cues listeners can localise low passed noise in the lateral vertical plane. Although at elevations greater than 45° this breaks down, contrary to this high passed noise can be located accurately at elevations up to 60°.

It has been explained previously that the duplex theory allows localisation of sources in the horizontal plane, and that other cues like HRTF's (head related transfer functions) allow localisation of sources in the vertical, it has also been proposed by Wallach (1939) that dynamic head movements also facilitate the localisation of sources in the vertical dimension, due to head movements which produce variations in inter-aural differences to resolve the position of sources in the vertical, this movement of the head not only on the lateral axis right and left, but also from tilting the head from side to side.

This was found by Thurlow and Runge (1967) who investigated the effect of head movement in assisting a listener to locate a sound source placed in the vertical and horizontal dimension. Different types of sound source were used which included transient high and low frequency tones and broadband high and low frequency tones. The authors found that head movement plays a critical role when there are front back localisation ambiguities. It was also found that head movements for sources in the horizontal reduces the localisation error, and that rotation of the head for vertical sources does not significantly reduce the vertical error, however, that a combination of pivoting and head rotation does reduce this error. To discern between the click stimuli and broadband noise in the vertical it was found that the click stimulus was harder to localise than the broadband stimulus. Further, the authors found that the departure from the horizontal for low frequency stimulus was harder to localise, yet the high frequency broadband noise was easy to locate in the vertical without head movement. The previous factors pointing to the combination of high frequencies and the pinna for accurate localisation of elevated sources which has been stated by several other researchers including Begault (1991) who states that the types of sound source input into a 3D system will affect

performance, with broadband and impulsive sources providing best performance, contrary, that low frequency sound sources and sources with slow amplitude envelopes being worse.

Recent research by Ashby et al. (2011) investigated the influence of head rotation on the localisation of elevated sound sources. They used a multi microphone sphere consisting of four pairs of omni-directional microphones distributed around the equator of a plastic sphere. The sphere was placed on the top of a simulated torso and mounted on a rotating table. A single speaker was used moving in increments of 5° from -20° to +70° and the turntable rotated in increments of 5° so that the impulse response of each microphone could be captured. Analysing each pair of microphone responses separately it was found that each had its own cone of confusion, however that there was some indication of the source position possibly from torso reflections. Yet when the responses from all the microphones were combined the source could be positioned accurately confirming the work of Wallach (1939) that head rotation does help in localising elevated virtual sound sources. Further, narrow and broadband noise sources were also tested spanning the frequency ranges that coincide with the duplex theory. For broadband sources these were localised well, however for noise sources coinciding with inter aural time differences these were degraded but could still be positioned with some accuracy. This coincides with the work of Butler and Humanski (1992) who found that elevated sounds containing high frequencies were localised with more precision than those containing solely low frequencies.

Following up the objective results Ashby et al. (2013) utilised subjective testing to investigate head movement in assisting three dimensional localisation. They used a 3D speaker array consisting of 17 speakers distributed from 0-180° in azimuth and from -35° to 55° elevation. Three noise signals were used broadband white noise, 2kHz low pass filtered noise and 2kHz low pass filtered noise with an additional 5.7-8kHz band-pass filtered noise component. Three different conditions were established, free head movement, no head movement and forced head movement. Listeners reported their judgement using a laser pointer and head tracker. It was found that the broadband noise source was localised accurately the most and the other filtered versions degraded elevation localisation accuracy. It was also found that for all types of sound source that head movement significantly reduces the localisation error, in addition it was noted that pinna cues contribute more to elevation localisation accuracy than head movement.

It would seem then that a combination of a number cues are used by the auditory system to locate the direction of a source in the vertical. In addition Davis and Stephens (1974) study the effects of stimulus intensity in the median plane. Two different stimuli were used, one being white noise and the other speech, participants were also observed and no head movements were made. Repeats of the stimuli were used. It was found that the noise stimuli could be localized significantly more accurately than the speech stimuli. It was also found that an increase in amplitude also lowered the localisation error, and that subsequent repeats of stimuli did not reveal any learning effects for vertical localisation. It was also interesting to note that the authors did not report any elevation offsets of sound sources, which have been reported by others concerning sounds containing high frequencies.

2.5 Summary

From the previous review of human hearing with elevated sources it would seem that there are a number of different ways in which humans can resolve the position of elevated sources. This can be achieved with methods which include tilting and turning of the head, it has also been identified that the spectral content of the sound source also plays a significant part, with sound sources including high frequencies being easier to locate than sources containing low frequencies when placed at an elevation. These factors would certainly have an impact when considering reproduction of sound sources over surround systems with height. In addition the auditory system can be fooled by manipulating the high frequencies (which mimics the effects of the pinnae on high frequencies) sources can be made to appear at different elevations independent of the actual source producing the sound, this would have an impact on the panning of sources in the vertical, since the intended positioning of a source may not match its perceived position.

Finally, because of the limitations of human hearing in the vertical it may be possible to reduce the number of channels and still provide a significantly enhanced listening experience.

In order to investigate with some accuracy the localisation of elevated sources it is necessary to record judgements from participants, however this is not without its complication and is not without error. A number of researchers have investigated how this could be achieved and also the accuracy gained

2.6 Recording Localization Judgements

Setting out to find the spatial precision of a sound reproduction system has complications. Evaluation of 2D systems itself is problematic however 3D brings further problems and a number of methods have been used to record localisation judgements from participants.

A method used by Wightman and Kistler (1989) to investigate the spatial position in elevation and azimuth of noise bursts presented over speakers or headphones consisted of allowing the participant to call out using spherical coordinates the apparent position of the virtual source, for example a source directly ahead with no elevation would produce a response of 0,0 or a sound heard to the right of the participant with a slight elevation would produce the response 90,0. The authors state that they were concerned that this would produce unreliable results, however they found that participants quickly became accustomed to the method and gave reliable results.

A more in-depth review of different directional reporting methods is given by Evans (1998) this review details several methods used by others to overcome difficulties in giving directional estimates of a sound source. It is stated that allowing subjects to answer using angular coordinates can give a high degree of resolution. However, that sometimes this can be counterintuitive to a listener. Instead it may be better to use in terms of a 2D system the clock-face method, where the 360 degrees around the listener is divided into 12 positions, although this is more intuitive it only has 30 degrees of resolution. Another method can be used similar to the clock face method for reporting of a 3D system, where, each point around a listener can be divided into for example in the front: forwards above, forwards below and so on, after the participant has given their vote a marker sound is then played, and the subject reports the direction of the marker sound relative to the virtual sound. The drawback with this method is that the time taken to complete the test. Possibly a better way of directional reporting is the use of a custom GUI (Graphical User Interface), this method of reporting can speed up the test, however this involves the subject mapping 3D space to a 2D GUI, investigations into reporting methods for localization tests using specialized GUI's has been studied by Pernaux et al. (2003) and it seems that this comes with problems too, In their investigations three different reporting methods have been tested to find which method gives the least error. Using individual data sets of HRTF for each participant the authors utilized three different GUI reporting methods. The reporting methods the first of which consisted of a 2D visual representation and a mouse to report the judgement. With the other two GUI's consisting of a 3D visual feedback system, however one relied on a mouse input whilst the

other relied on finger pointing. The 3D visual feedback systems relied on a laser scanning of the participants bust to derive the anatomical shape of the participant. Front back confusions were found for their data, and also down/up confusions. Comparing their different methods, the authors found that the 3D visual feedback system provides less error than the 2D visual feedback system with a reduced angular error of three degrees. Although the finger pointing method with 3D feedback produces the least error, and also provides the fastest response time compared with the other methods.

Similar to the previous author an investigation into user interfaces for localisation tests has been carried out by Schoeffler et al. (2014). Two different interfaces were investigated one used a 2D view whilst the other allowed a 3D view using a movable camera. To evaluate the two methods a six channel speaker array was used. Three signals were utilised pink noise, castanets and a sine wave of 220Hz. The speakers were screened from the listeners view. Using each of the GUIs (graphical user interface) in turn, 30 listeners reported their localisation judgement with them. The 3D GUI was found to be slightly more accurate than the 2D one, but the 3D method was not statistically significantly more accurate than the 2D version. Further, it was explained that listeners were faster at reporting with the 2D GUI than the 3D, reasons for this was that the 2D GUI had two views at the same time and for the 3D system to have the same perspective the camera had to be moved increasing the response time.

2.7 Summary

It would seem from the previous investigation that to a certain extent the use of sophisticated reporting methods in localisation test do reduce the reporting error. However, it would seem that there is still no perfect method for reporting the position of a sound without some inaccuracies and also hindrance to the user reporting their judgement.

The positioning accuracy of a number of rendering methods, mainly ambisonics, have been investigated using 3D surround systems, and as will be described in the next section a number of different techniques have been adopted in order to record localisation judgements.

2.8 Localisation in Reproduced Sound

It was illustrated at the beginning the peculiarity of human hearing in the vertical dimension, particularly in the median plane. As a consequence this will certainly have an impact when considering positioning sources using speaker arrays with elevated channels. Limited research in this area has been carried out.

A study of the localisation of sources in the median and horizontal plane was carried out by Pieleanu (2004) varying the elevation of pink noise in 5° increments, utilising a dodecahedron configuration of speakers. Reverberation was also used to simulate a more reverberant environment. Rendering of the sources was handled using 1st and 2nd order ambisonics. In the initial evaluation of the results they found that there was no trend to each of the listener's responses. For sources on the horizontal plane there were large deviations for 1st order for both anechoic stimuli and reverberant one and no significant improvement with 2nd order. Taking the total average over all subjects for the anechoic environment 1st order was found to be slightly better than 2nd order. Evaluation of sources in the median plane found similar results where no significant difference was found between orders, no matter where the virtual sources were positioned using 1st order ambisonics, the virtual sources grouped around 10-11° in elevation and slightly higher for 2nd order, it could be that no matter the ambisonic order used there would still be no improvement because of the factors outlined by Hebrank and Wright (1974b) in that localisation of sources in the median plane are dominated by spectral cues alone. Ioana concludes that for horizontal errors training would provide better results, although for elevated sources in the median plane this would not be the case, possibly due to factors found by Pulkki (2001) in that not one panning law could explain the localisation of virtual sources in the median plane.

An investigation of the NHK 22.2 system was carried out by Liebetrau et al. (2007). The authors tested two systems. The first system tested was the 3D NHK 22.2 system, the second being WFS (wavefield synthesis system) with speakers only located in the horizontal. Since this study is mostly concerned with 3D systems, focus was directed on the testing of the 22.2 system. Three listeners took part simultaneously, with one listening position being in the sweet spot, and two other positions diagonally at either side. In order to help subjects record their responses a coordinate system was projected on the walls and the reproduction systems were screened from listeners view. Three different test signals were utilised, pink noise, guitar, and female voice. The authors state that the voice and guitar were recorded in a non-neutral environment, therefore containing some spatial information. Participants were asked to note the sound's position, and also how defined it was. The authors do not describe the panning method used, instead state that it is a method under development. It was found that if the source is between speakers and there is a large angular distance between each of the speakers, the source is pulled toward the nearest speaker. This was more pronounced for the 22.2 system, the authors stating that this was because of the varying density of the speaker

array. There was also a tendency to overestimate the angle of elevated sources, however this could have also been caused by the type of sound source used, certainly in the case of pink noise, and could be linked to factors outlined by Roffler and Butler (1968) and Blauert (1996) in that the sound source assumes an elevated position based on its frequency content not its actual elevation. Unfortunately there was no breakdown regarding the performance of the systems using the different sound sources.

Capra et al. (2007) Investigated 1st order ambisonics, and a triple stereo dipole configuration using trans-aural techniques. A six speaker configuration was used for the triple stereo dipole, with 16 speakers used for the 1st order ambisonic system arranged in a horizontal octagon, with four speakers arranged above and below the listener in a cube. The triple dipole system had a set of speaker's front and rear, and a set positioned above the listener. In order to test the localisation of the two systems pink noise was used as the stimulus. The subjects were asked to simply write down the direction in elevation and degrees the position of the source using a small plastic sphere, which had elevation and azimuth marked on it. Twenty listeners took part in the test using 21 different source locations of which 8 were real speaker locations. Horizontal and elevated sources were used, however of more interest in this study are the elevated sources. No clear statistical significance was found between systems, although the triple stereo dipole system had slightly better localisation by 3% than the ambisonic system. It was also noted that localisation of sources not on the horizontal were much harder to achieve for both systems. It was found that five of the real speaker positions were localised correctly by 80% of the participants and by 100% of participants within 30°. The imprecise localisation of the ambisonic system could have been due to the use of a 1st order system, yet the dipole system was worse for sources in the lower sector, this would be because the dipole system had no speakers located below the listener. These are similar results to that of (Pieleanu, 2004) and in fact were not improved with the use of 2nd order material, however it may be that the increase to 3rd order may provide more precise results.

Morrell and Reiss (2009) expanded their comparisons to other methods such as VBAP (Vector Based Amplitude Panning), ITD inter-aural time delay, and 3rd order ambisonics to find which method using manipulation of distance, and position, places the sound source at the desired location with least error around a 3D 16 channel speaker array. The ambisonics panning method used all 16 channels at once to pan a source whilst the VBAP and ITD method used only 6 at once. To pan sources using the ITD method delays were calculated for

each speaker, for example to pan a source in a 2D array to the left speaker, a delay was inserted into the right speaker, for the 3D implementation a triplet of speakers is used and delays and gains are calculated for all 3 speakers to pan the virtual source. A variety of audio samples were used, these included jazz recordings and pop music, which were used over noise bursts, the authors stating that these were more representative of the type of sounds people would listen to. Further, it is not stated whether the speaker array was screened from the listeners view and the sound samples were long in duration >20seconds. Only four positions were tested, and it was found that overall the 3rd order ambisonic method and VBAP method gave similar results and were more reliable than the ITD method.

The localisation of two ambisonic orders was evaluated by Braun (2011) utilising a Soundfield microphone which covers 1st order, and an Eigenmike which recreates 4th order ambisonics. Instead of recording arbitrary sources impulse responses were used, which were measured in a non-anechoic environment, however windowing of the impulse responses was used to remove reflections. A number of directions encompassing a variety of elevations and azimuths were measured. The impulse responses were then convolved with pink noise which created the test material. Further the author then utilised virtually encoded sources of 1st and 4th order to compare with the pink noise sources. To reproduce the audio, a non-regular array of speakers covering the hemisphere was used consisting of 24 loudspeakers and the decoder was optimised for this configuration. To record participants judgements, a toy gun was used which had laser tracking, further 5 real loudspeakers were utilised at select positions. Their results showed that as expected the 4th order provided better positioning of sources, with the 1st order material showing a high number of errors and the 1st order synthetic sources showing fewer errors. Yet the 4th order synthetic sources and Eigenmike recordings were comparable. Overestimation of elevated sources was found for 1st order soundfield, possibly due to previously outlined factors concerning human hearing in the vertical. And because lower order ambisonics has a larger active loudspeaker number compared to higher order ambisonics. However, the author states that it was due to the absence of the bottom layer of speakers which causes an imbalance of energy in the speaker array causing sources to be pulled up.

It could be said that the previous evaluation utilising pink noise may not have a high degree of ecological validity, however a study into several aspects of Ambisonic decoding, which included localisation was carried out by Horsburgh et al. (2012) in this study three different

decoders are evaluated utilising horizontal 1st, 3rd, and 7th order decoding with 1st order utilised for height. Further to this two different environments were utilised, this was a room with similar properties to that of a domestic environment, and a semi anechoic chamber. The participants were asked to consider sound source direction and relative perceptual quality, this allowed the simultaneous comparison of stimuli similar to a MUSHRA (ITU, 2001) (Multiple Stimulus with Hidden Anchor and Reference) test. It was not stated if the speaker array was hidden from the listeners view. In each case using the 1st, 3rd and 7th order decoder, two female samples and a male speech sample were distributed within the horizontal plane, in addition a musical excerpt was also used with the instruments being distributed around the horizontal in a similar manner to the speech samples, tones of 400Hz, 1.5kHz and 7kHz were also used and distributed around the horizontal plane also, no exact intended directions for the sources were given. The 1st order cube included was used to add reverberation. In terms of the distributed speech samples it was found that the increase in decoding order from 3rd to 7th order did not provide a significant difference for immersion. It was also found that in the domestic listening room environment there were larger perceived deviations for the female speech sample than the male sample, with an average deviation +-20°. However large deviations were shown for the pure tone stimuli and it was stated by participants that this was the hardest to localise and it was found that in the domestic environment there was a smaller localisation error compared with the semi anechoic environment. For the music sample the highest preference was for the 3rd order decode with and without height, with the 7th order decode being least preferred. Overall it was found that the 3rd order decode without height offers the highest degree of immersion, however that for localisation of speech and music the increase to 7th order provided more accurate results and that the introduction of height did not improve the preference scores significantly.

2.9 Summary

The previous review of localisation tests demonstrates that in the case of using ambisonics as the rendering method, using 1st order does not provide accurate localisation, although it may be that increasing the ambisonic order may provide more accurate localisation, it would seem that this would need to be greater than 2nd order. It was reported by Pieleanu (2004) that there was no significant difference in localisation accuracy when increasing from 1st to 2nd order, therefore it may be necessary to extend the ambisonic order beyond 2nd order, some evidence of this was found by Braun (2011) that utilising a 4th order decode provided significantly better localisation accuracy, in addition it was also found by Horsburgh et al.

(2012) that an increase from 1st to 3rd order provided significantly better localisation accuracy and that the further increase to 7th order provided further increased accuracy, but that the increase from 3rd to 7th order did not give the best sense of immersion. Further it was found by Morrell and Reiss (2009) that the use of a number of different rendering methods which included VBAP, ITD panning and also 3rd order ambisonics, that 3rd order ambisonics and VBAP provided the greatest accuracy, therefore it may be that 3rd order ambisonics would be sufficient to provide significant localisation accuracy over 1st order.

The previous review of localisation tests illustrate that it may be necessary to consider a number of factors to provide accurate localisation. Although good localisation is not the only factor which makes for a highly rated listening experience, for example looking towards concert hall acoustics there are five main important subjective factors these are: clarity, reverberance, spatial impression, intimacy and loudness (Barron, 2009). It has been noted that there has been much research into the abilities of horizontal surround systems in terms of spatial impression (Morimoto, 1997) , particularly 'envelopment' (Soulodre et al., 2002), (Soulodre et al., 2003), (George et al., 2010).

The next section will explore the research carried out into the perception of with height systems in terms of the number and positioning of the height channels focussing on the attribute envelopment.

2.10 Number of Channels and Positioning in 3D Surround Systems and its Impact on Envelopment

Asking participants about the overall sound quality of a system provides limited information, therefore it may be more useful to focus on an important attribute. One such study was conducted by Oode et al. (2011) using de-correlated white noise as the stimulus and varying the vertical spacing of the speakers in regular and irregular intervals and asking subjects to compare the resultant sound-field with that of a reference of 13 loudspeakers. The authors found that the difference grade worsened when the number of loudspeakers in the vertical were reduced to three with larger than 90° spacing in the vertical, So it was concluded that at least four loudspeakers of less than 45° vertical intervals were needed to produce what the authors term spatially uniform sound (a continuous soundfield with no gaps). In terms of an irregular layout, which would probably be found in most consumers homes, the results showed that when no loudspeakers were located directly above the listeners head e.g. +90° a minimum of five speakers were required at an elevation of 60°.

A similar study has been carried out by the national broadcaster in Japan (ITU, 2012) (NHK), who evaluated several different speaker configurations utilising decorrelated noise signals. This was to determine the minimum number of speakers using a height configuration that was needed to give significant improvements over horizontal systems. It was found that a configuration of two rings of 8 speakers, one at ear height and another above the listener at 45° was enough to provide a significant improvement in terms of envelopment. Further, they found that an increase in speakers beyond 12 in the horizontal plane saturated the sensation of envelopment providing no further improvement. Yet a further increase in speakers in the vertical plane continued to give further improvements. These results differ from the recommendations of the previous author, since the speakers were placed at an elevation of 45° whereas Oode et al. (2011) recommends that when no speaker is placed above the listeners head the elevation of the height channels should be 60° to produce spatially uniform sound.

More recently a study by Sawaya et al. (2013) investigated the effect of changing the angles of elevation and azimuth of a 22.2 system on the perception of envelopment. In order to evaluate the perception of envelopment two different sound sources were used one was a recording of an organ in a cathedral the other source was of de-correlated white noise. A reference system was used which was the 22.2 system. The first experiment focussed on changing the azimuth angles of speakers on the top layer using the angles 15°, 30°, 60° and 75° the second focussed on changing the elevation angles of the height channels. A double blind triple stimulus with reference method was used and listeners were asked to assess the impairment of spatial impression on a five point scale compared to the reference. The changes in azimuth for the two different sound sources for the horizontal layer showed significant differences, however for the top front layer this showed no significant difference. Their second experiment on changes in elevation angle on envelopment used a reference scenario of 8 speakers placed at an elevation angle of 45°. Three other conditions were compared against the reference which were 15°, 30°, and 60° elevations. Using the same test method as before participants compared the different elevation configurations to that of the reference. It was found that as long as the centre top channel was held constant the changes in elevation made no significant difference to the perception of spatial impression.

Another evaluation was carried out by Kim et al. (2013) this was similar to the previous author, in that the angular positioning of the two elevated channels was investigated for its

effect on the perception of envelopment. The test content was created by convolving impulse responses which were captured in a concert hall using a multichannel microphone configuration with three different elevations for the microphones capturing the height content. The reproduction system consisted of an ITU 5.1 system in the horizontal with two height channels placed at 25° elevation and three different azimuths 90° 110° 130°. Note that these angles are greater than those used by the previous author and no mention was made of the speakers being screened from participants view. Five test conditions were created which consisted of anechoic music convolved with the impulse responses of the three different elevated speaker positions, and also a virtual height surround created from transaural reproduction. Fifteen listeners ranked the reproductions using the appropriateness of reproduction as the criteria. The authors found that the positioning of the elevated channels, in terms of azimuth, had a significant effect, contrary to the previous author who used a different set of azimuth angles for the elevated channels and found no significant difference. They do not state which azimuth angles were significant. Further, listeners were also asked to provide descriptors for the reproduction which was ranked highest, these were categorized into five attributes which were naturalness, spaciousness, envelopment, immersiveness and image dimension.

2.11 Spatial Audio Quality in 3D Surround

A large amount of research has been carried out into 5.1 surround systems, and it has been found that a there are a number of attributes that make up the overall listening experience. Previously localisation and envelopment have been the focus of research by others, however more detailed studies have went further and identified that the spatial qualities are multidimensional (Bech, 1999), (Zacharov, 2000).

With the view of increasing the listening experience by adding height channels it is necessary to investigate the effect this has on spatial perception. However a review of some of the previous research into horizontal surround systems will be given, since this highlights some of the preferential attributes of horizontal surround reproduction.

There are a number of ways of evaluating the spatial quality of systems (Berg, 2006) highlights the areas that need attention when evaluating the spatial quality of systems. Berg considers that the sound scene can be evaluated as a whole or broken into its constituent parts meaning that a global assessment of the overall sound quality can be made or that specific attributes can be evaluated. Identifying important attributes that could be used to assess sound

quality can be achieved in two ways. The first method can be derived by considering previous research and using attributes identified by other researchers, or by utilising elicitation techniques which may provide more reliable results. The former method providing constructs defined by previous research may not be useful, since participants may not be able to relate to these and elicitation techniques whilst being more valid since subjects can relate to them, take longer to execute. Other considerations like the reproduced sound scene may be complex and will excite a number of perceptions confusing the participant or masking other attributes, so it may be necessary to use simple sound scenes instead. The listeners experience is also critical since an experienced listener will be more sensitive to different perceptual parts of a sound scene.

The aforementioned difference between listeners in evaluating sound-scenes was found by Paquier et al. (2011) when evaluating different types of surround recording techniques, using a Soundfield microphone with four capsules and an Eigenmike with 32 capsules and other different 5.1 recording techniques. In assessing the resultant sound-fields two types of listener were used, naive and expert. It was found that there were significant differences between expert and naïve listeners in terms of preference of the recording technique, with higher scores given for the non-coincident techniques by the expert listeners and first order ambisonics receiving the lowest score, although this was not true for naive listeners who tended to prefer the HOA coincident mic to the non-coincident techniques. It was stated that the expert listeners were disturbed by the high amount of rear sources reproduced by the coincident microphone techniques, however this was not mentioned by the naïve participants. The authors point to the fact that expert listeners are used to a full frontal soundstage with the rear channels being used for ambiance, and that the reproduction using the higher order microphone provides a more realistic sound scene.

Similar to the previous author Berg and Rumsey (1999a) evaluate several different recording techniques ranging from mono to stereo to 5.1, and also different electronic spatial processing techniques. Different sound scenes were recorded which ranged from orchestral recordings to outdoor environments reproduced over mono, stereo and 5.1 systems. In order to explore the multidimensional spatial attributes the authors employ the repertory grid technique, the participants were presented with a triad of each stimuli and were asked to use their own constructs to describe what they perceived. In the analysis of their results it was found that non spatial attributes were also elicited, these were mostly concerned with spectral properties

of the sound. However, mostly spatial attributes were found even if participants were not given instructions to look for spatial properties. The most frequent construct was related to naturalness, the next common attribute was related to source width and also a feeling of being surrounded. Summarising the most frequently elicited attribute these were found to be naturalness/authenticity, lateral positioning/source size, and envelopment/ depth.

Developing their previous study Berg and Rumsey (2001) seek to verify the attributes elicited from their previously described experiment, some of which were naturalness, envelopment, source distance, and room size. Several sound samples consisting of single instruments played back in reverberant environments were used. In order to explore the data further a factor analysis was carried out. This showed that there were three factors to the reproduced sound scene. The first was related to general width aspects, naturalness and preference. The second was related to source distance, room level and presence, then finally the source image focus. A correlation carried out between attributes found that there was a high correlation between envelopment/ preference, naturalness/presence, and room width/room size.

The previous investigations solely focussed on the perception of surround sound on the horizontal. However, it is highlighted that important preferential attributes like naturalness, envelopment, source size, are concerned with the source. And there are other attributes linked to the perception of an environment in terms of room width and room size.

The next detailed investigations include perceptual evaluations which include the height element providing an insight into how this impacts on the listening experience.

A number of reproduction methods were investigated in terms of spatial and general audio quality by Zacharov and Koivuniemi (2001) amongst which included 5.1, and an 8 channel periphonic system. Using an elicitation method they were able to extract information regarding the spatial attributes of each system. Using several samples of different spaces indoor and outdoor, and employing several different recording techniques of which included stereo microphone techniques, Soundfield microphone, and a head and torso simulator. The authors were able to extract several attributes of which related to spatial and also timbre properties. Some of the attributes were broadness, direction, and distance. The timbre properties were hardness, tone colour, and naturalness. When looking at the attribute hardness in terms of systems it is noted that the HATS (head and torso simulator) system is perceived as being hard, whilst the periphonic system is noted as being soft, possibly due to

the low frequency content of the periphonic system which could be attributed to the decoding method used, since the decoder used by Zacharov *et al* was only a basic decoder which only optimises the low frequency content. Whilst the HATS system has a large amount of high frequency content, which could have been caused by the crosstalk cancellation method used to reproduce the binaural recordings over loudspeakers. Investigating the attribute broadness it is found at the lower end of the scale the mono system is found having a narrow image, whilst at the other end of the scale the periphonic system is found providing broad or enveloping sound. In terms of direction it was found that the mono systems provided the best sense of direction whilst the periphonic and hats system provided ill-defined sense of direction. In terms of naturalness it was found that the anechoic or heavily reverberant samples were found to be most unnatural whilst outdoor samples or familiar samples were found to be more natural.

Investigating methods of soundscape reproduction Guastavino and Katz (2004) evaluate several methods of reproducing soundscapes, using a six channel system and also a 12 channel 3D system. Two other variants of the systems were also used which included a subwoofer. Five Parisian soundscapes were recorded by means of a Soundfield microphone and replayed over each of the systems using 1st order ambisonics. 27 participants took part and were asked to describe each of the different reproductions and pick which one sounded the most true to life and justify their choice. The descriptive words elicited were reduced to presence, readability, distance, and colouration. It was noted that even though only the spatial properties were changed i.e. the system presentations, attributes related to timbre also appeared. Participants described the 2D systems as well defined and enveloping and providing a good immersion in the scene. Whilst the 3D systems were described as poorly enveloping and sounding further away and that the reproduced space was poorly defined and the timbre was described as being muffled. This is contrary to the findings of Zacharov and Koivuniemi (2001) who found the 3D system broad, and enveloping using sound-field recordings, but with a slightly different speaker array. One reason for this is the fact that in the experiment conducted by Zacharov and Koivuniemi they utilised a low frequency decoding scheme was used and in the Guastavino and Katz experiment an in-phase decoding scheme was used which creates a larger listening area primarily designed for live sound events this could have caused the sound events to have appeared further away from the listener. A second evaluation was then carried out which this time included a stereo configuration using the same test regime was used as before with scales developed using the

descriptors from the previous test. The authors found that the rating of the different attributes depended heavily on the type of soundscape used, this also separated the respective systems. For example the stereo and six channel surround system were preferred for the concert music scenes, whilst for indoor sound scenes the 3D system was preferred. A principal components analysis showed that there was a clear difference between the stereo and the surround 2D and 3D system in terms of 'distance', 'presence', and 'stability'. Correlations of the attributes with preference found that there were high correlations between sense of space, sense of depth, and sense of movement, which reinforce the statements of Rumsey, whilst penetration and timbral emphasis were negatively correlated to preference.

Investigating future surround height system Kim et al. (2010) used the NHK 22.2 system as a reference with 9 elevated channels and several other systems with 4, 3, and then 2 elevated channels. The audio was created for the 22.2 system using VBAP (vector based amplitude panning) and then down- mixed to the other systems. Another method involving recordings made by a Soundfield microphone were also used and processed for the loudspeaker arrays using DIRAC (directional audio coding). It was found that the reproduction system configuration had a significant effect, as did the audio material used. Ratings were given using a MOS (Mean Opinion Score) based on the overall quality. Configurations with 4 and 3 height channels received high MOS scores. It is explained however that the system with 3 height channels, was perceived to be not much different than that of the NHK system with 9 height channels, illustrating that a large number of height channels may not be economical or needed. Further, their test methods were limited since they only asked participants to judge the overall audio quality, which is unfortunate as it is difficult to pinpoint exactly in what way the speaker arrays were affecting the listening experience.

A more revealing method was used by Ko et al. (2010) into three systems one being 5.1 which was used as a reference and the other two utilising 3D,these being a 10.2 system and a 22.2 system. They explore three main groups of sound quality which they term as timbre, localisation, and spaciousness. The main attribute 'timbre' features sub attributes naturalness and distinctness, with the main attribute 'localisation' having sub attributes, direction , distance movement and volumetric. Finally the main attribute 'spaciousness' has sub attributes which were envelopment and apparent source width. The stimuli consisted of several sound-scenes which included movie sound scenes and music, there was no mention of the rendering method used. It was found that the 22.2 system was rated significantly better

than the 10.2 system in terms of all the attributes tested including envelopment. This result is contrary to the previous author Kim et al. (2010) who found that there was not much difference between the 22.2 system with 8 elevated channels and that of a system with only 3. Further analysis of their results using principal component analysis found that for moving sources 'distance' and 'naturalness' were major attributes for sound quality. Conversely for non-moving sources 'envelopment' and 'apparent source width were major attributes for sound quality. However the validity of their results is questionable since they only used 7 listeners. A final comment from the author's state that the attributes such as envelopment and naturalness may be reproduced better with a large amount of speakers, however no reasons for this were given.

Another evaluation of 3D systems and horizontal surround systems was carried out by Silzle et al. (2011) In the evaluation several systems were tested which included stereo, 5.1, NHK 22.2 and a 9.1 system using VBAP. Several sound scenes were evaluated which consisted of music and atmospheric content. Down mixing of signals from the reference 22.2 system was used as well as selective muting of channels to create content for the other systems. In the first test a MUSHRA test methodology was used using the 22.2 system as a reference and listeners were asked to rate the overall sound quality, then in the second test no reference was used. From the evaluation it was found that listeners rated the 22.2 system as having the best sound quality for the test with a reference, although in the test without reference some excerpts for the 9.1 system with four elevated channels were scored slightly higher or the same as the 22.2 system, this being a similar outcome to that of Kim et al. (2010) where it was found that a system with only three height channels gave a similar performance to that of the NHK system with nine height channels. Finally it was stated that overall the preference was for surround with height, however that this was stimulus dependent and that further research was required.

A variant of descriptive analysis called repertory grid technique was used by Avni et al. (2013) This method uses triads of stimuli in which the listeners use descriptive words to describe the differences and similarity between stimuli and generate opposite terms. To generate stimuli the authors simulated a reverberant environment using the image source method. In order to capture the soundfield two methods were used, the first consisted of using a spherical microphone array of finite number of capsules and the second involved calculating the pressure on a rigid sphere to avoid spatial aliasing. Impulse responses of the

simulated room were captured using the two methods and decomposed using spherical harmonics. To reproduce the test items which were castanets and speech, HRTF's were convolved with the sound samples and reproduced over headphones with head tracking using various spherical harmonic orders of which were 0, 1,2, 5, 12 and 22.After the elicitation phase listeners then used their own descriptive words to then rate each of the stimuli. However the test lacks rigour since only five listeners were used, of which three were the authors of the paper. The analysis of their results found that as the spherical harmonic order increased, the perception of the timbre was more balanced and brighter, the source was easier to localise and the room was perceived as more spacious. It was also stated that most of the attributes gathered in the experiment could be divided into three groups: timbre (dark vs bright), spaciousness (wide vs narrow) and artefacts (ringing vs no ringing). It is concluded that reproduction using binaural synthesis using higher order spherical harmonics will lead to a more externalised, thinner and further away source with timbre that is more balanced.

2.12 Summary

Some studies have shown that the rendering method used for example 1st order ambisonics recorded with a Soundfield microphone have provided different results. One study suggesting that 1st order ambisonics with height provides an improvement whilst another suggesting that horizontal surround is better in this context. This lack of clarity in the literature needs to be resolved for 3D sound reproduction to be used to its full potential.

It is also suggested that using a higher order ambisonic reproduction produces a sound scene which is perceived to be further away and produces a more spacious reproduction, with less high frequency attenuation creating a more balanced timbre. Therefore it is of interest to establish if using higher order ambisonics provides a significantly better reproduction than lower order and in what way.

It seems that overall there is disagreement on the loudspeaker configuration in terms of number and positioning for optimal reproduction. Therefore another area of investigation would be to evaluate the number and positioning of speakers on preferential attributes. Finally, most of the studies reviewed concerning 3D surround have used the attributes gathered from horizontal surround research, which is not an unreasonable starting point, however it may be better to investigate from a less biased viewpoint using elicitation based methods to allow listeners to provide constructs.

2.13 Initial Area for Investigation

From the literature review it had been identified that human localisation in the vertical plane is not as accurate as the horizontal plane, in addition there were also peculiarities in human localisation dependent on where the source was positioned in the vertical plane. It has been discussed that there are some studies into the use of ambisonics for rendering elevated virtual sources and there has been a direct comparison between 1st and 2nd order ambsionics in the median plane which found no significant reduction in localisation error between the two orders. It may be that an increase to 3rd order may provide a significant reduction in localisation error over 1st and 2nd order. The first area of investigation will seek to answer the questions:

1. Does the rendering of sources in 3rd order ambisonics help in the ability to localise elevated sources?
2. How is localisation affected by placing virtual sources at an elevation?
3. What role does the spectral content of a source play in its ability to be localised?

3 Surround Sound Development

3.1 Introduction

This chapter starts with an account of the development of spatial audio systems from mono to the current 5.1 system and the perceptual benefits with the increase in reproduction channels for each system. The current systems which extend the listening area into the vertical plane are detailed and the claimed perceptual benefits to be gained from extending the soundfield into the vertical plane are also discussed. Details of the rendering methods used to pan virtual sources in the vertical plane are also explored along with their perceptual consequences, detailing the mechanisms used in human hearing to localise sources in the horizontal and vertical plane and the differences.

Finally the parallels between concert hall acoustics and the 5.1 system are detailed, describing the perceptual effects of lateral reflections leading to the perceptual effects of reflections from the vertical plane in concert hall acoustics and the consequence of utilising height speakers to render the vertical component recorded in a live space are also discussed.

3.2 Development of Spatial Audio Systems

3.2.1 Mono

The first available speaker reproduction systems were monaural, where a single speaker reproduced all of the recorded sound with no separation in the reproduction, using one microphone to record the sound scene (Watkinson, 1998). It is also stated by Steinberg and Snow (1934) that when only a single channel is used some of the depth properties of the original sound may be produced, but the spatial character of the original sound is imperfectly preserved, Steinberg and Snow then suggest that to reproduce the spatial character there are two ways, using two or three speaker channels or binaural reproduction.

3.2.2 Binaural

It was not until the late 1930's that monaural technology reached maturity to allow inventors to start considering multichannel audio that had the ability of portraying spatial information to the listener, however as far back as 1880 Bell labs were experimenting with binaural reproduction. However, the first notable demonstration of spatial audio by Clement Ader in 1881 at the Paris opera. This was achieved by setting up microphones across the stage and the onstage performance was then transmitted to remote listeners who were listening using headphones in other parts of the building (Davis, 2003).

Although Ader's reproduction was only in binaural utilising only two channels of audio, which provided limited spatial information. Several years later in1940 a conductor named Leopold Stokowski along with Disney and RCA engineered a system called 'Fantasound'. This was a commercial system for theatres which included three frontal channels coupled with speakers dispersed around an audience, the rear speakers were used sparingly to enhance envelopment. The additional speakers provided further complexity in terms of panning, so Disney's engineers were also responsible for developing a method of panning sources reliably from the front to back using a device called "Togad" (Tone Operated Gain Adjusting Device", Which worked by using separate control tracks to handle panning of sources (Davis, 2003).

3.2.3 Stereo

The commercial multichannel systems beginnings can be attributed to two men namely Alan Blumlein and Harvey Fletcher. Alan Blumlein was born in Manchester in 1903 and was said to have developed over 128 patents one of which was for stereo reproduction, recording and transmission (Blumlein, 1931), however the patent was not put to practical use until the 1950's (Torick, 1998)

Harvey Fletcher born in 1884 among some of his achievements were in the establishment of the Fletcher Munson curves along with Wilden Munson who was also one of the founders of the acoustical society of America. Although most notably known for his experiments into stereophonic sound, yet different from Blumlein, was more concerned with three channel techniques, with his first experiments using a wall of microphones which picked up an orchestra at one side, with the reverse side covered in speakers which created the virtual image. However, this was later whittled down to three channel techniques due to the poor localisation of human hearing in the vertical plane (Streicher and Everest, 1998).

The stereo speaker layout utilises an equilateral triangle with the listener placed at the apex of this triangle, the loudspeakers forming the base of the triangle, with speakers dispersed at ±30° in front of the listener. *See Figure 3.1.*

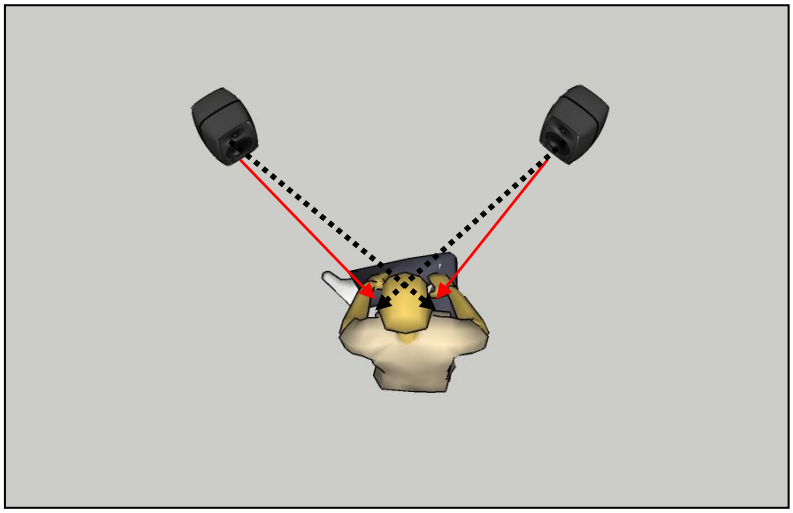


Figure 3.1 Stereo Configuration

The stereo system has limited capability to produce spatial effects, since only a limited frontal reproduction can be created (Rumsey, 2005).

It can also be seen (dashed black lines) that a delayed version of the left speaker also goes to the right ear and vice versa for the right speaker called crosstalk (Bartlett, 1991). This means that for frequencies up-to 700Hz the signals will differ in phase between the ears and for frequencies of 1200Hz and greater there are level differences between ears due to head shadowing of higher frequencies, similar to listening to a real source (Talbot-Smith, 2001). Further, the phenomenon of summing localisation can be observed by driving the speakers with two identical signals, this will cause the listener to perceive a centrally located virtual source between the speakers in the median plane (Blauert, 1996).

Further manipulation of the signals in the stereo field can be created by using amplitude or time differences between the speaker signals, allowing the impression of virtual 'phantom sources'(Rumsey, 2005). There are also underlying psychoacoustic principles which allow the formation of phantom sources which will be explained in *Section 3.8*.

Blumlein was also responsible for inventing methods of recording stereo sound scenes by firstly using a spaced pair of omni directional microphones in order to mimic how the human hearing systems worked (Blumlein, 1931). This method was then refined with the development of figure of eight microphones which were then arranged as a coincident pair with their grills crossed at a 90° angle, this technique is known as the Blumlein pair. In addition to capturing the frontal sound source, the rear facing negative lobes of the figure eight microphones used in the Blumlein pair also pick up the reverberant room sound

(Rumsey, 2005). This is said to cause a spread in the captured room response outside of the speaker angle providing a greater source width (Wells, 2010).

Another variation on the coincident microphone technique described by Blumlein (1931) is the mid and side, this utilises a forward facing cardioid microphone (the mid component) with a figure of eight microphone (the side component) with its null facing the sound source. The stereo signals are derived by the mid+side for the right channel and the left channel by mid-side. The signal can be further manipulated to create a more spacious reproduction by boosting the bass in the left or right signal (Bartlett, 2013)

The remaining stereo microphone technique is that of the spaced pair, this utilises a spaced pair, usually several feet apart, of omni directional microphones, this configuration utilises time differences to encode sounds, one characteristic of this configuration makes off centre images diffuse and unfocussed, but provides a warm sense of ambience and a spacious sound (Bartlett, 2013).

3.2.4 Quadraphonics

Quadraphonics was spawned from a series of test recordings by Robert Berkovitz at the Acoustic Research Corporation with the intent to show what happens when a pair of rear channels were added to a front stereo pair, thus avoiding the compromise in stereo reproduction by adding enveloping ambience behind the listener (Torick, 1998). Quadraphonic reproduction used four speakers, two placed at ±45° in the front and two at ±135° behind the listener, *See Figure 3.2.*

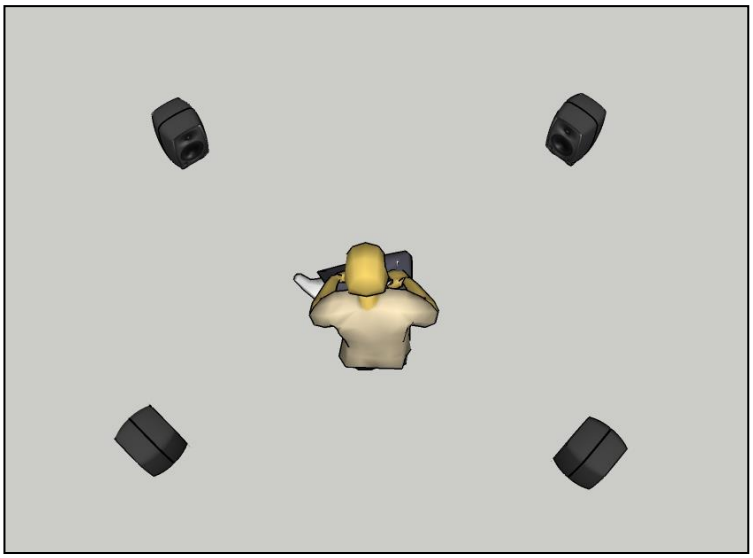


Figure 3.2 Quadraphonic Layout

However with four speaker reproduction came with it the problem of broadcasting and placing four tracks on two track stereo media.

An audio engineer named Peter Scheiber derived a matrix system, others had suggested four track tape systems however Schreiber's was unique as it used a method of taking four channels down to 2 by using a 4-2-4 matrix, which used wideband phase shifting methods to derive the rear channel signals, although quadraphonic reproduction never took off due to its increased production costs and the requirement of consumers to purchase more equipment (Torick, 1998).

There were also perceptual issues with quadraphonic layout, it has been shown by (Ratliff, 1974) that the quadraphonic layout has consequences when trying to create images to the side of the listener causing frontal dominance, this was also found by (Philipson et al., 2002). In addition the larger than stereo frontal angle causes phantom images to be perceived as elevated (Ratliff, 1974), this effect was also reported by (Choisel and Wickelmaier, 2007).

3.2.5 5.1 Surround

In 1975 Dolby developed Dolby stereo, a method of placing two track information on a film track instead of the mono system already in place. To do this required placing four channels of information on two tracks, which was achieved using the similar matrix technology previously explained that was used for quadraphonic surround systems. This then provided a left, right and centre channel and a single surround channel, which cemented the new formats place in film and audio entertainment (Dolby, 2011)

However there was a need to move to a discrete system, since problems caused by matrix encoding of channels meant that speech reproduced from the centre channel caused background music to modulate (Holman, 2008) and that there was only one surround channel which was not a problem for film sound where sound effects were used, but was problematic for conveying spatial properties of room sound (Steinke, 1996). Therefore a discrete system was suggested, which included in addition to the three frontal channels, one at 0° and two at ±30°, with two surround channels at ±120° and a subwoofer *See Figure 3.3*, this was first characterised by Thomlinson Holmann as 5.1 and then standardised in 1991 as the ITU-R Rec. B.S775-1. The 5.1 system was also designed to be compatible with the current stereo format (Rumsey, 2005).

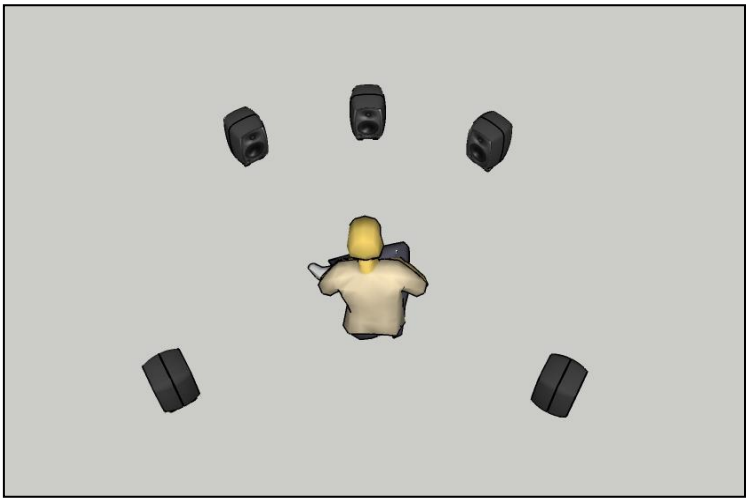


Figure 3.3 ITU 5.1 Layout

The discrete 5.1 system provided perceptual benefits over the other multichannel formats like stereo and quadraphonics. The front centre channel anchored the stereo soundfield for off centre listeners and provided better tonal balance over a phantom centre (Holman, 2008).The rear channel positions providing a balance between the reconstruction of lateral energy and panning of sources to the rear of a listener (Rumsey, 2005) and in addition the ability to envelope the listener using the rear channels (Morimoto, 1997), (Martin, 2005), or to generate a new artificial room acoustic (Theile, 1993). Further, it is stated by Stuart (1996) that the move from stereo to multichannel is significant, providing better sound source segregation with lower masking thresholds. However, the orientation of the speakers in the 5.1 system meant that panning of sources to the side of the listener caused images to be unstable. This was because of the large gap between the front and rear channels (Martin et al., 1999b) and in addition, it is stated by Holman (2008) that 5.1 coarsely quantizes a soundfield and panning between the front and back speakers to the side of the listener causes sounds to jump between front and rear channels, similar to the problems with quadraphonics.

Small developments have been made to 5.1 surround systems by adding more speakers, like 7.1 which adds two extra speakers centre left and centre right, primarily designed for cinema applications with wide screens to fill in the gaps between the centre and left/right channels (Rumsey, 2005) In addition further up-mixing schemes have become available like the Dolby EX system and Lexicons Logic 7 system which allow 5.1 material to be up-mixed to derive additional channels, these generally include a rear channel directly behind the listener and also additional side channels, to help alleviate the problem of creating images to the side of a listener. Some of these up-mixing systems have been tested by Chaffey and Shirley (2007) to

investigate the improvement in localisation over the current 5.1 system, and it was found that the up-mixing systems did significantly improve localisation for lateral sources over the current 5.1 configuration.

Investigations have also been carried out into increasing the number and positioning of horizontal channels, in order to enhance the spatial impression experienced in concert hall soundfields and it was found that the 5.1 system was the optimum number of channels to achieve this (Hiyama et al., 2002).

3.2.6 Summary

There has been significant development of sound systems in the horizontal plane from the mono system to the surround systems available today and it would seem that small improvements have been made by adding channels to the standardised 5.1 system. Recently there have been a number of developments in order to extend sound systems to include height in an attempt to develop surround systems further.

3.3 Surround With Height

There have been notions of adding height to surround systems for many years now. Full 3D systems can be traced back as far as the 1970's where Michael Gerzon was proposing not only a reproduction system, but a complete recording, transmission and reproduction system encompassed in what is termed ambisonics (Gerzon, 1973). There is renewed interest in 3D surround systems and there are some emerging systems becoming available with each system having various numbers of channels and configurations and different justifications for this.

3.3.1 NHK 22.2

A surround system with height, which has been the subject of a large amount of research is the NHK 22.2 system. This system uses 10 channels at ear height, 9 channels above the listener and 3 frontal lower channels along with 2 subwoofers *See Figure 3.4*. It was first presented at an Audio Engineering Society convention (AES) in 2004 and was accompanied by what was termed ultra-high definition video.

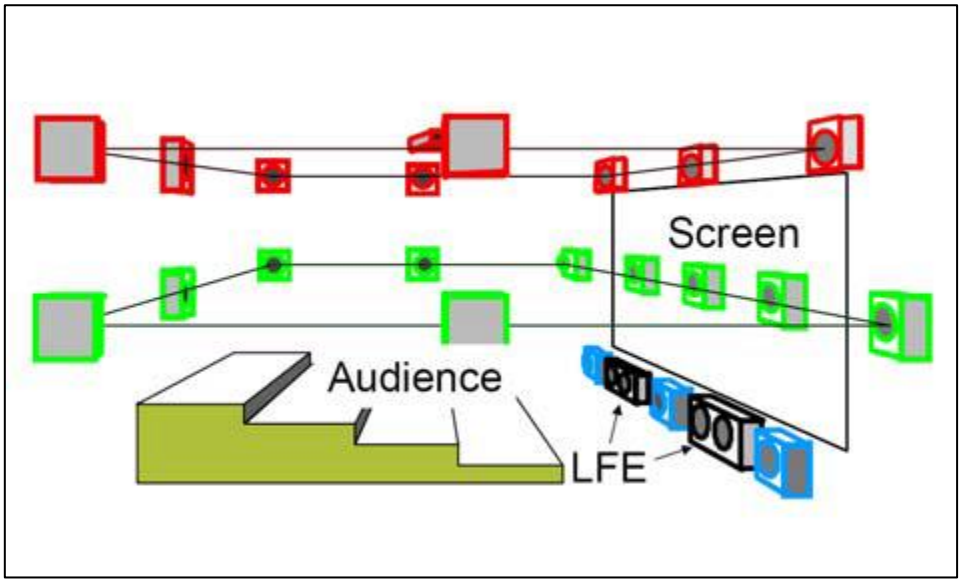


Figure 3.4 NHK 22.2 Layout (Hamasaki et al., 2004a)

This system is ultimately designed for commercial use i.e. cinema presentations. The system has been the subject of intense research which has included the development of custom panning systems and ultra-high definition video (Hamasaki et al., 2004a). In addition, further perceptual studies claim that the large amount of frontal channels and overhead channels improve the reproduction of presence, localisation and depth over a larger area than the current 5.1 configuration, however this was found to be dependent on the sound scene used (Hamasaki et al., 2004b). More recent studies into the effect of the height channels on the size and shape of the listening area, also showed that increasing the number and positioning of speakers increased the listening area size (Sawaya et al., 2011).

3.3.2 Dolby Pro Logic IIz

The first commercial surround system to include height was the Dolby Pro Logic IIz system *See Figure 3.5*. This system utilises two height channels above the two front stereo channels at an elevation of 45°. The audio for the two front height channels is said to be mainly non directional diffuse content and are encoded in a 5.1 or 7.1 soundtrack using a matrix encoding scheme where the two front height channels are contained in the rear speaker stems. The extension of the soundfield using the height channels is said to produce a better sense of depth, greater envelopment and an improved sense of acoustic space (Tsingos et al., 2011). Up-mixing technologies have been developed that allow two frontal height channels similar to the Pro logic IIz system and in addition add extra wide front and rear channels from standard 5.1 material, this system developed by Audessy, is called dynamic sound xpansion

(DSX) (Audessy, 2014). I t is claimed that this allows the blending of front and rear channels to create a seamless and more enveloping soundstage.

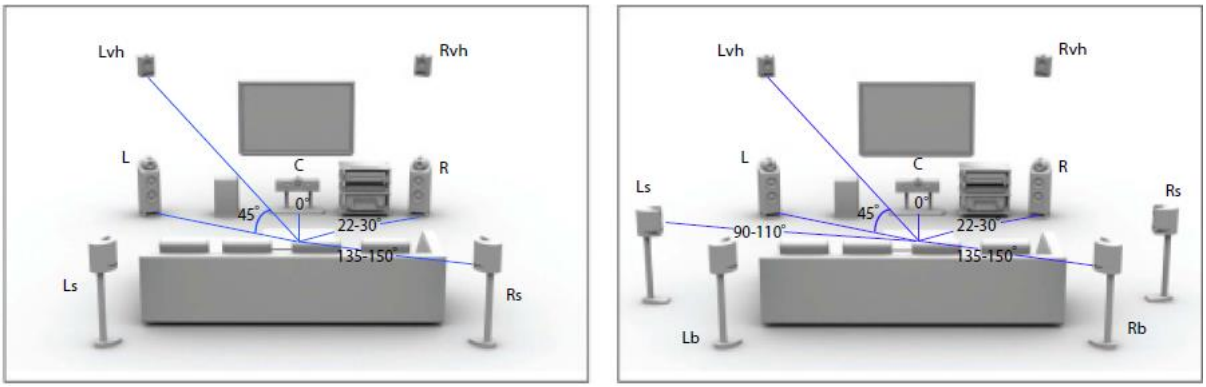


Figure 3.5 The Dolby IIz Configurations Left 5.1 and right 7.1(Tsingos et al., 2011)

3.3.3 10.2

A further number of systems have recently been standardised and are included in the international telecommunications union report (ITU) (ITU, 2011) which are the 10.2 system proposed previously by Holmann (Willis, 1999), (Holman, 2008) along with the NHK 22.2 system. The 10.2 system *See Figure 3.6* has also been the subject of investigation by Kim et al. (2010) and it was found that the 10.2 system was not significantly different to that of the NHK 22.2 system in terms of audio quality. Yet, Ko et al. (2010) using different sound scenes to that of Kim found that for the attributes used in his test which included envelopment and localisation the NHK 22.2 system was superior.

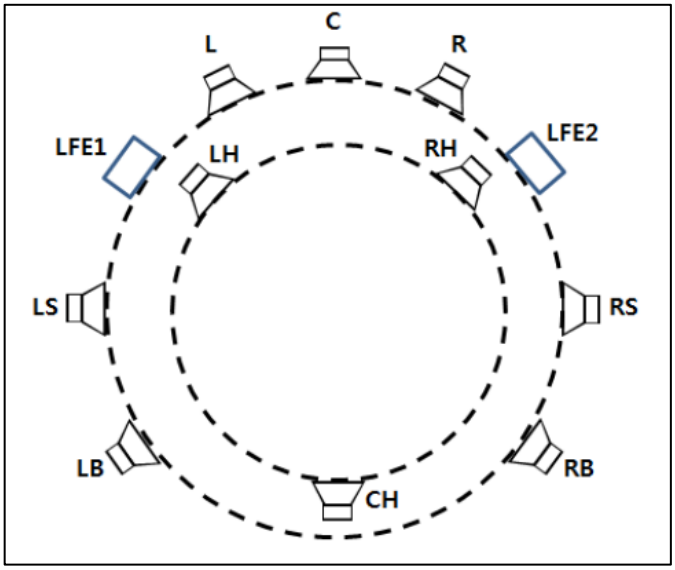


Figure 3.6 10.2 Configuration (Kim et al., 2010)

A variation on the 10.2 system has recently been developed called Auro 3D *See Figure 3.7.*This system utilises the current 5.1 configuration but includes four height channels each at a 45° elevation above all the channels excluding the centre channel. However other channel counts exist

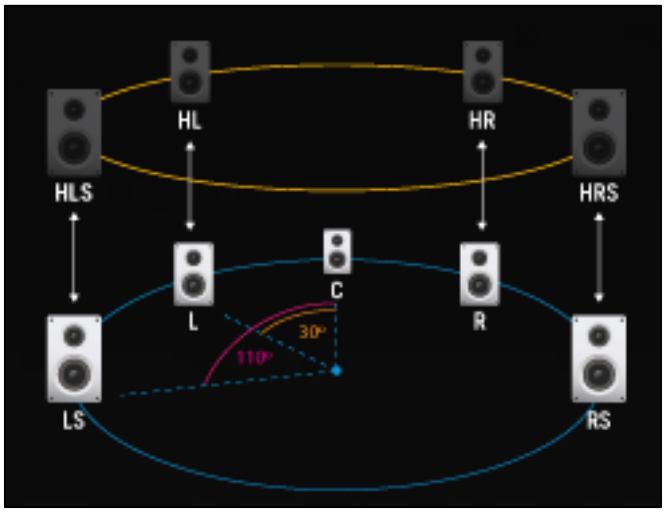


Figure 3.7Auro 3D Layout (Auro3D, 2014)

including 11.1 and 13.1, these systems are mainly for use in commercial cinema applications and include three layers, the standard ear height surround system and then two height channels, one at an elevation of 40° and also a channel directly above the audience at a 90° elevation called the ‘voice of god’ (Auro3D, 2014).

It is claimed that the height channels at a 40° inclination create the immersive feeling as heard in real life. It is also claimed that other leading 3D commercial systems do not have height channels between the ceiling channels and the ear height surround channels, only overhead stereo channels, as a result do not create a natural hemisphere over the listener, leaving the overhead height channels feeling disconnected from the rest of the system (Claypool et al., 2012)

3.3.4 Dolby Atmos

A similar system to the previously mentioned Auro 3D has become widespread in a number of cinemas worldwide called Dolby Atmos. This system utilises a similar horizontal surround

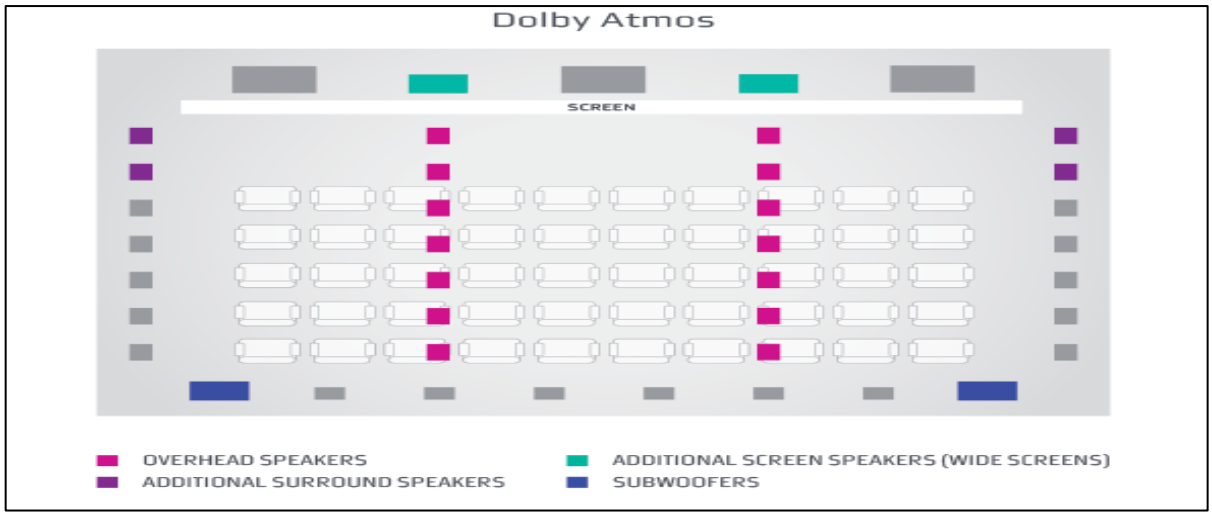


Figure 3.8 Dolby Atmos Layout (Dolby, 2012)

system but only provides an array of speakers overhead *See Figure 3.8*. With a total number of 64 channels available in larger theatres, it is claimed that the Atmos system compared to the 5.1system provides a more lifelike sound, maintains tonal sound quality as sources are panned and provides a consistent experience from any seat in the theatre (Dolby, 2014). The Atmos system utilises an existing 7.1 system but adds the opportunity of height through object orientated panning, this is where the sound sources are independent of the reproduction format and carry metadata regarding their orientation in 3D space along with amplitude and other features, this hybrid system based on the 7.1 cinema channel configuration and an object orientated system, means that cinemas without Atmos can use the 7.1 mix instead therefore maintain backward compatibility (Robinson et al., 2012).

3.3.5 Summary

An overview of the current 3D surround systems has been given and has illustrated that there are a number of systems which differ in speaker channel numbers and positioning. Further it is also reported that each system justifies their different arrays by claiming the perceptual benefits each configuration brings over the current reproduction systems which include, greater sense of acoustical space, greater sense of depth, more natural reproduction and a greater sense of envelopment. Furthermore the small amount of research that has been carried out is inconclusive and seems to contradict each other making it hard to get a clear picture of

the benefits of surround with height. In addition some poorly implemented research which is also commercially biased to a single system.

3.4 Spatializing Methods for 3D Reproduction

For the two previously detailed systems Auro 3D and Dolby Atmos it is undisclosed how these systems render their sound scenes. However, there are two well established 3D panning methods one is ambisonics and the other vector base amplitude panning.

3.4.1 Vector Base Amplitude Panning

Vector base amplitude panning (VBAP) for horizontal layouts utilises amplitude panning similar to that used in stereo between pairs of speakers. However the advantage of VBAP is in its ability to render sources in 3D speaker layouts.

The method behind the 3D panning works using triplets of speakers instead of pairs of speakers used in 2D. Triplets of speakers can be constructed above and in addition below the listener in triangles. The gain factors for each speaker can be calculated by considering vectors pointing to the loudspeaker from the listening position and the virtual source is created by considering a vector length which is represented by the weighted sum of the loudspeaker vectors *See Figure 3.9* (Pulkki, 2001). Similar to amplitude panning in the horizontal plane whenever a source coincides with a speaker position only that speaker will be operational, however there are drawbacks, for sources in the centre of the triangle all of the speakers will be active and it is stated by (Pulkki, 2001) that this causes a spreading of the virtual source, making the image less sharp.

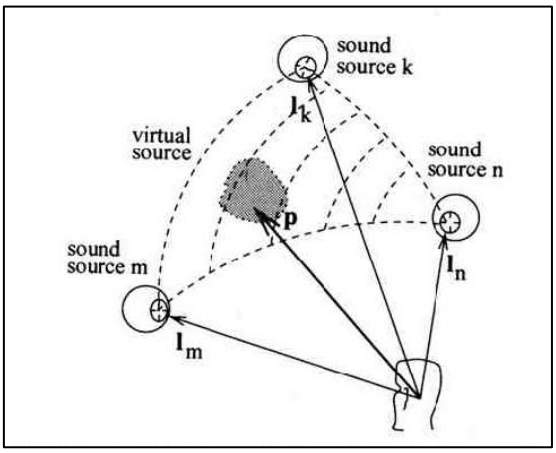


Figure 3.9 VBAP Panning Principle (Pulkki, 2001)

Further, for triangles placed laterally to a listener it will not be possible to place a virtual source there, similar to the problems with 5.1 systems, this is shown below in *Figure 3.10* where it is not possible to reconstruct stable virtual sources in the white circle. This is due to the lack of inter-aural level or time differences since the sound is arriving from one side of the head, this will be further explained in the following sections. Due to the aforementioned factors it has been shown by Blauert (1996) that the greatest localization blur is for positions to the side of a listener. Although it was demonstrated by Theile and Plenge (1977) that for what was termed an all-round effect or in other words stable side sources to the side of a listener required a speaker to be placed at this location.

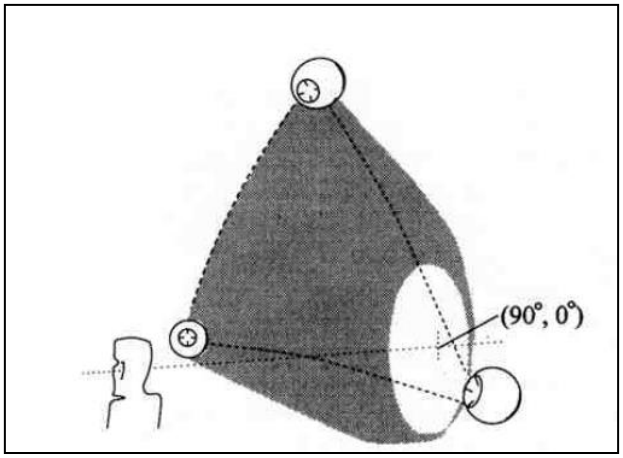


Figure 3.10 Area Where Virtual Images Are Not Stable (Pulkki, 2001)

VBAP is very flexible in terms of use with different speaker layouts and only requires that non overlapping triangles are formed around the listener. However, there are limits to how well these virtual sources will be perceived and it is explained that there are a few conditions that have to be met in order for correct reproduction of virtual sources. These include, non-overlapping triangles as this will allow smooth gain transitions between triangles and that triangle sides are a short as possible, since larger speaker distances affect virtual source quality (Pulkki, 2001).

Psychoacoustic testing has also been carried out using VBAP with elevated sources. It was found that for sources elevated in the median plane, the spread of the virtual source was perceived differently by each of the listeners and the timbre was different to that of the reference source (Pulkki, 2001). One reason for this is that VBAP relies on inter-aural differences and for elevated sources in the median plane there are no inter-aural differences,

this is not only peculiar to VBAP based systems but has also been found in systems utilising ambisonics (Pieleanu, 2004) and amplitude panning (Barbour, 2003b).

3.5 Ambisonic Encoding/Decoding

3.5.1 Introduction

The methods for the creation of 3D sound were outlined in the review section and it was noted that there are only two rendering methods available for the creation of surround with height, one being VBAP and the other ambisonics. VBAP is not a complete system i.e. it does not include a method of 3D recording, however ambisonics does allow the capture of 3D sound scenes and reproduction of surround with height.

Ambisonics was conceptualised by Gerzon (1973) and was a complete method of recording transmission and reproduction of not only horizontal surround but also what is termed 'periphonics' or full sphere reproduction. The recording and reproduction system created by Gerzon was based on the principles of human hearing in order to optimise the system, where for low frequency localisation ITD (inter-aural time differences) are used or for high frequencies where ILD (inter-aural level differences) are used (Gerzon, 1992).

The general 'Metatheory' described how the low frequency termed (rV) the velocity vector and high frequency regions termed (rE) energy vector of the soundfield could be optimised in line with how the human hearing system works at low and high frequencies, (albeit a simplified model), based on the duplex theory. The equations to calculate these vectors are shown below in *Equations 3.1, 3.2*.

Equation 3.1

$$P = \sum_{i=1}^{n} g_i$$

$$Vx = \sum_{i=0}^{n} g_i \, (\theta_i)/P$$

$$Vy = \sum_{i=0}^{n} g_i \, \sin(\theta_i)/P$$

$$E = \sum_{i=1}^{n} {g_i}^2$$ Equation 3.2

$$Ex = \sum_{i=0}^{n} {g_i}^2(\theta_i)/E$$

$$Ey = \sum_{i=0}^{n} {g_i}^2(\theta_i)/E$$

Where g_i represents the gain from the i[th] speaker, n is the number of speakers $and\ \theta_i$ is the position of the i[th] speaker.

It was also stated by Gerzon that the value of the velocity vector (rV) should be 1 and that the energy vector should be as close as possible to 1, however it would only ever be possible to have an energy vector of 1 if there was only one speaker playing. Further, that the energy vector (rE) should not be less than 0.5 since this will cause image instability and inferior reproduction (Gerzon, 1980)

In designing a speaker array for the reproduction of ambisonics, Gerzon stated three requirements in order to ensure correct reproduction based on the previous energy and velocity theory. These were termed the diametric decoder theorem and are as follows (Gerzon, 1980)

1. *All speaker are the same distance from the centre of the layout*
2. *Speakers are placed in diametrically opposite pairs*
3. *The sum of the two signals fed to each diametric pair is the same for all diametric pairs*

3.5.2 Ambisonic Encoding

There are two ways in which ambisonics can be utilised to reproduce a sound scene:

1. By recording using the Soundfield microphone or Eigenmike
2. By synthesis using ambisonic encoding equations

3.5.3 SoundField Microphone

The Soundfield microphone was created by (Farrar, 1979) using the design specifications by Michael Gerzon. This microphone uses four capsules mounted in coincidence with each other on the faces of a tetrahedron, *See Figure 3.11*.

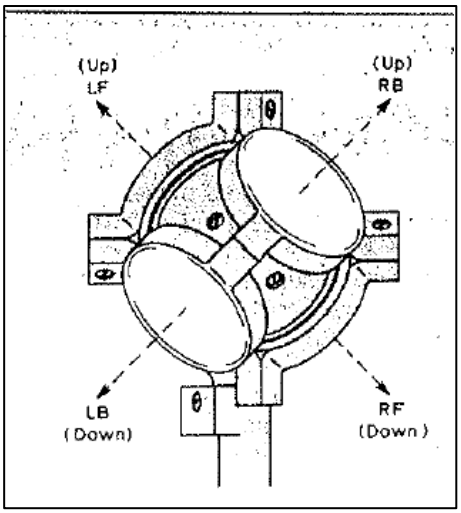


Figure 3.11 Soundfield Microphone Capsules (Farrar, 1979)

The output of the microphone includes four signals these are in the so called A format, which include, left front up (LFU), right back up (RBU), left back down (LBD) and right front down (RFD) each signal is then processed into the B format by correcting for the fact that each capsule in the microphone is not absolutely coincident. The B format signal components are achieved using the combination of signals captured by the microphone as shown below in *Equation 3.3.*

$$W = LFU + RFD + LBD + RBU = Omni \qquad \text{Equation 3.3}$$

$$X = LFU + RFD - LBD - RBU = Figure\ of\ Eight\ (Front/Back)$$

$$Y = LFU - RFD + LBD - RBU = Figure\ of\ Eight\ (Left/Right)$$

$$Z = LFU - RFD - LBD + RBU = Figure\ of\ Eight\ (Up/Down)$$

From the B format signals these can then be decoded to any speaker layout, therefore at the recording stage no knowledge of the reproduction layout needs to be known. However, there are some rules to ensure correct reproduction, which will be explained in the next sections.

3.5.4 Eigenmike

A newly developed microphone called the Eigenmike allows for the capture of higher order ambisonics from 32 capsules distributed around a sphere, *See Figure 3.12*. This allows the derivation of up to 4^{th} order ambisonics. However, to derive the ambisonic components requires processing the (A) format components of the mic (transcoding) and also further processing, because of the size of the sphere used to house the microphones is an appreciable

size compared to the wavelength of sound at high and low frequencies (Ihle, 2003), (Meyer and Elko, 2008).

Figure 3.12 The Eigenmike

3.5.5 Theoretical Encoding

The second method allows synthetic sounds, samples or even close miked recordings to be panned into a reproduced soundfield (B- Format). This can be done by using the encoding equations shown below in *Equation 3.4.* Where θ represents the azimuth of the source and φ represent the elevation of the source.

$$W = 1$$
$$X = \cos\theta \cos\varphi$$
$$Y = \sin\theta \cos\varphi$$
$$Z = \sin\varphi$$

Equation 3.4

These functions will then provide a 1st order B-format signal, as previously explained for the Soundfield microphone.

Once the signals have been recorded or encoded synthetically they then have to be decoded to the speaker layout of choice. This is handled by using the decoding equation. *See Equation 3. 5.* Where θ represents the azimuth of the speaker and φ represents the elevation of the speaker.

$$LS = W + \sqrt{2}\ (\mathrm{X}\cos(\theta) + Y\sin(\theta) + Z\sin(\varphi)) \qquad \text{Equation 3.5}$$

Where for a configuration of an octagon of equally spaced speakers around 360°, a speaker at 45° azimuth and 0° elevation, the speaker signal would be found using the equation shown below in *Equation 3.6.*

$$LS = W + \sqrt{2}\ (\mathrm{X}\cos(45) + Y\sin(45) + Z\sin(0)) \qquad \text{Equation 3.6}$$

This would then provide three signals which are combined for each speaker position to reconstruct the soundfield at the listening position. As can be seen for horizontal layouts the Z component will always be 0, therefore this part of the equation can be dropped for horizontal only layouts.

The previously explained method created what is termed a 1st order decode this is the most basic version of ambisonic reproduction. However it is possible to create higher resolution reproductions (termed higher order), for example using 2nd order. The equations to do this build on the previously described 1st order encoding equations by adding further components shown below in *Equation 3.7.* Where θ represents the azimuth of the source and φ represents the elevation of the source.

$$R = (3\sin^2\varphi - 1)/2 \qquad \text{Equation 3.7}$$

$$S = \sqrt{3/2}\cos(\theta)\sin(2\varphi)$$

$$T = \sqrt{3/2}\sin(\theta)\sin(2\varphi)$$

$$U = \sqrt{3/2}\cos(2\theta)\cos(\varphi)^2$$

$$V = \sqrt{3/2}\sin(2\theta)\cos(\varphi)^2$$

This adds further resolution to the soundfield, however as a final stage a further number of components can be added which allow the reproduction of 3rd order ambisonics which are shown below in *Equation 3.8.* Where θ represents the azimuth of the source and φ represents the elevation of the source.

$$K = \sin(\varphi)(5\sin^2\text{-}3)/2 \qquad \text{Equation 3.8}$$

$$L = \sqrt{3/8}\cos(\theta)\cos(\varphi)(5\sin^2\varphi - 1)$$

$$M = \sqrt{3/8}\sin(\theta)\cos(\varphi)(5\sin^2\varphi - 1)$$

$$N = \sqrt{15/2}\cos(2\theta)\sin(\varphi)\cos^2\varphi$$

$$O = \sqrt{15/2}\sin(2\theta)\sin(\varphi)\cos^2\varphi$$

$$P = \sqrt{5/8}\cos(3\theta)\cos^3\varphi$$

$$Q = \sqrt{5/8}\sin(3\theta)\cos^3\varphi$$

The previously shown ambisonic equations upto 3rd order encompass the full set of equations as specified by Malham (2003), however equations to derive soundfields beyond 3rd order are detailed in a paper by (Chapman, 2009).

3.5.6 Resolution of Different Ambisonic Orders

The effect of adding further ambisonic components on the directional resolution of an ambisonic panning function can be viewed by plotting the gains of all the speakers in the array used to pan the virtual source. Ambisonics belongs to a family of amplitude panning techniques, however uses a slightly different method of panning, because all of the speakers in the surround sound array work at once to pan a source. However, with the increase in ambisonic order the contributions of the other speakers becomes less. Below it can be seen moving from left to right in *Figure 3.13* the expected polar pattern for 1st order, 2nd and finally the 3rd order can be seen. This demonstrates that increasing the ambisonic order a finer directivity can be achieved.

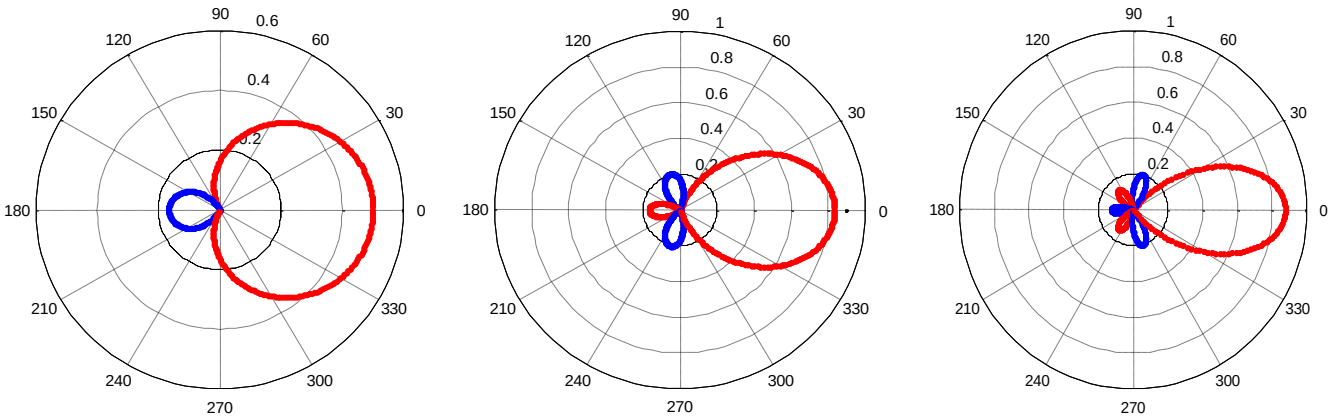


Figure 3.13 left to Right 1st, 2nd and 3rd Order Polar patterns for a source at 0°

The previously described method of 1st order decoding utilises a linear equation and it has been found that to derive a 2nd order decode it is simply a case of adding the 2nd order components. A number of authors have used this method (Bamford and Vanderkooy, 1995), (Martin et al., 1999a), (Southern and Murphy, 2007), (Neukom, 2007). See below in *Equation 3. 9.*

$$LS = W + X\cos(\theta) + Y\sin(\theta) + U\cos 2(\theta) + V\sin 2(\theta) \qquad \text{Equation 3.9}$$

However it has been found that there is a more efficient way of deriving the decoding gains by use of matrix inversion (Hollerweger, 2006).

To use this method firstly a vector $(\overrightarrow{B})$ of the encoded components has to be created using the encoding equations shown previously. Then a matrix of speaker positions is created which is shown below in *Equation 3. 10.*

$$\begin{pmatrix} Y_{00}^{1}(\theta_1,\phi_1) & Y_{00}^{1}(\theta_2,\phi_2)\cdots & Y_{00}^{1}(\theta_j,\phi_j)\cdots & Y_{00}^{1}(\theta_L,\phi_L) \\ Y_{11}^{1}(\theta_1,\phi_1) & Y_{11}^{1}(\theta_2,\phi_2)\cdots & Y_{11}^{1}(\theta_j,\phi_j)\cdots & Y_{11}^{1}(\theta_L,\phi_L) \\ Y_{11}^{-1}(\theta_1,\phi_1) & Y_{11}^{-1}(\theta_2,\phi_2)\cdots & Y_{11}^{-1}(\theta_j,\phi_j)\cdots & Y_{11}^{-1}(\theta_L,\phi_L) \\ \vdots & \vdots & \vdots & \vdots \\ Y_{M0}^{1}(\theta_1,\phi_1) & Y_{M0}^{1}(\theta_2,\phi_2)\cdots & Y_{M0}^{1}(\theta_j,\phi_j)\cdots & Y_{M0}^{1}(\theta_L,\phi_L) \end{pmatrix} \qquad \text{Equation 3.10}$$

Where Y_{00}^{1},Y_{11}^{1},Y_{11}^{-1} refers to the 1st order W, X, and Y component etc. and $(\theta 1, \varphi 1)$ refers to the loudspeaker position in azimuth and elevation.

It is then necessary to pseudo invert this matrix termed the C matrix in order to derive the decoder matrix D.

Finally to derive the speaker gains needed to reproduce the soundfield, the full matrix operation is shown below in *Equation 3.11.*

$$C^{-1} \cdot \vec{B} = D \cdot \vec{B} \qquad \text{Equation 3.11}$$

Several types of decoders are available with each of these decoders having a certain application and also an effect on the perception of the soundfield.

3.5.7 Decoder Types

There are three main types of ambisonic decoder, each has an effect on the perceptual qualities of the reproduction. To decode using one of the three main decoder types requires the final multiplication of the (D) decoder matrix with the diagonal matrix (Γ) of per order correcting gains. *See Equation 3.12* and *Table 3.1* for decoder type correcting gains which are used in the diagonal matrix.

$$D \cdot \Gamma \qquad \text{Equation 3.12}$$

Decoder Type	Ambisonic Order	Correcting Gains
Basic	1 2 3	1 1 1
Max rE	1 2 3	0.577 0.775,0.400 0.861,0.612, 0.305
In Phase	1 2 3	0.333 0.500, 0.100 0.600,0.200, 0.029

Table 3.1 Decoder Type and Correcting Gains for Each Order (Daniel, 2000)

3.5.7.1 Basic Decoder

A basic decoder only satisfies one half of Gerzons metatheory, thereby providing a limited sound-field reconstruction. A number of different names have been given to this type of decoding like mode matching, basic and low frequency solution (Heller et al., 2012). Since the low frequency decoding will be applied across all frequencies, with no separation of low and high frequencies, producing a phasey reproduction, because at high frequencies the wavelengths are smaller than the inter-ear distance (Heller et al., 2008). *See Figure 3.14.*

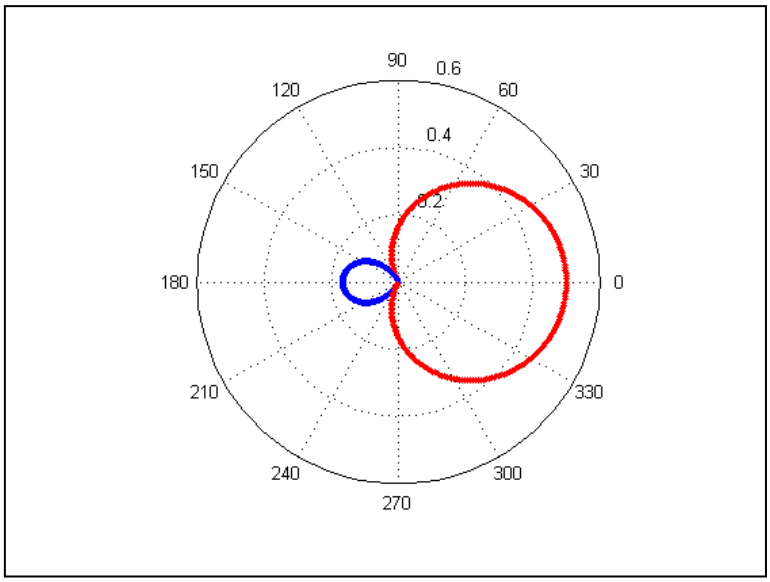


Figure 3.14 Directivity of 1st Order Polar Pattern Using a Basic Decode (blue is negative phase)

3.5.7.2 Maximum Energy

The maximum energy decoder aims at concentrating the energy in the expected direction, therefore reducing the secondary rear lobe, providing a sharper virtual source (Bertet et al., 2006). *Figure 3.15* shows the plot for 1st order max rE decode.

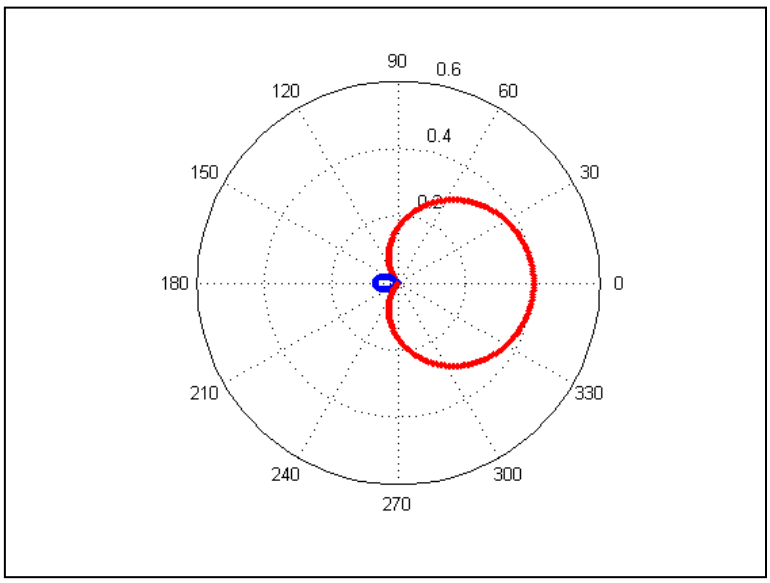


Figure 3.15 Directivity of 1st Order Polar Pattern Using a Max rE Decode (Blue is Negative Phase)

3.5.7.3 In Phase

A further modification to the ambisonic decoder can be implemented by using the in phase decoder. This was first conceptualised by Malham (1992) and was designed to be used for large listening areas like live concerts. The perceptual consequence of this is that the size of the listening area is extended, this works by getting rid of the anti-phase components from diagonally opposite speakers in the array, although at the expense of other factors, like less definition of reproduced sources. *See Figure 3.16.*

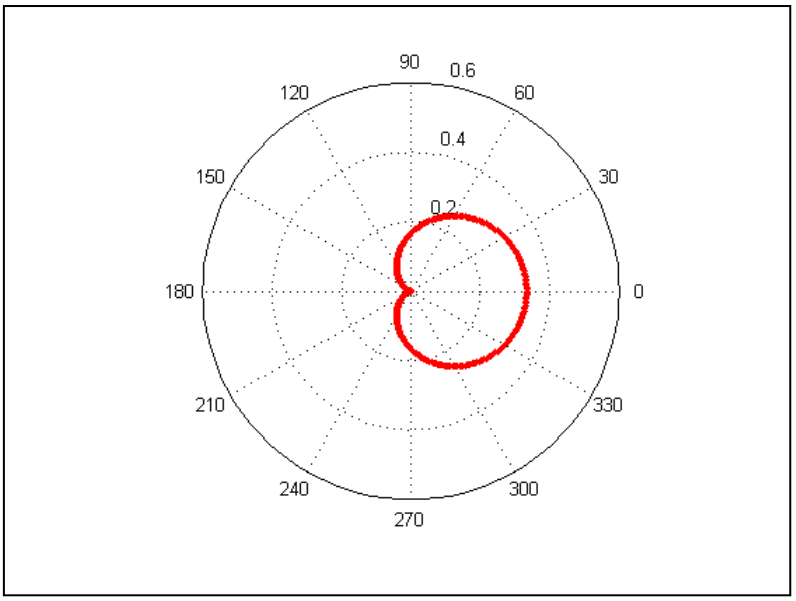


Figure 3.16 Directivity of 1st Order Polar Pattern Using In Phase Decoding

Ambisonic reproduction works on the fact that some speakers in the array will be reproducing out of phase signals which are based on the diametric decoder theorem. And that optimum reproduction will only be created for a centrally located listener. Therefore in a concert situation some listeners will be nearer some speakers than others which will cause a breakdown in imaging because of the precedence effect by using the in phase decoder this gets rid of any out of phase signals negating this problem.

3.5.8 Dual Band Decoder

The previously mentioned basic and maximum energy decoders can be combined. This involves using the basic decoder and the maximum energy decoder and combining the two matrices using a phase matched crossover network similar to that found in loudspeakers creating a dual band decoder (Gerzon and Barton, 1992) with the transition point at 400Hz. This allows the low frequencies and high frequencies to be maximized in line with Gerzons meta-theory for optimal soundfield reproduction, although it is stated by Daniel et al. (1998)

that in anechoic listening spaces a single band solution such as that described in *Section 3.5.7.2* will give the same perceptual results as a dual band decoder. In addition Heller et al. (2008) states that if a dual band decoder is not used it is better to use a single band energy optimised decoder.

3.5.9 Limits to Correct Soundfield Reconstruction

It has been illustrated by Daniel (2009) that there are limits for each order up to which correct frequency reconstruction can be guaranteed.

Order (M)	1	2	3	4
f Lim	700Hz	1300Hz	1900Hz	2500Hz
aE	45°	30°	22.5°	18°

Table 3.2 Frequency limit for correction reproduction

Table 3.2 shows that depending on the ambisonic order used, correct reconstruction of frequencies is only guaranteed up to a frequency limit (f Lim) thereafter this frequency limit angular dispersion (aE) causes a spread of the sound source, the spread of the sound source decreases with an increase in ambisonic order.

3.5.10 Near Field Compensation

Another important aspect of an ambisonic decoder is to apply near field compensation (NFC) which is achieved using a set of filters (Daniel, 2003). This compensates for the finite distance of the speakers which causes a perceived bass boost and is analogous to the bass boost which occurs due to close microphone placement using a directional microphone. The NFC compensates for the curvature of the wavefronts and is a function of the distance of a speaker from the centre of the array and the order of reproduction (Heller et al., 2012).

3.5.11 Speaker Array Configuration

In order for correct reproduction it is necessary that there are enough speakers to reproduce the soundfield. A simple equation covers this *See equation 3.13*, where the M stands for the ambisonic order.

$$N = 2 * (M + 1) \qquad \text{Equation 3.13}$$

The same equation governs the reproduction for height layouts with a slight difference since a greater number of speakers is required, *See Equation 3. 14*.

$$N = (M + 1)^2$$ Equation 3.14

A minimum number of speakers is required for each order to reproduce the soundfield correctly, however it is stated by Gerzon (1983) that the more loudspeakers used the better the reproduction, since this prevents the speaker detent effect where sounds are pulled to the nearest speaker when panning sources. However no maximum number of speakers for a given order has been specified.

As previously mentioned in addition it is necessary to ensure that the speakers in the array are equally spaced, for horizontal layouts this is easily achieved by choosing regular intervals around the unit circle. But for height layouts this becomes problematic above 1st order since it is difficult to dispearse speakers evenly around a sphere for practical and mathematical reasons.

3.5.12 Platonic Solids

A small number of 3D shapes which meet the criteria of ambisonic reproduction exist for 3D systems, these are known as the platonic solids and are shown below in *Figure 3.17.*

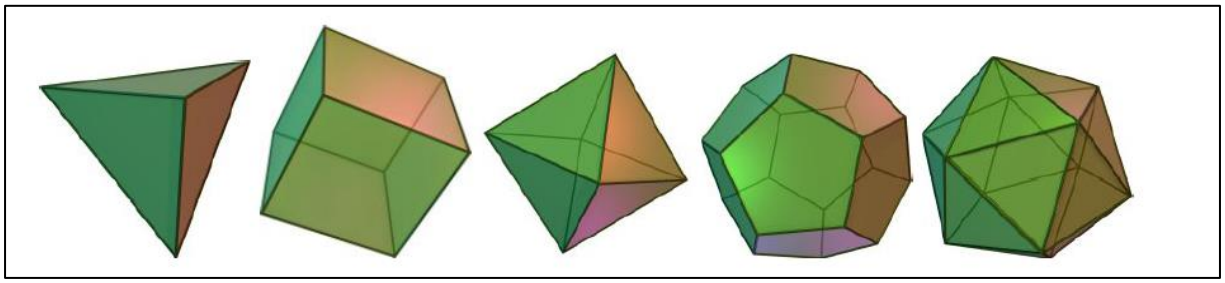


Figure 3.17 The Platonic Solids

Although creating layouts using for example the dodecahedron can be difficult, since speakers have to be placed at the vertices of each of the shapes. Further, some of these layouts, for example the cube, do not provide the ability to have speakers on the horizontal at ear height where human hearing is most accurate (Blauert, 1996).

Therefore the motivation is to look for layouts that are semi regular and can accommodate speakers on the horizontal at the listener's ear height.

3.5.13 Irregular Layouts

The aforementioned prerequisites for the correct reproduction of ambisonics over loudspeakers have a consequence for the current surround standard e.g. 5.1. Therefore compromises have to be made. As previously discussed Gerzon stated several criteria which had to be met in order for correct ambisonic reproduction. However in terms of the 5.1 speaker layout this poses a problem and deriving a decoder solution for this layout is cumbersome, as it requires an iterative process and has been the topic of extensive research (Wiggins, 2004), (Benjamin et al., 2006), (Moore, 2009). Stated in *Section 3.5.12* there are only a small number of solids which provide a regular layout for height systems, as a result a number of decoding schemes for irregular layouts have also been created (Treviño et al., 2011), (Boehm, 2011).

It is also stated by Gerzon himself that it is not possible to satisfy the criteria of energy and velocity vectors for all positions around a listener (Gerzon and Barton, 1992), therefore because human hearing is best to the front it is better to optimise for this section rather than the rear or sides of the listener (Boehm, 2011).

3.5.14 With Height Irregular Layout Solutions

A method used to overcome irregular with height layouts has been created, see (Zotter and Frank, 2012), (Heller et al., 2012). This involves decoding to a virtual array of regular speaker layouts using the so called “T Designs” (Hardin and Sloane, 2014) which are optimal distributions of points on a sphere to derive an optimal decoder matrix and then mapping these signals to the real speaker layouts using VBAP. Since it was explained in *Section 3.4.1* that for real speaker layouts VBAP places less constraints on the physical layout of the speaker array, this alleviates the problem of creating a regular with height layout, which as explained before is difficult to achieve due to practical problems of mounting speakers in strict regular positions above and below a listener.

3.6 Normalisation Schemes

Explained previously in *Section 3.5.6* ambisonics uses different orders which are based on differing numbers of audio signals which make up the resolution of the soundfield. As the orders increase so does the numbers of components or ultimately the audio signals, it is therefore necessary to control the levels across all of these orders or signals to avoid clipping, therefore normalisation schemes are used which scale each ambisonic component accordingly, in order to ensure that all of the levels are consistent and do not peak. Several

normalisation schemes exist for use in ambisonic encoding/ decoding in 2D and 3D formats, these can be summarised into three main types.

3.6.1 FurseMalham (FuMa)

The Furse/Mahlam set utilises normalisation values so that each of the coefficients can only have a maximum of 1.Without this normalisation, as the order increases the maximum value each coefficient can take on increases. Whilst for reverberant or diffuse soundfields this would not be an issue, for point sources this would cause problems in terms of clipping. Therefore the Furse/Mahlam normalisation coefficients are used to prevent this (Malham, 2003). The Furse/Mahlam normalisation also reduces the W component by $\frac{1}{\sqrt{2}}$ this was done in order to reproduce the same levels in the W,X,Y and Z components and to keep the higher order components compatible with the older B format analogue hardware.

Furse/Mahlam set deviates from the mathematical derivation this causes problems concerning the manipulation of the soundfield in terms of tilt and tumbling and the development into higher order formats (Bates, 2009).

3.6.2 N3D

N3D is said to give the best possible use of channels for a natural soundfield and it is stated it is easier to implement and understand (Chapman et al., 2009). N3D also ensures equal power of the encoded components in the case of a perfectly diffuse soundfield, where it is necessary to have an even level across the soundfield (Chapman et al., 2009)

3.6.3 SN3D

This scheme is different to the Furse/Mahlam scheme since the W component has a value of 1 and different normalisation weightings are applied to the higher order components, this ensures that panning of sources using higher orders the levels of the higher orders never exceed 1.*See Equations 3.4, 3.7, and 3.8* previously shown as they use the SN3D normalisation.

A number of other normalisation schemes have been developed each with their own method of operation, however in standardising the ambisonics format termed ambiX (Ambisonics exchangeable) it was proposed to use the SN3D normalisation (Nachbar et al., 2011). *See table 3.3* below for the peculiarities of each normalisation scheme.

Criteria	Furse-Malham	N3D	SN3D
Extensible to higher orders in a straightforward manner. Hence able to meet expanding user's needs.	No	Yes	Yes
Consistent with 1st-order B-format recordings, equipment and software.	Yes	No	No
Consistent with higher-order horizontal-only FMH format recordings, equipment and software.	Yes	No	No
Consistent with the Ambisonics portion of MPEG-4	Yes	No	Yes
Consistent with FMH format recordings, equipment and software having periphonic order P greater than 1.	Yes	No	No
Components of the same degree have the same 3D gains. So that typical processing (such as rotation) can be coded reasonably cleanly.	No	Yes	Yes
The Mathematical Formulation is Consistent across degrees. No special cases	No	Yes	Yes
The harmonics are orthonormal, which eases such things as beamforming.	No	Yes	No
The scheme benefits from direct support in general maths libraries	No	Yes	Yes
The scheme has mathematical grounding in the 3D world.	No	Yes	Yes
Higher-degree components have headroom for when the sound field is not spherically isotropic. Means that 1st order monitoring can be used over higher orders (with more accuracy).	Yes	No	Yes
Greatest dynamic range when the major sources are in/near the horizontal plane	No	No	No
Greatest dynamic range when the major sources are distributed over the sphere	No	Yes	No

Table 3.3 Different normalisation Schemes comparisons (Bates, 2009)

3.7 Summary

A detailed overview of two rendering methods used to create 3D sound fields over loudspeakers has been given. Both methods are classed as amplitude panning methods, however differ slightly, since VBAP is solely a panning method and ambisonics is a complete system allowing the recording of sound scenes using one microphone, then the subsequent reproduction and transmission of 3D sound scenes. Further strengths of ambisonics is that with the use of the B format signal, which can be used to decode to any loudspeaker layout, therefore allowing the recording of sound scenes without having prior knowledge of the playback configuration. Ambisonics also allows the tailoring of different types of decodes for

a number of applications and also the ability to reproduce the sound scene using different orders which have an effect on the resolution of the soundfield.

3.8 Human Hearing Mechanisms in Spatial Audio

Considering stereo listening there are two main mechanisms used to determine the position of a sound source.

The duplex theory in human sound localization is well known and was first discovered by Rayleigh (1875). There are two main cues which are used to determine the direction of a source, the inter-aural time difference (ITD) and the inter-aural level difference (ILD).

The ITD is said to occur due to a time difference between the sound arriving at the ear nearer the source (ipsilateral) to the other ear further away (contralateral) with the ILD occurring due to the different intensity levels occurring at each ear due to the effect of shadowing that the head has on the sound at high frequencies (Howard and Angus, 2006).

There is also a cut off frequency for each mechanism as Howard and Angus (2006) has reported are 700hz and below for the ITD and about 2.8khz and upward for the ILD and between these frequencies our ears are not as good at detecting direction. It is also stated by Long (2005) that in the vertical plane the ability of the brain to interpret time delays is much weaker, it is also reported that there is a greater tolerance for wider speaker layouts when the speakers are elevated compared to stereo in the horizontal plane.

When considering listening to a stereo system it is only possible to place sources in front of a listener. It is stated by Blauert (1996) and Carlile (1996) that human hearing is most precise for sounds directly in front in the horizontal plane and the smallest detectable change in azimuth, called the minimum audible angle (MAA) for a sound source is 1°, however for sources in the vertical direction the MAA is 4-5° (Strybel and Fujimoto, 2000).

Considering horizontal only surround systems like the ITU 5.1, it is possible to place sources in a number of directions, however as sources are moved further from the front the MAA increases, in addition the accuracy of which the sound source will be localised will also depend on its spectral content (Yost, 1994). In 2D surround systems it is possible for sounds to arrive at the listener from a number of different directions, not just the front, therefore at high frequencies the sound's will be modified due to the folds and ridges of the pinna creating constructive and destructive interference and also the effects of the torso of the listener. These are classed as head related transfer functions (HRTF) (Rumsey, 2005). This is

also reinforced by Blauert (1996) who states that the direction to which a sound source arrives at the ear causes linear distortions, which are specific to the direction of incidence and distance. Further, it is explained by Kendall (2010) that front back perception also relies on the acoustics of the pinna which face forward and provide attenuation to sounds arriving from the rear.

It is clear that there are a number of cues for localising a source in 2D surround systems, however considering 3D surround systems that can place sources in the vertical plane, it is necessary to investigate how accurately the human hearing system interprets sources in the vertical plane.

3.8.1 Median Plane

There are two main anatomical sections to consider when looking at localisation with height, the first being the median plane. The median plane divides the head into two sections. *See Figure 3.18.*

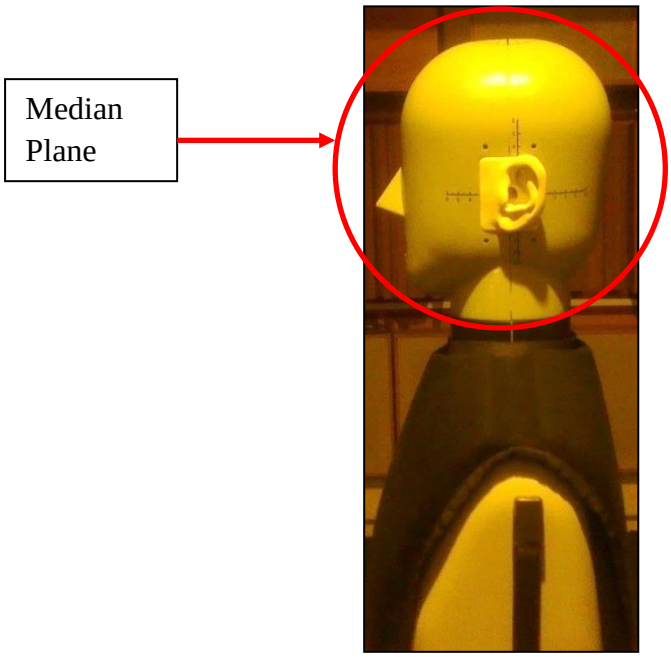


Figure 3.18 Median Plane

Sound sources which are placed in the median plane will reproduce an identical ITD and ILD cues and as a result are considered monaural cues (Warren, 2008). This has a consequence when considering sources which appear in the median plane but with an elevation, for example in 3D surround systems. Since it has been explained by Blauert (1996) that the localisation blur is in the order of 9° in the median plane for a speech source by a familiar person, in addition it is also stated that localisation in the median plane also relies heavily on

the spectral content of the source, this was demonstrated by Roffler and Butler (1968) who found that the sounds pitch dominated where the listener localised the sound source, with sound sources assuming a position based solely on its spectral content, furthermore, this is also reported by (Blauert, 1996) in which he describes 'directional bands' where narrow band sources assumes a position based on its centre frequency.

3.8.2 Frontal Plane

The frontal plane divides the head in two sections at 90° from the front *See Figure 3.19* and for sources located with an elevation on the frontal plane there are more cues available compared with the median plane, like the ITD and ILD cues.

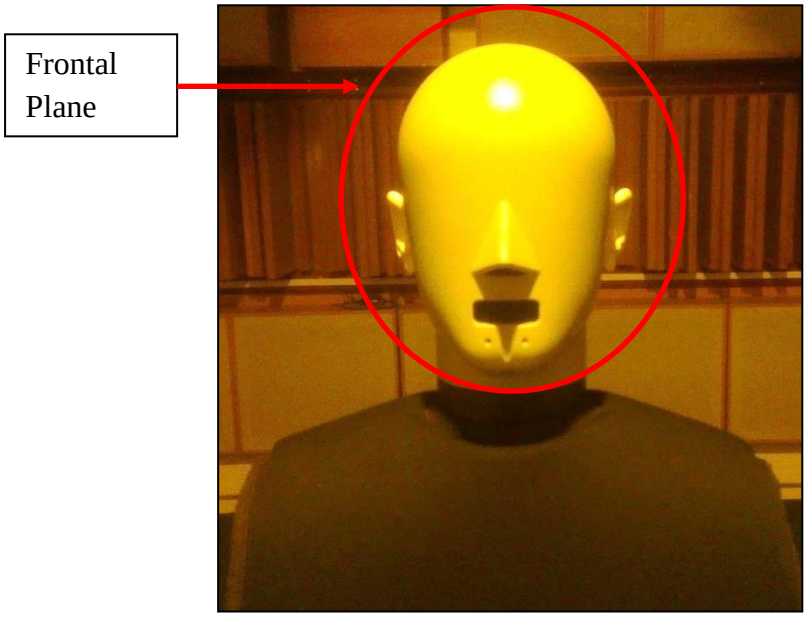


Figure 3.19 Frontal Plane

One of the main cues available for elevated sources as previously explained are the HRTFs caused by the folds and ridges of the pinna, which is true for high frequency sources. Since at higher frequencies the wavelength of the sounds are closer to the size of the ridges and folds of the pinna causing greater attenuation or boost and for complex noise sources a larger number of higher frequencies will be attenuated providing the listener with a greater number of cues. This is also dependent on the direction of arrival of the sound source (Yost, 1994). Further it is explained by Popper and Fay (2005) that for elevation localisation there is a greater dependancy on the source spectral content and bandwidth, this is reinforced by Butler and Humanski (1992) and Ashby et al. (2013) who found that localisation of high passed noise in the vertical plane surpassed that of low passed noise.

3.8.3 Cone of Confusion

Peculiar to both the horizontal and vertical localisation is the cone of confusion, so called because any sound which lies on the cone will produce an identical interaural difference (Yost, 1994). *See Figure 3.20* where the positions of the blue dots show where equal interaural differences will be perceived.

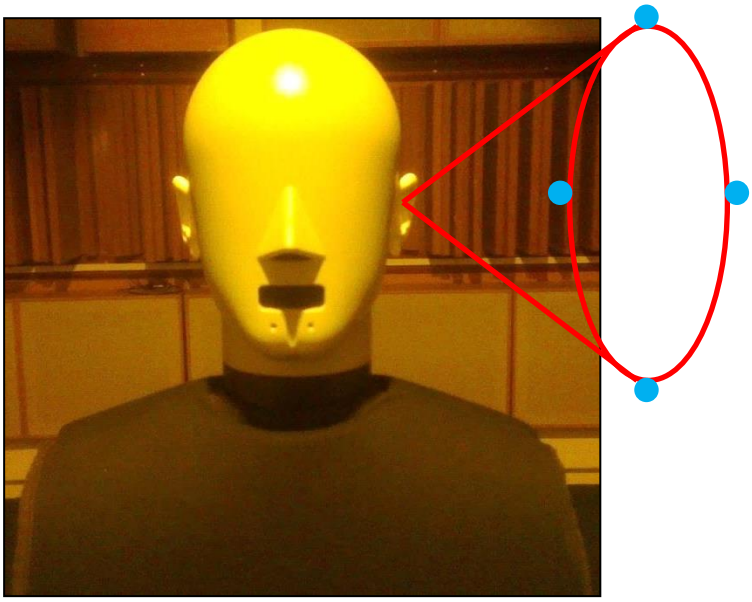


Figure 3.20 Cone of Confusion

To resolve sources placed on the cone of confusion head rotation is used which asissts in locating the sound source position, ultimatley changing the ITD or ILD of the source reducing front/back confusions (Wallach, 1939), (Thurlow and Runge, 1967), (Blauert, 1996), (Ashby et al., 2013). In addition it has also been noted by Wenzel et al. (1993) using non idividualised HRTF data to place sources in the vertical plane, that it is also possible for a listener to experience up-to-down and down-to-up reversals, due to the degradation of high frequency cues in the sound source.

3.8.4 Precedence Effect

Listening environments like concert halls will cause the sound to be reflected off walls, floors and ceilings causing a number of reflections to be generated. This means that the direct sound which reaches the listener first will then be followed by reflected versions at different times and directions, therefore the ear/brain mechanism has to have some way of sorting this information out, this mechanism is known as the precedence effect (Wallach et al., 1949).

The listener hears the direct sound first in a reverberant room followed by reflections, if the reflections arrive within a 30 msec window they will be fused with the direct sound and not heard seperatley, however if after hearing the direct sound the reflections arrive after the 30msec window the reflection will be heard as separate sound or echo (Howard and Angus, 2006). One form of the precedence effect was identified by Haas (1972) who was investigating the effect of a single echo on the perception of speech and found that the delayed sound had to be substantially louder than the direct sound before it was perceived as the same loudness.

The aformentioned factors also come into play when in enclosed spaces since the time of arrival of the first reflection after the direct sound allows the listener to make judgements about the size of the space. Concert halls with a so called short time delay gap being perceived as what Beranek termed intimate within the 20msecs time window (Beranek, 2004).

It has also been hypothesized that the precedence effect is related to spatial impression, (spatial impression will be explained further in *Section 3.9.1.)* Experiments were carried out by Morimoto (2002) which illustrated that components which arrive within the upper limit of the precedence effect add to the apparent source width (ASW) whilst those that arrive later than the upper limit add to the perception of envelopment.

3.8.5 Masking

This phenomenon can be experienced in a music mix where a louder instrument can mask a quieter instrument, until the louder instrument is turned down or the quieter instrument is turned up louder, frequency components can also mask each other not just whole instruments. The previous description of masking where the louder sound or different frequency component is called the masker and the softer sound is called the maskee (Zwicker and Fastl, 1999).

3.8.5.1 Basilar Membrane

The basilar membrane of the inner ear allows humans to carry out frequency analysis of sounds. The basilar membrane is wide and thick at one end becoming thinner and narrow at the opposite end, this allows maximum stimulation at the thin end for high frequencies and maximum stimulation for low frequencies at the thick end, with different frequencies in between stimulating different places along its length (Zwicker and Fastl, 1999).

The structure of the basilar membrane gives rise to masking, since it is explained by Zwicker and Fastl (1999) that a loud response on one place of the basilar membrane will mask softer sounds in the critical band around it, unless the stimulation of another tone breaks the masking threshold, similar to the previous description of sounds in music mix

Temporal masking can also occur with tones sounded close in time, for example a masking signal can stop before the maskee starts with this being known as forward masking, in addition for short impulsive signals the phenomenon of pre-masking can also happen this is where the maskee stops before the masker starts (Zwicker and Fastl, 1999).

It is stated by Yost (1994) that there is no clear explanation for the effects of forward and backward masking, however a theory is provided by Pohlmann (2005) in that the brain could be integrating sound events over time and processing the information in bursts or that the brain processes loud sounds faster than soft sounds.

3.8.5.2 Cocktail Party Effect

The term cocktail party effect was first coined by researcher Colin Cherry, who was researching the problem of the ability to select one voice of a talker in a room where there are many concurrent talkers (Cherry, 1953). Cherry Carried out experiments over headphones, in his experiments he found that it was much easier to distinguish what concurrent talkers were saying when one voice was fed to one ear and another to the other ear. This was an important discovery, since this showed that when the spatial position of concurrent talkers was different this helped in interpreting what was being said. However, he also found that differences in pitches of voice, difference in speed of concurrent talkers and differences in accents also helped.

The cocktail party effect allows humans to separate interfering noises or sounds known as maskers because of differences in pitch, differences in amplitude or most relevant in spatial audio because of different spatial location. An example of this being given in *Section 3.2.5* where it is explained that the move from stereo to 5.1 provides better source separation leading to lower masking thresholds.

It could be that the move from surround in the horizontal plane to include also the height dimension will further enhance source separation and provide lower masking thresholds.

3.8.6 Summary

From the previous theory of human hearing in the horizontal plane and the vertical plane it would seem that there are less cues available to the listener for sources in the vertical plane compared to the horizontal plane. Further, it would also seem to be dependent on the spectral content of the sound source, since it has been reported that sound sources which are spectrally rich in high frequencies being easier to locate than lowpassed sounds due to the increase in head, torso and pinna influence. In addition, the number of auditory cues available to a listener to localise elevated sound sources will depend on its direction of arrival, since sounds in the vertical median plane witll produce equal interaural differences therefore will only rely on spectral differences, whereas sounds arriving from the frontal plane will also utilise in addition to spectral cues, ITD and ILD cues.

The theory of localisation has also been provided for environments which generate multiple reflections, where it was detailed that the ear/brain mechanism uses the time of arrival when attending to sound sources based on the direct and reflected sound.

A brief summary has also been given on the theory of masking of sounds and also the cocktail party effect. This phenomenon provides listeners with the ability to make sense of concurrent sounds when the sounds come form different positions in the space around a listener.

The previous details encompass the mechanisms a listener would use to localise a sound source, or the abilty to make sense of sound scenes in which sources are spatially spread out as opposed to coming from the same direction. Although, the goal of spatial audio reproduction is not soley to reproduce point sources.

3.9 Assessment of Spatial Audio Quality

It is stated by Letowski (1989) that sound quality can be divided into two separate concepts these are timbral and spatial. It is also explained by Rumsey (2002) that different reproduction systems can produce varying degrees of timbral and spatial qualities.

In assessing 2D spatial audio reproduction systems a number of attributes have been investigated for example the ability of a system to reproduce accurately a virtual source in terms of localisation (Liebetrau et al., 2007), (Ashby et al., 2013), (Bertet et al., 2013). Although as stated by Rumsey (1998a) localisation may be impotant but it is not the only

factor of importance in the design of reproduction systems. It has also been found by Berg and Rumsey (2000), Zacharov and Koivuniemi (2001) that precise localisation of sources has weak correlation with subjective preference.

Other researchers have focussed on the timbral qualities of multichannel reproduction (Zielinski et al., 2003), (Zielinski et al., 2005) in assessing the effects of bandwidth limitations on timbral aspects of the multichannel reproduction. It was found that band-limitation of the two front channels left and right of the 5.1 system by lowpass filtering to remove the high frequency content gave rise to significant perceivable deteriation of audio quality, yet band-limitation of the centre channel or surround channels was perceptable but not annoying. The authors hypothesize that the effects observed could be as a result of masking or perceptual streaming and suggest that for some types of material it might be possible to limit the bandwidth of the centre or surround channels without a significant perceivable deteriation of audio quality, it was also noted by the authors that the downmix algorithms used affected not only timbral quality but also spatial quality.

Early experiments by Nakayama et al. (1971) where mulitchannel recordings of an orchestra in a concert hall were reproduced using one through to eight channel speaker arrays. Found that multichannel sound can be charaterised by three attributes, fullness, clearness and depth of image sources. Further it was found that clearness was related to the objective measure deutlichkeit[1] and that the fullness of the sound was related to IACC and that this was due to low interaural coherence.

A large amount of research has continued on from Nakyama and evaluated the ability of surround systems to reproduce the perception of spatial impression (Morimoto, 1997), (Bech, 1998), (Berg and Rumsey, 2002), (Hiyama et al., 2002), (Hamasaki and Hiyama, 2003), (Hirst, 2006), (Kamekawa and Marui, 2010). Whilst the previous authors have assessed spatial impression as a whole, others have been more specific and focused on the perception of envelopment (Soulodre et al., 2002), (Soulodre et al., 2003), (George et al., 2010).

It has been explained by Morimoto and M. Maekawa (1989) that spatial impression consists of two attributes, one being apparent source width, which is described as the width of the

[1] Deutlichkeit (distinctiveness) is measured from the impulse response of a room from source to receiver and provides a measure for the clarity of speech or music. BARRON, M. 2009. *Auditorium acoustics and architectural design,* Oxon, Routledge.

sound stage fused with the direct sound and envelopment is defined as the degree of fullness of sound image surrounding the listener. It is also explained that listeners are able to distiguish between the perception of apparent source width and the perception of envelopment.

Other studies have used methods of subjective evaluation which are more exploratory in nature, since listeners are urged to provide descriptors based on the playback of different reproductions, these methods being called elicitation. The value of these methods are that the listeners are not given the experimenters derived atributes to scale, instead agree on a common set of attributes through a series of meetings. Such methods have been used by Zacharov and Koivuniemi (2001) in assessing a number of spatial audio systems, which have included 5.1, stereo and a system with height, several prominent attributes have been derived which were subsequently used to score the different systems. Some of the attributes found were sense of direction, width, broadness, distance, naturalness, tone colour and emphasis.

Similar methods have been used by Guastavino and Katz (2004) when assessing surround systems with height and horizontal only systems. The attributes found in Guastavino and Katz study were coloration, presence, distance, localization and readability.

Comparable attributes were found by Berg and Rumsey (2006) utilising elicitation techniques in evaluating stereo and surround systems. From the elicitations a number of descriptors have been generated which were localisation, depth/distance, width, room perception and envelopment.

Apart from the ability to localise virtual sound sources accurately it would seem that another highly rated percept is being totally surrounded by sound, or the fullness of sound scene which could be assumed as reffering to envelopment. It is described by George et al. (2010) that the sub attribute of spatial impression (envelopment) is one of the main attributes that seperates multichannel systems from mono and that it is one of the main attributes driving multichannel development. It would therefore be necessary to investigate if the reproduction of surround with height enhances the perception of envelopment. From the previous studies there are similar attributes which appear in each of the investigations, for example localisation, width/broadness and envelopment. It would seem that these are important

attributes, since these are the most pertinent and lead to listeners picking them out in the previously described free elicitation experiments. *See Table 3.4* below which summarises a number of attributes derived from descriptive analysis using multichannel spatial audio systems.

Author	Attribute	Description
Koivuniemi & Zacharov, 2001	Sense of direction	How easily the locations of events can be discriminated. Also measures whether several sound sources can be distinguished, a negative value means that the source is ill defined or enveloping
	sense of depth	Describes how strongly the sensation of distance is. Describes whether several sound events can be discriminated in terms of distance. A negative value could mean that the distances for all events are ambiguous except those originating from the transducers position
	sense of space	Scales how well the space where the recording was made is perceived. A positive value could mean a strong sensation of being in a certain kind of environment
	broadness	Describes how wide an area the perceived sound events seem to have. A strong positive value would mean that the sounds are coming from all around the listener i.e envelop the listener
	Naturalness	Describes how well the perceived events conform to what the subjects consider realism. Perception of something that isn't possible in reality yields a negative value
Berg & Rumsey,2002	Source envelopment	For the listener to be surrounded by the sound source (the instrument/voice)
	Ensemble width	To experience that the sound sources are dispersed in space as an opposite of being positioned together
	Room Envelopment	Refers to the extent the sound coming from the sound sources reflections in the room (the reverberation) envelopes/surrounds/exists around the listener
Choisel & Wiklemaier,2006	width	Imagine an area occupied by the sound sources (e.g the instruments) for every pair of sound's you should indicate for which of the sound's this area is wider
	Elevation	Some sounds might appear to be positioned at the same level as your ears . Some others might be lower (closer to the floor) or higher (towards the ceiling). Indicate which two sounds you perceive as being higher in space
	Spacious Enveloping	A sound is said to be spacious when you have a good impression of the space in which it was played. Try to imagine this space it can be a small room for example or a large hall.

		A sound is enveloping when it wraps around you. A very enveloping sound will give you the impression of being immersed in it, while a non-enveloping one will give you the impression of being outside it.
	Natural	A sound is natural if it gives you a realistic impression as opposed to sounding artificial
Martens & Sungyoung, 2007	Wide-Narrow	The apparent horizontal spatial extent of the sound source
	Focussed-Diffuse	The apparent integration of the sound source into a single unified image
	Distant-Close	The apparent spatial distance of the sound source from the listening position

Table 3.4 Attributes Derived by Other Researchers Using Descriptive Analysis

3.9.1 Spatial Impression in Concert Hall Acoustics

In a concert hall there are a number of factors which recreate a highly rated listening experience. One of these is the attribute spatial impression, which was first realised by Barron (1971) and is related to a broadening of the frontal sound scene caused by early reflections between the time limits of 0-80ms from the lateral direction called apparent source width (ASW). However, following this listener envelopment was realised by Beranek et al. (1996) and was related to later arriving lateral reflections which envelopes the listener between the time limits of 80ms-∞ms.

3.10 Purpose of Spatial Audio Systems

The extension of the stereo system to 5.1 includes the extra centre and surround channels which provides a number of benefits and were explained in *Section 3.2.5*. In summary these were that the centre channel allowed anchoring of the centre stereo soundfield for off centre listeners, including better tonal balance. The extra rear channels provide two uses, one for the creation of rear sound effects and two for lateral energy, with lateral energy being improtant for envelopment. Finally, it was also stated in *Section 3.2.5* that the 5.1 system is the optimum layout and number of speakers for reproducing the spatial impression of a diffuse soundfield in the horizontal plane and is closely related to listeners envelopment (Hiyama et al., 2002).

The term envelopment or more specifically listener envelopment was first coined in relation to concert hall acoustics, where the perception of envelopment is judged highest when late reverberant sound is arriving equally from all directions (Beranek et al., 1996).

In multichannel systems envelopment can be recreated in two different ways, the first can be by reproducing reverberant sound which has been recorded with appropriate techniques in a real space, or produced using digital signal processing or by using dry direct sources placed around a listener which exhibit special qualities like, soundfield density (number of mono sources), interchannel correlation, location or position of the sources and frequency content (Conetta, 2011). Examples of dry direct sources might be random noise signals which exhibit low correlations like that used by Hiyama et al. (2002) to approximate a diffuse field.

The distinction between the direct and indirect envelopment is made clear by Rumsey (2002) where he explains that in reproduced sound subjects described being enveloped when surrounded by dry direct sources, however in terms of concert hall acoustics this is never the case since envelopment is a property of late reflected sound, however if the listener were placed in the middle of the orchestra they would claim to be enveloped or inside the music.

The perception of envelopment in multichannel reproductions has been explained in that two different types of envelopment can be expereinced one of which has a different context to that experienced in concert hall acoustics.

The clear benefits of the extra channels in a 5.1 system are established, in the fact that the rear channels can provide the sensation of envelopment such as that experienced in a concert hall environment and is a highly rated attribute (George et al., 2010).

3.10.1 Objective Measures

A measure which has been shown to be closely related to ASW and envelopment is called IACC (Inter-aural Cross Correlation) (Nakayama et al., 1971), (Hiyama et al., 2002). This measurement has been used extensively in concert hall acoustics and in reproduced sound (Mason, 2002), (Hirst et al., 2006).

IACC is defined as the maximum value of the normalised cross correlation function for the time interval of ± 1ms. *See Equation 3.15* and *3.16* where Pl and Pr are the impulse responses at the left and right ears, τ is the offset between ±1ms and is set to account for the maximum

inter-aural time difference. t1 and t2 are the time limits of the integration, different variants of IACC exist depending on the time limits. The version used in this work is IACC for which t1 would be 0 and t2 = inf. This IACC is based on the degree of correlation between the impulse responses at the left and right entrance of the ear. A well-defined source has a sharp maximum of 1, whilst subjective diffuseness has a low value ~<0.15 (Ando, 1998)

$$IACF_{t1t2}(\tau) = \frac{\left[\int_{t1}^{t2} Pl(t) \cdot Pr(t+\tau)\, dt\right]}{\left[\int_{t1}^{t2} Pl^2(t)\, dt \int_{t1}^{t2} Pr^2(t)\, dt\right]^{\frac{1}{2}}}$$ Equation 3.15

$$IACC_{t1t2} = max\,|IACF_{t1t2}(\tau)|, for -1ms < \tau < +1ms$$ Equation 3.16

A large amount of research has been carried out into IACC and the measurement is not a perfect solution since it does not take into account the effect of low frequencies due to the wavelength of the signal and the distance between the ears (Tohyama and Suzuki, 1989). As a consequence, it has been explained by Blauert and Lindemann (1986) that in terms of concert hall acoustics early lateral reflections which contain mainly frequencies up to 3 kHz contribute solely to the expansion in the front back direction, referring to the attribute envelopment.

3.10.2 Importance of the Direction of Reflections in Concert Hall Acoustics in Terms of Subjective Preference

It has been explained in *Section 3.10* that lateral reflections are the most important, therefore what part do ceiling reflections have? It has been reported by Marshall (1967) that diffusion of ceiling reflections are advantageous since this lowers the masking level which provides audibilty to lateral reflections which are important for what he terms spatial responsivness, which could be assumed to be reffering to spatial impression.

In the same way it is also reported by Furuya et al. (2001) that if reflections arriving from directions other than lateral are excluded in terms of envelopment, this could be detrimental,

in that late sound from other directions like overhead are also effective for experiencing the three dimensional fullness of sound, this was also confirmed by Wakuda et al. (2003).

However,it is stated by Barron (2009) that ceiling reflections will not add to spaciousness, this sentiment is also echoed by Griesinger (1996) who states that it is the lateral or diffuse components which matter, reveberation from the front or overhead is not perceived as spacious.

The previous statements underline the fact that lateral reflections in the horizontal plane are important in concert hall acoustics and there is evidence that reflections from the vertical plane in concert halls do not add to envelopment/spaciousness, however the direction of arrival of reflections from the vertical plane are still important. Since it has been explained that reflections from the vertical plane ariving in the median plane are not useful in lowering the IACC, because reflections arriving at the listener from the median plane will lead to identical signals at each ear increasing the IACC (Schroeder, 1979), therefore it would seem that in order to reduce the IACC the reflections from the vertical plane should arrive at the listener from non central directions similar to the horizontal case.

3.10.3 Clarity

There is some uncertainty regarding the importance of ceiling reflections in concert hall acoustics, although there is some evidence that ceiling reflections add to the clarity of sound for listeners. This is stated by Barron (2009), however care has to be taken with reflections from overhead since it has been established that these reflections can cause tone colouration effects. In the same way, it has been reported by Silzle et al. (2004) for elevated speakers that the comb filtering effects of higher frequencies are much more audible in the vertical plane than the horizontal plane, this due to the binaural processing of the two ear signals in the horizontal plane which makes it much less audible compared to the vertical plane. The theory being that sources arriving from the vertical plane are subject to unmasking making comb filtering more audible.

It is also stated by Lokki and Pätynen (2011) that early lateral reflections should reach the listener from slightly elevated angles as this helps to optimize high frequency content and harmonics. This theory is supported by Hartmann (1983) who found that localisation accuracy was affected by the direction of which early reflections appeared and that fewer localisation errors were noted when early reflections came from above and below instead of

lateral directions. Hartmann also makes the distinction between lateral reflections for a sense of surround, but not for good localisation of sources in the horizontal plane.

3.11 Reproduced Sound with Height

It was illustrated in *Section 3.8* that there are two main mechanisms used to localise sources in the horizontal plane these being the ITD and ILD, further to this there is an additional cue which is used for source's in the vertical plane which utilises spectral differences to identify the sound sources position. It has also been illustrated that depending on the sound sources position in the vertical plane will dictate the cues available to localise the source, this will have an impact on the reproduction of elevated virtual sources using 3D speaker arrays. For example it was also illustrated in *Section 3.8.1* that for virtual sources in the median plane there are no interaural differences therefore the major cue used in this position is solely the spectral differences of the high frequency components of the source. Therefore in terms of the quality of localisation in this position will be dependent on the frequency content of the virtual source signal used. In addition it was also discussed in *Section 3.5.6* that ambisonic reproduction has different orders of encoding and decoding which ultimately effects the resolution of the source therefore depending on the ambisonic order used will have a further effect on the localisation accuracy.

It was noted earlier in *Section 3.10* that the 5.1 system is able to provide a sense of envelopment in a similar manner to that encountered in concert hall acoustics. Looking towards the development of current surround systems to include height it is necessary to explore how the height dimension in real listening spaces adds to the experience. An account of research in concert hall acoustics in terms of reflections from the vertical plane has been given and there are some convincing arguments that reflections from positions other than lateral do not add to envelopment, but at the same time can enhance the timbral properties of sound for example clarity by reducing spatial masking, although in the same way can cause tone colouration.

3.12 Summary

3.12.1 Current Reproduction Systems

This chapter has started with an account of the development of reproduction systems, from mono leading to surround systems and the benefits of the move from mono to stereo and then to 5.1. Further new emerging systems have also been detailed which include height channels

and a number of reasons have been provided as to the perceptual benefits the addition of height channels bring. A small amount of research has been carried out on these new systems but the results are inconclusive and do not provide a clear picture as to the benefits of the height channels.

3.12.2 Human Hearing

It is also necessary to understand the mechanisms used in human hearing. It has been illustrated in *Section 3.8* that human hearing in the horizontal plane is more accurate than the vertical plane. In addition it has also been detailed that the cues available to a listener when sounds are presented in the vertical plane depend firstly on the elevated position, whether it is in the frontal plane or the median plane, since sources in the frontal plane in addition to ITD and ILD cues will also provide spectral cues, but at the same time this will be dependent on the frequency content of the sound source. Contrary to this sources in the median plane will rely solely on the spectral content of the source and the possibility that the listener can move their head.

These points will have an impact when considering creating content for replay over with height systems.

3.12.3 Masking

The theory of masking has also been discussed, this illustrates the problem of competing sounds coming from the same direction which will not be perceived separately for a number of reasons. It was also detailed in *Section 3.8.5* that sound sources could be more clearly perceived if they were more separated in space for example.

An example was given of the current 5.1 system whereby it has been discussed that lower masking thresholds have been attained over the current stereo configuration. Therefore it is not unreasonable to expect that 3D surround systems will be able to achieve even further sound source separation with the addition of height channels allowing further separation of sources.

3.12.4 Importance of the Direction of Reflections

Finally, It has been illustrated that the current stereo system was developed into 5.1 to include rear channels, which provide two uses the reproduction of lateral energy important for envelopment and rear sound effects *See Section 3.2.5.* Looking towards the development of horizontal surround systems to include height channels it remains to be established whether height channels enhance envelopment, since there is some conflicting information as to

whether ceiling reflections in concert halls add to (spaciousness) or envelopment. Although, previous research has shown that ceiling reflections in concert hall acoustics do add to clarity, since it is explained that ceiling reflections lower the masking threshold allowing greater audibility to lateral reflections.

3.12.5 Rendering Methods for 3D Surround

Two different rendering methods which facilitate the creation of 3D sound scenes have also been detailed which were VBAP and ambisonics. It was explained that VBAP allows the creation of 3D sound scenes using triplets of speakers arranged in triangles. However the drawback of VBAP is the fact that it is only a panning method and sound scenes have to be created by panning sources into the desired location. Ambisonics is different from VBAP in the fact that ambisonics can capture real spaces in 3D using the soundfield microphone or the newly developed Eigenmike. Furthermore ambisonics has the ability to pan sources using different resolutions of decoding which could provide a benefit when creating virtual sources in the vertical plane.

3.13 Conclusion

A number of areas have been identified in relation to the development of surround systems from the current systems in the horizontal plane to surround systems which include the vertical plane. These include newly developed 3D surround systems and the different layout configurations, possible methods for rendering 3D surround and the benefits of each method. Possible benefits of reproducing 3D surround are also explored which include greater source separation, enhancement of spatial impression/envelopment, possibility of panning sources off the horizontal plane, however it was detailed that human hearing in the vertical is not as accurate as in the horizontal plane and that there were a number of other factors that can cause inaccurate localisation due to the number of cues available to the listener, since sources in the frontal plane will provide the listener with a greater number of cues compared to sources in the median plane which will rely solely on spectral differences and head related transfer functions to localise the virtual source. It was detailed that ambisonics provides different resolutions of reproduction which may be an advantage when panning sources in the vertical plane as the greater resolution offered by higher order ambisonics may provide more accurate panning, especially for sources in the median plane which will be investigated in the next chapter.

4 Experiment 1 Localisation in the Median Plane

4.1 Introduction

In this chapter an experiment is detailed which evaluates the accuracy of virtual sources panned in the median plane using a 3D surround system. In music production or sound design it may be desirable to place sources at an elevation directly in front of a listener, however it has been detailed in *Chapter 3 Section 3.8* that there are less cues available to a listener when localising virtual sources in the median plane and that sources in the median plane would assume a position depending on its spectral content, independent of its intended position.

It has also been outlined in the literature review there has been some work carried out into the localisation of virtual sources in the vertical plane and more specifically in the median plane using 1st and 2nd order ambisonics. It was found by Pieleanu (2004) that there was no significant difference in terms of localisation accuracy between 1st and 2nd order ambisonics for virtual sources in the median plane, however extending the ambisonic order beyond 2nd order may provide improvements.

Described in *Chapter 3 Section 3.5.6* ambisonics utilises different orders of reproduction which in turn has an effect on the resolution of the sound source, therefore does the increase in resolution offered by higher order ambisonics help localising elevated sources in the median plane?

4.2 Median Plane

In order to investigate if the use of higher order ambisonics improves the positioning of virtual sources in the median plane a subjective test was carried out. The test utilised the previously described ambisonic reproduction to render sources at various elevations in the median plane.

4.3 Ambisonic Encoding/ Decoding

In order to encode the virtual source positions the encoding equations shown in Chapter 3 *Equations 3.4, 3.7* and *3.8* were used to encode pink noise in 1st, 2nd and 3rd order. The highest ambisonic order that will be used is 3rd order, it was shown in Chapter 3 *Equation 3.14* that to reproduce 3rd order at least 16 speakers will be needed. A speaker array configuration which has been used before by Capra et al. (2007) and Churnside et al. (2011) was utilised since this provided the best compromise in terms of creating a regular array of

loudspeakers and to keep the decoding fairly straightforward. A theoretical analysis of the aforementioned array has been carried out by Gorzel et al. (2014) using non optimised 3rd order max rE decoding and the array was found to give good energy distribution.

The ambisonic decoding used was the max rE decoding, using a one band solution. The Matlab script used to achieve this can be seen in *Appendix 1*, as previously described max rE decoding focuses the energy in the intended direction and provides the narrowest polar pattern and reduces the size of the negative rear lobe. It is also explained by Daniel et al. (1998) that in ideal conditions like anechoic spaces single band solutions will provide the same reproduction as a dual band decoder using shelf filtering, in addition Heller et al. (2008) states that if a dual band decoder is not used it is better to use a single band energy optimised decoder. The magnitude and uniformity of rE for each of the soundfields in 1st, 2nd and 3rd order reproduced over the loudspeaker array can be seen below in *Figures 4.1, 4.2 and 4.3*.

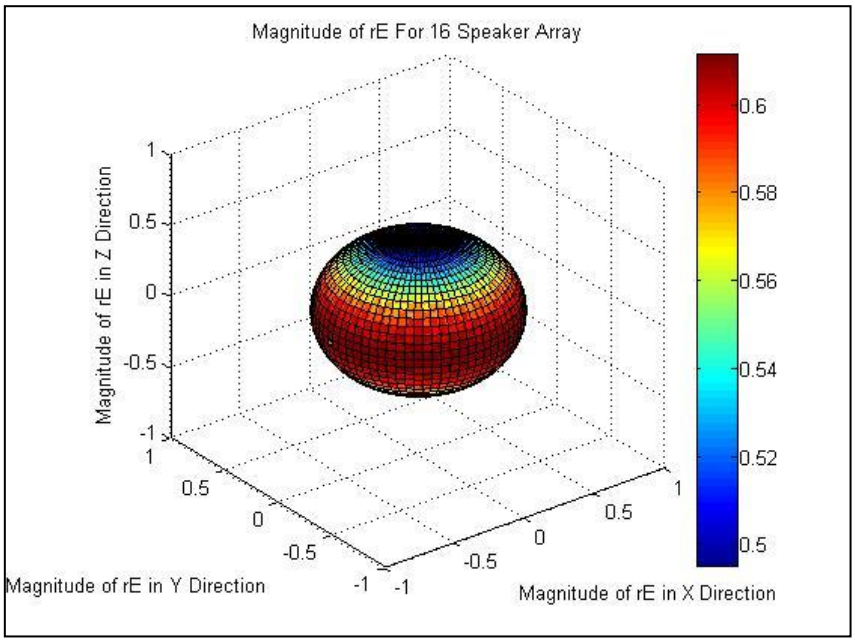


Figure 4.1 Magnitude of rE For 16 Speaker Array 1st Order

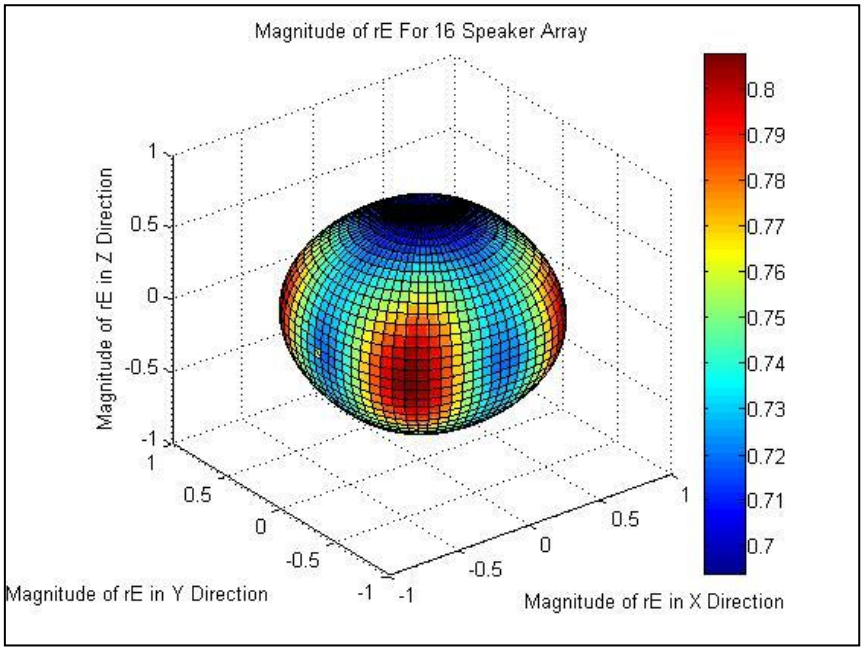


Figure 4.2 Magnitude of rE For 16 Speaker Array 2nd Order

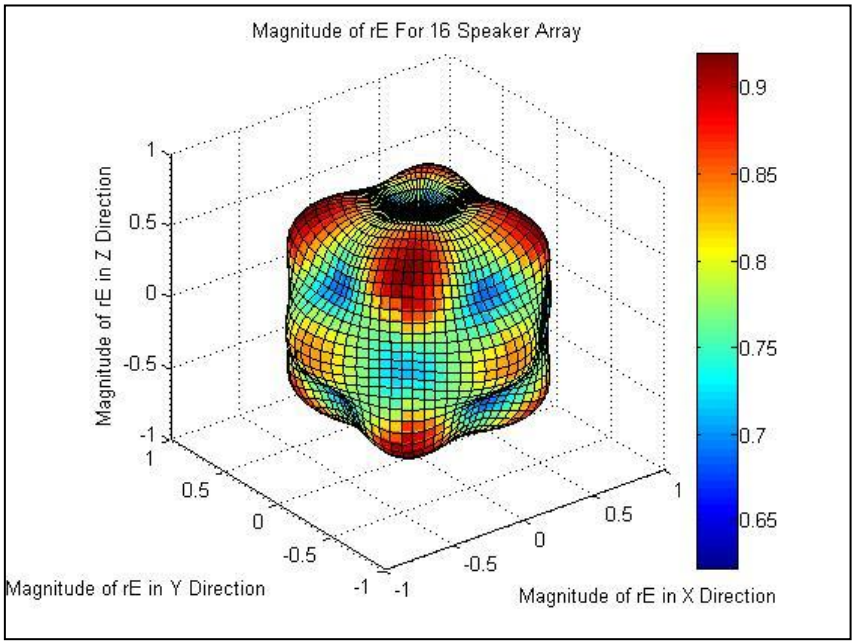


Figure 4.3 Magnitude of rE For 16 Speaker Array 3rd Order

It is stated by Gerzon (1980) that ambisonic decoder with values lower than 0.5 cause inferior reproduction and image instability. It can be seen in *Figures 4.1, .42* and *4.3* that the rE values are greater than 0.5, although for the 2nd and 3rd order decoders there is some variability in the rE values around the soundfield.

The 16 channel speaker array of Genelec 8030A speakers *See Figure 4.4* was connected to a computer via an RME MADI (multichannel audio digital interface).

There was no availability of a curtain to screen speakers so they were in full view of the listener. The test was carried out in a semi anechoic chamber which conforms to the ISO 3744, ISO 3745, and BS 4196 standards.

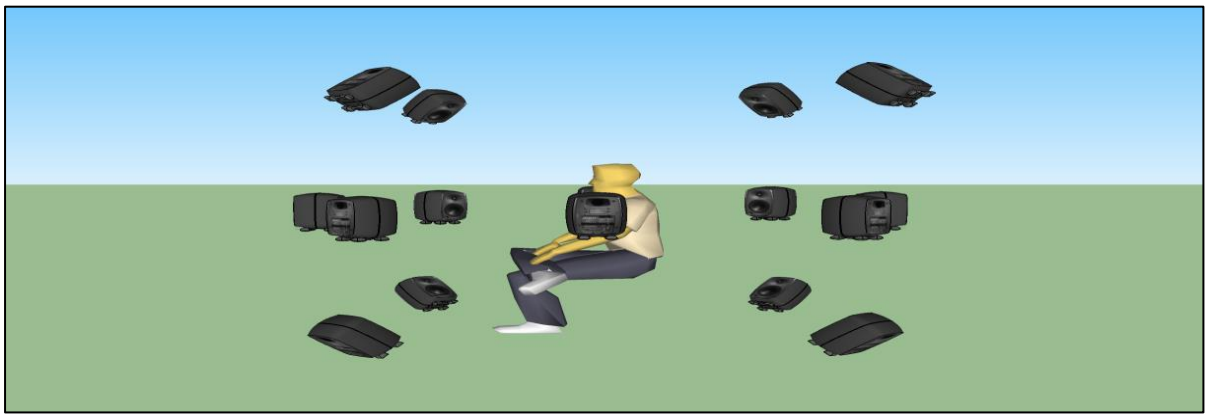
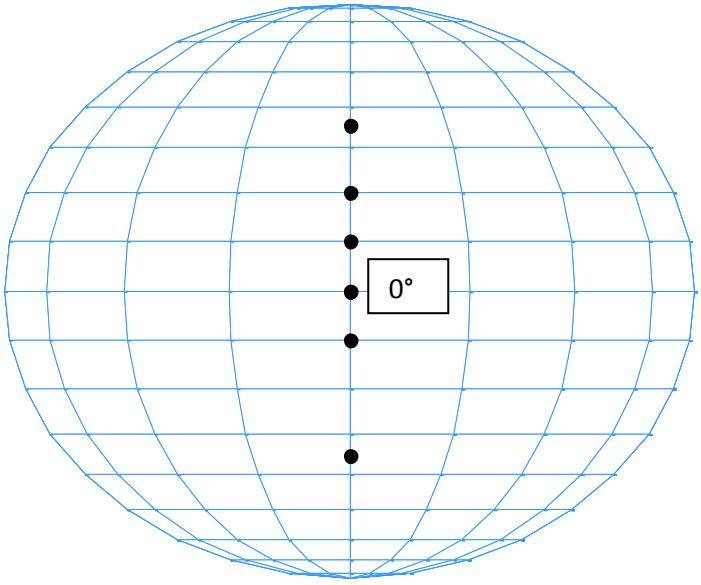


Figure 4.4 Speaker Array Used in Subjective Test

The positions tested are shown below. The increments in source positions were chosen due to factors explained by Strybel and Fujimoto (2000) that the minimum audible vertical angle is 4-5° therefore panning increments larger than this should in theory be detectable by listeners. The extremes of the test positions were governed by the speaker stands, which only allowed a maximum and minimum elevation of +- 35°. *See Table 4.1*.



Azimuth	0	0	0	0	0	0
Elevation	0	10	20	35	-10	-35

Table 4.1 Positions used in test

4.4 Stimulus

Pink noise was used as the test stimulus, since this has been used previously by Keiler and Batke (2010), Capra et al. (2007), Barbour (2003a) and also due to factors explained by Carlile (1996) that for accurate localisation spectral information across a wide range of frequencies is required. It has also been stated by Wightman and Kistler (1992) that localising the direction of broadband sound source the inter-aural time differences dominate.

A pre run of the test had shown that listeners were satisfied with their judgement of sound source position after three bursts of the pink noise. Therefore the pink noise was made to be three 2 second bursts with a silent gap in between with a short fade in/out.

All of the speakers in the array were aligned inputting pink noise of -20dB and measuring the resulting output at the listening position to be 70dB(A) since it was found by Airo et al. (1996) that most comfortable listening levels ranged between 52 and 88dB(A) and that 69dB(A) was the average most comfortable listening level for music.

Further, in order to level align each of the ambisonic renderings subjective level alignment was used, subjective loudness data has been used previously to validate the accuracy of loudness meters (Crockett et al., 2004). To achieve equal subjective loudness between

renderings, each of the renderings were aligned relative to the centre channel which reproduced the pink noise at 70dB(A). This was achieved by two listeners comparing the level of the multichannel reproduction to that of the centre channel and adjusting the level of the multichannel reproduction to be of the same subjective level, a similar method has been used by Zacharov (2000), Hirst (2006), it was also found by Zacharov (2000) in his studies of mulitchannel level calibration, that listeners were capable of very accurate level alignment. The input levels relative to the centre channel were. *See Table 4.2.*

1st ORDER	2nd ORDER	3rd ORDER
-17.5	-12.5	-10.5

Table 4.2 Input levels to the speakers relative to the centre channel input 0dB

4.5 Angle Referencing

A scale was placed in front of the listener on the wall *See Figure 4.5* with elevation angles going from 0° at eye level to +- 40° measured using an inclinometer. This was to assist in judgement of the virtual source position.

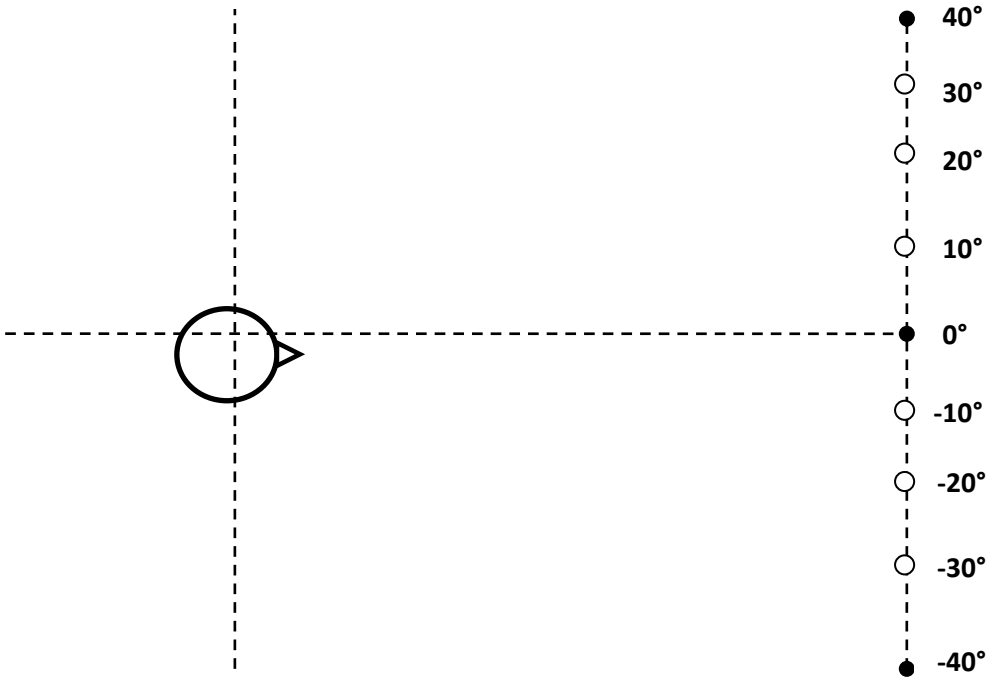


Figure 4.5 markers used in test 0° referenced to listeners gaze

The reference coordinate system was set up so that the 0° was straight ahead in line with the acoustic centre of the speaker at 0°.

4.6 Listeners

Nine listeners took part who were postgraduate researchers from the acoustic research department, who had a keen interest in experimental work and had taken part in listening tests before so were classed as selected assessors (Bech and Zacharov, 2006). Before commencing the test each listener was provided with a printed instruction sheet See *Appendix 2* and also given verbal instructions, each listener was also told to face forward and not to move their head whilst the sound was playing, however there was no way of checking that listeners were not moving their head when making judgements. Before starting the test listeners were also free to ask questions about anything they did not understand.

The reporting method used was similar to the method used by Wightman and Kistler (1989) where the participant communicated their answer stating “left 30, up 20” via an intercom system which was set up using a boundary microphone placed in the semi anechoic chamber. The listener was seated in the semi anechoic chamber alone whilst the playback of noise signals was carried out from outside the chamber by the experimenter. Each session was randomised for every participant. A short familiarisation was provided for each listener where they could become acquainted with the reporting method before starting proper.

4.7 Results

The localisation error was calculated by subtracting the participant's response from the intended position and taking the absolute value to be the localisation error. *See Equation 4.1.*

$$\textit{Localisation error} = (\textit{Intended position} - \textit{reported position}) \qquad \text{Equation 4.1}$$

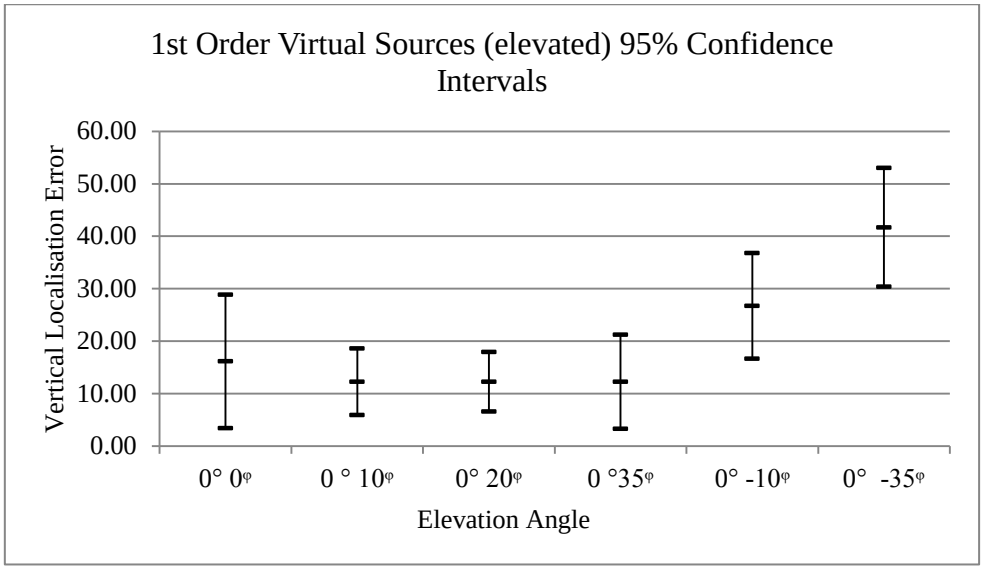


Figure 4.6 Mean and 95% confidence intervals for virtual source positions 1st order (where ° refers to angle in azimuth and φ angle in elevation)

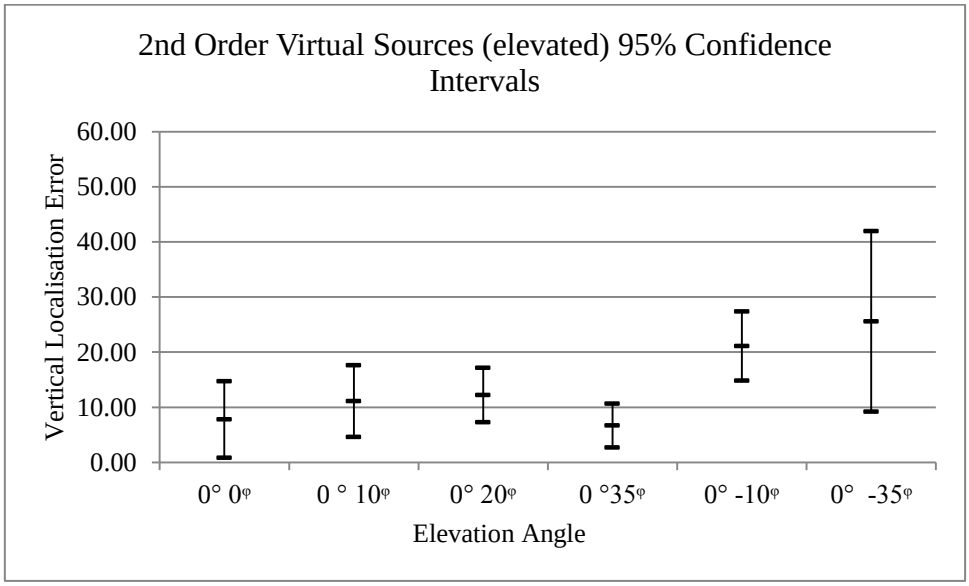


Figure 4.7 Mean and 95% confidence intervals for virtual source positions 2nd order (where ° refers to angle in azimuth and φ angle in elevation)

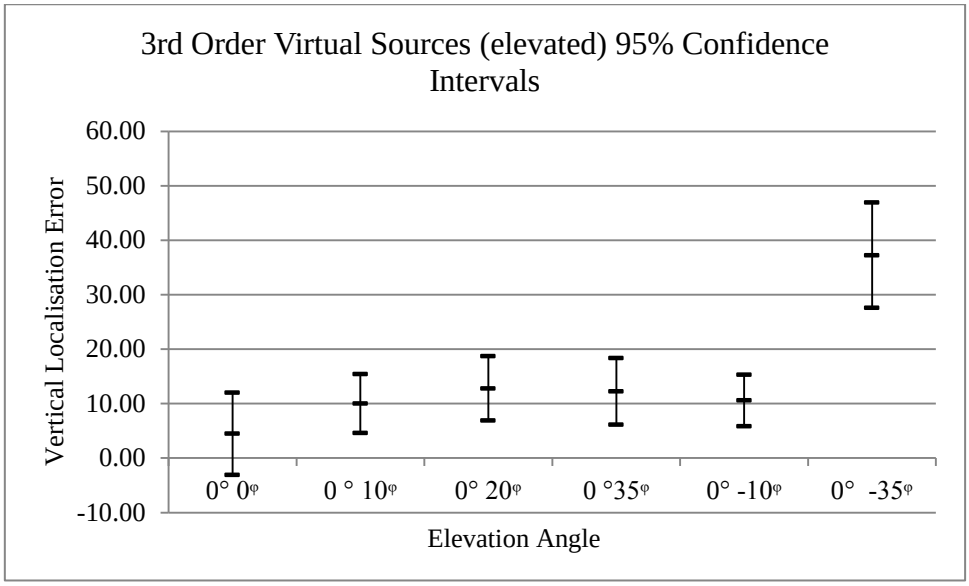


Figure 4.8 Mean and 95% confidence intervals for virtual source positions 3rd order (where ° refers to angle in azimuth and φ angle in elevation)

It can be seen in *Figure 4.6* that for 1st order ambisonics the confidence intervals are large and that moving to 2nd order ambisonics in *Figure 4.7* the confidence intervals show less variation. Then finally increasing to 3rd ambisonics generally for all positions have a similar mean score *See Figure 4.8*. However across the three orders it appears that the two sources which are placed below the listener are difficult to localise, this is apparent across all orders but most prominent in the 1st order case. There is an exception for the 2nd and 3rd order case where it seems that the widest confidence intervals are for the source at 0° -35_{φ} indicating a large variability in localisation for this position and that the increase in ambisonic order did not alleviate the variability for this position.

Since the experiment set out to find if higher order ambisonics could improve localisation accuracy in the median plane, the average error was calculated for each of the positions. The mean and 95% confidence limits are shown for each order *See Figure 4.9.*

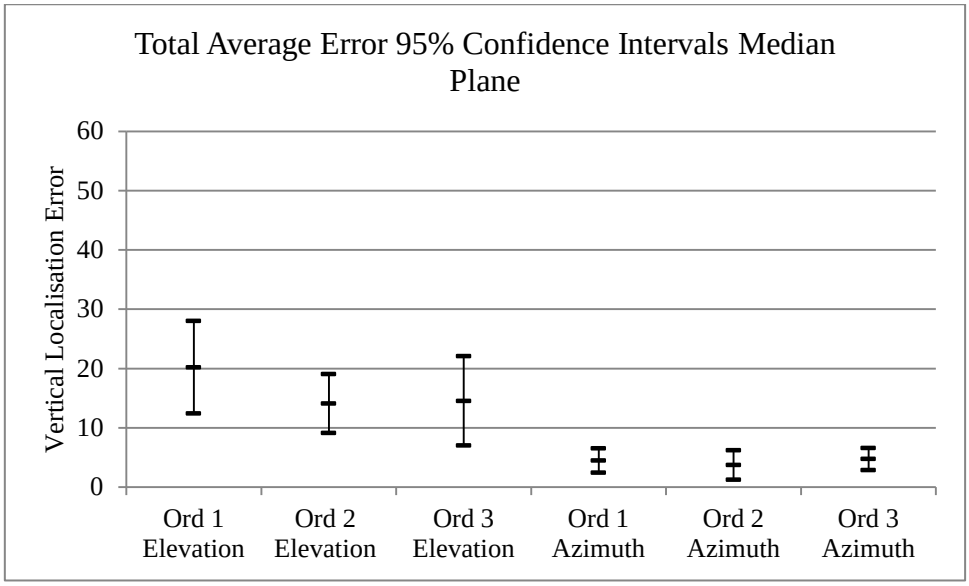


Figure 4.9 Mean and 95% confidence intervals for the 3 orders tested

It can be seen that the confidence limits overlap which could indicate that there are no significant differences between the orders for the positions tested. However to verify this the mean error for each position for each order was entered into a one way repeated measures ANOVA to confirm this, with the factor Order. A Mauchly's test of sphericity was non-significant indicating that the data meet the assumption of the ANOVA i.e.(the variances are not significantly different) (Field, 2009).

The subsequent repeated measures ANOVA showed that there were no significant differences between orders. *See Table 4.3*

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
order	Sphericity Assumed	138.670	2	69.335	3.038	.093	.378
	Greenhouse-Geisser	138.670	1.833	75.657	3.038	.100	.378
	Huynh-Feldt	138.670	2.000	69.335	3.038	.093	.378
	Lower-bound	138.670	1.000	138.670	3.038	.142	.378
Error(order)	Sphericity Assumed	228.217	10	22.822			
	Greenhouse-Geisser	228.217	9.164	24.902			
	Huynh-Feldt	228.217	10.000	22.822			
	Lower-bound	228.217	5.000	45.643			

Table 4.3 Repeated Measures ANOVA Summary Table for Significance Test between Orders

4.8 Discussion

This initial pilot test illustrated some points regarding localisation in the median plane. When looking at the individual results for each order and position as expected it was found that due to the larger confidence intervals the 1st order renderings had the most variability and improvements could be seen across positions for 2nd and 3rd order due to the tighter confidence intervals.

Across all orders it appears that for the virtual sources panned below the listener these have more variability, especially for 1st order, with improvements being noticed for 2nd and 3rd order, yet for all orders it appears that creating a virtual source at 0° -35φseems to be problematic.

Looking at the total mean error for each order for all elevated positions it has been found that certainly 1st order ambisonics has a higher mean localisation error as would be expected, and the mean error for 2nd and 3rd order is nearly identical. However a subsequent repeated

measures ANOVA found that there is no significant difference between the 1st, 2nd and 3rd order, although the spread of results for the 2nd order are much less than the 1st and 3rd order.

A similar result was found by Pieleanu (2004) who investigated 1st order and 2nd order in the median plane and found no significant difference between orders similar to this case using pink noise in addition Pieleanu (2004) also used artificial reverberation to simulate a more reverberant environment and found that this did not have a significant effect on the localisation error compared to the anechoic environment. Listeners stated after the test that it was difficult to discern between samples but that there were timbral differences between some of the stimuli.

Criticisms could be levelled on the grounds of the experimental procedure, since the speakers were in full view of the participants this could have affected the results by causing listeners to quantize their judgements to speaker positions, it is also interesting to note that there were no front/back confusions possibly because the speakers were in full view of the listener. The reporting method used could also have caused errors because of the non-intuitive way it was carried out. Therefore improvements could be made to the test design to obtain better results by including a curtain to hide speakers and using a more reliable reporting method. A further improvement to the test design would be to include a real speaker at the virtual source position as this would help validate the listener's response.

As previously highlighted in the literature review localisation in the median plane is based mainly on spectral cues with the sound source being positioned depending on its spectral content. In this study broadband pink noise was used therefore listeners would have a large number of spectral cues available to them which could have influenced their perception of source location. It was also noticed that the distribution of the listeners errors were quite large suggesting that the sample of listeners were not certain of the source position. This was also found by Pieleanu (2004) who stated that there was no consistency between listeners judgements. This would then suggest that the effort of using higher order ambisonics in the median plane may not be required, although, it would be interesting to extend the evaluation to higher orders to observe if this is still the case, since it was noticed in this research that increasing the ambisonic order reduced the mean error, therefore it may be that extending the ambisonic order may provide further improvements. It would also be interesting to include different filtered versions of pink noise and render these using different ambisonic orders to

investigate if the positioning of the sound source is independent of ambisionic order and dependant on the sound source spectral content.

4.9 Summary

A pilot test has been carried out to investigate localisation in the median plane using a surround system with height utilising pink noise as the test signal. It was also investigated if the addition of higher order ambisonics would improve localisation, however it was found that there was no significant difference between ambisonic orders, however that the extension of higher orders greater than 3rd may provide different results.

4.10 Conclusion

A localisation test has been carried out to investigate if the addition of higher order ambisonics can reduce the localisation blur in the median plane. Utilising pink noise and rendering virtual sources at several different positions in the median plane using 1st,2nd and 3rd order ambisonics it has been found that the use of higher order ambisonics does not significantly reduce the localisation error.

As outlined in the literature review there are less cues available to the listener for sources in the median plane compared to the frontal plane. Therefore the next chapter will investigate if higher order ambisonics can reduce the localisation error in the frontal plane where there are more cues available to the listener.

5 Experiment 2 Localisation in the Frontal Plane

5.1 Introduction

Since the previous subjective experiment was limited to positions in the median plane it was decided to extend the test to positions off the median plane. Due to the aforementioned factors of the methods used by humans to localise a sound, placing virtual sources off the median plane it was hypothesized that this would provide the participants with more cues. As previously mentioned in *Chapter 2*, Capra et al. (2007) evaluated the localisation accuracy of 1st order ambisonics using a 16 channel 3D speaker array compared to a 3D stereo dipole system and found that 1st order provides inaccurate localisation. In addition Morrell and Reiss (2009) investigated the localisation accuracy of 3rd order ambisonics compared to VBAP rendering and found that 3rd order ambisonics provides comparable localisation accuracy to that of VBAP, however this was for only four positions and their comparisons were not between lower orders of ambisonics. As before, it was of interest to investigate the differences in localisation of virtual sources using different orders of ambisonic reproduction in order to find out if utilising higher order ambisonic improves the ability to localise virtual elevated sources.

Therefore this additional localisation experiment sought to investigate sources off the median plane. In addition to virtual source positions, real sources were also used which was a speaker placed at the virtual source position in order to compare to the virtual source. The positions for the real and virtual sources are shown below *in Figure 5.1 and Table 5.1*.

The majority of virtual source positions were placed in front of the listener so that the listener could make their judgement without moving their head. And because this is the area where listeners localisation ability is most accurate (Blauert, 1996). Unfortunately it was not possible to place a real speaker at each of the virtual source positions so a compromise had to be made to cover some of the virtual positions.

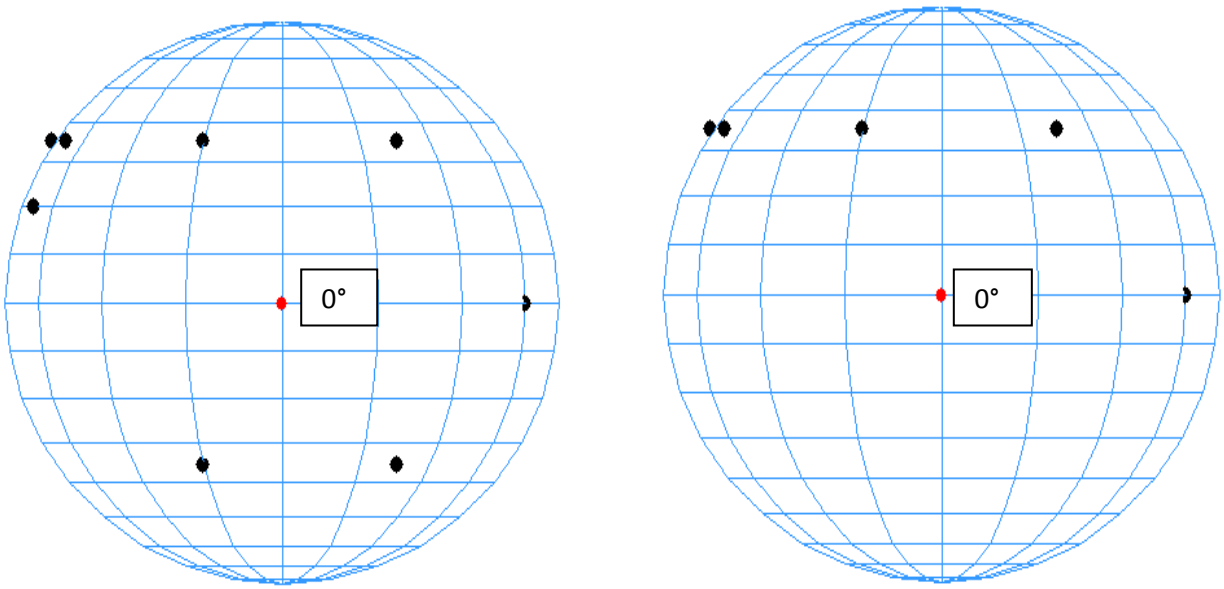


Figure 5.1 Left Virtual Source Positions, Right Real Source Positions

Virtual Sources	Azimuth	Elevation	**Real Sources**	Azimuth	Elevation
Pink Noise, Speech	20	35	Pink Noise, Speech	20	35
Pink Noise, Speech	20	-35			
Pink Noise, Speech	70	20			
Pink Noise, Speech	70	35	Pink Noise, Speech	70	35
Pink Noise, Speech	90	35	Pink Noise, Speech	90	35
Pink Noise, Speech	300	0	Pink Noise, Speech	300	0
Pink Noise, Speech	330	-35	Pink Noise, Speech		
Pink Noise, Speech	330	35	Pink Noise, Speech	330	35

Table 5.1 Stimuli and real and virtual positions tested

5.2 Test Set Up

The same speaker array was used as the previous test using Genelec 8030A speakers, however in order to make the test more rigorous, in addition to the sixteen speakers for the ambisonic reproduction, a further 5 speakers were also placed at a select few positions corresponding to where the virtual sound source would be positioned. This would be a

control condition, and would allow a comparison of the real and virtual source, and also determine the accuracy of the participants answer since this was lacking in the previous test.

Further to this an acoustically transparent but visually opaque curtain was also used, this would ensure no visual cues would be provided. The front panel indication lights on all of the speakers were also covered, including the front panel of the digital to analogue converter metering to negate any further visual clues.

Each of the individual active monitors were level aligned using pink noise of -20dB input and measuring the level at the listening position to be 70.1dB(A) since it was found by (Airo et al., 1996) that most comfortable listening levels ranged between 52 and 88dB(A) and that 69dB(A) was the average most comfortable listening level for music.

To align each of the ambisonic renderings the same method as previously described in *Section 4.4* was used, where each of the reproduced positions in 1^{st}, 2^{nd} and 3^{rd} order were compared to the centre channel of input 0dB. *See Table 5.2.*

1st ORDER	2nd ORDER	3rd ORDER
-9.8dB	-9.8dB	-5.2dB

Table 5.2 Input levels to the speakers relative to the centre channel at 0dB

The level of each of real sources (not rendered using ambisonics) were made to be 70dB(A).

Again the same marker set up used in the first localisation test shown in *Figure 4.5* was used to assist participants in the judgement of the virtual sources position, which was relayed to the experimenter via a two way intercom. A sheet of paper with azimuth angles marked on it was also placed in front of the listener and listeners were told to face forward and keep their head stationary whilst the sounds were played.

The reporting method used was similar to the method used by Wightman and Kistler (1989) where the participant communicated their answer stating “left 30, up 20” via an intercom system which was set up. The listener was seated in the semi anechoic chamber alone whilst the playback of noise signals was carried out from outside the chamber by the experimenter. Each session was randomised for every participant. A short familiarisation was provided for each listener where they could become acquainted with the reporting method before starting proper.

5.3 Stimulus

Since the previous localisation test utilised only pink-noise, it could be argued that this is not representative of the type of sound source that the general population would be exposed to. Further it has also been reported by Liebetrau et al. (2007) in terms of speech that humans are very sensitive to speech and as such should assist localisation. Also Neher (2004) states that human speech is very familiar and as a result very critical. Contrary to this it was found by Davis and Stephens (1974) that the use of noise and male speech in their tests that noise could be localised better than the speech sample in vertical localisation.

Therefore in addition to the pink-noise sample used previously detailed in *Chapter 4, Section 4.4* a female speech sample of 2 seconds duration was also used, this would also be of interest regarding intelligibility, since 1st order material does not provide a high resolution it would be of interest to see if this affected the ability to determine the speech sources location. The speech sample was taken from the music test CD for Archimedes (Bang, 1992) and was repeated three times for each location.

5.4 Ambisonic Decoding

The ambisonic decoding used was the max rE decoding for 1st,2nd and 3rd order, using a one band solution, as previously described. This method of decoding focuses the energy in the intended direction and provides the narrowest polar pattern reducing the size of the negative rear lobe.

The uniformity of rE for the soundfield reproduced by the 16 loudspeaker array for each ambisonic order can be seen in *Chapter 4 Figures 4.1, 4.2 and 4.3*. The Matlab script used for the encoding and decoding can be seen in *Appendix 1*.

5.5 Listeners

Eleven participants took part and they were a mixture of research students in audio technology or acoustics and also staff who had taken part in subjective experiments before, so were classed as selected assessors (Bech and Zacharov, 2006). The test instructions given to the listener can be seen in *Appendix 3*.

5.6 Results

The elevation error was calculated by comparing the participant's unsigned response from the intended panned location to obtain the localisation error. *See Equation 4.1*

Two responses were gathered for each participant, the azimuth position, and the elevation.

In order to evaluate the participant's performance before assessing the virtual sources, it was necessary to inspect the performance with regards to the real sources.

It has been explained by Strybel and Fujimoto (2000) that the minimum audible vertical angle is approximately $4-5°$ however vertical resolution is reduced at higher elevations. Inspecting the errors for the real positions it can be seen that the real sources are on average just slightly greater than this amount, using a different reporting method could have achieved more accuracy. *See Figure 5.2.*

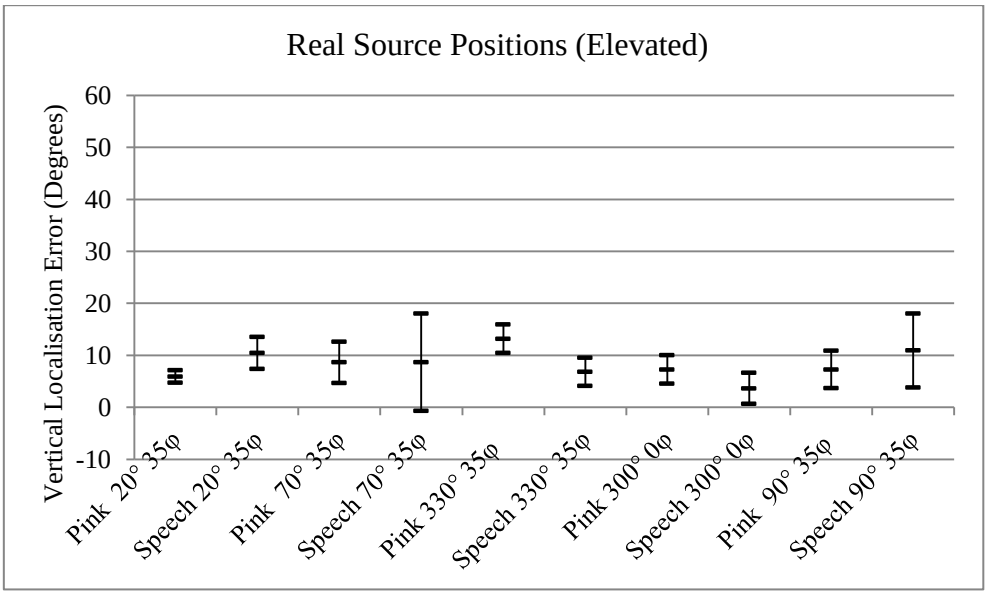


Figure 5.2 Mean and 95% confidence intervals for real source positions (where ° refers to angle in azimuth and φ angle in elevation)

For the elevated source localisation tests, looking at the differences between orders for each of the elevated positions it was interesting to note the degree of variation See *Figure 5.3, 5.4 and 5.5.*

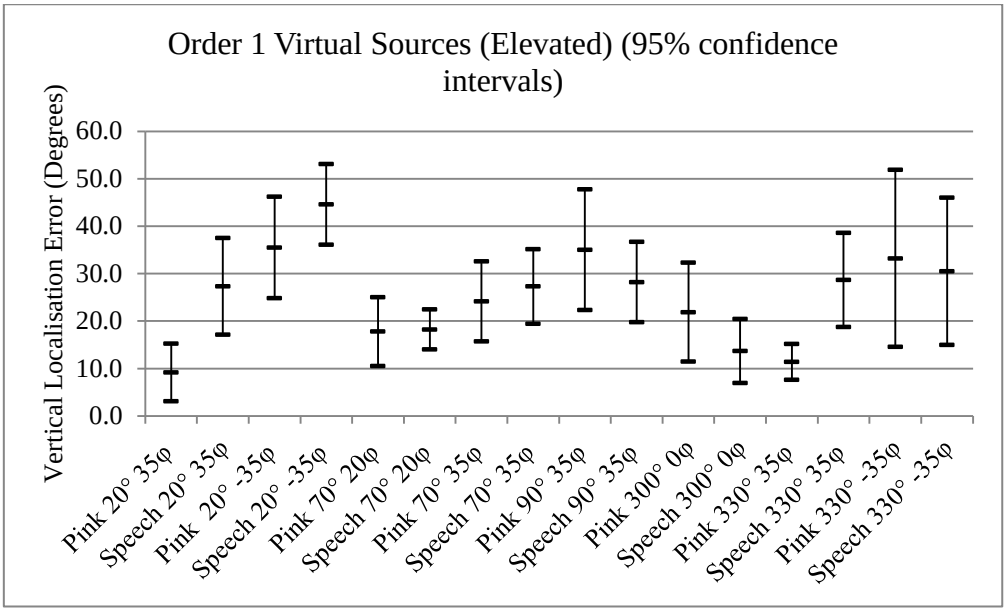


Figure 5.3 Mean and 95% confidence intervals for virtual source positions 1st order (where ° refers to angle in azimuth and φ angle in elevation)

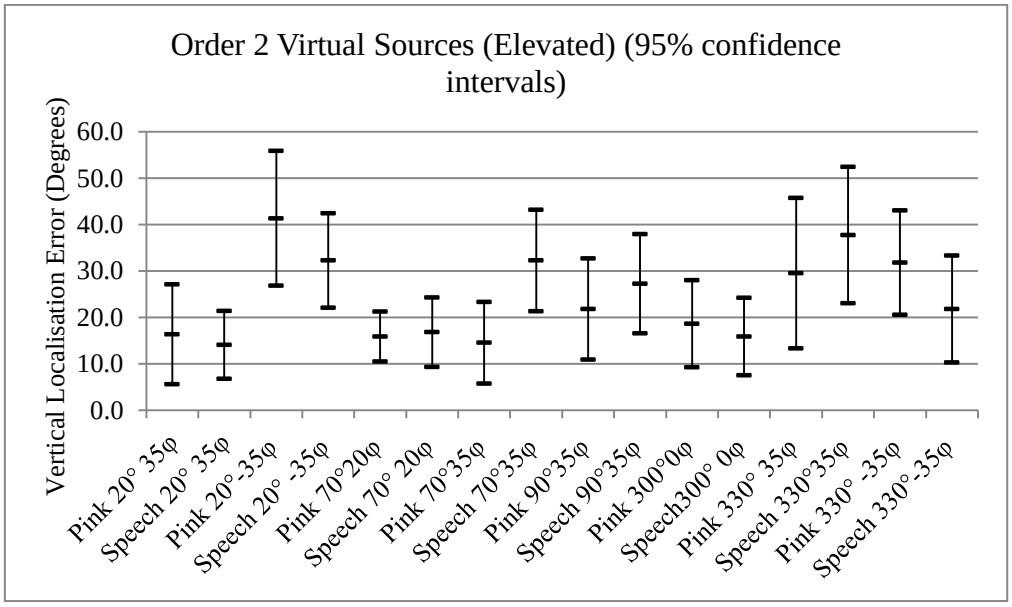


Figure 5.4 Mean and 95% confidence intervals for virtual source positions 2nd order (where ° refers to angle in azimuth and φ angle in elevation)

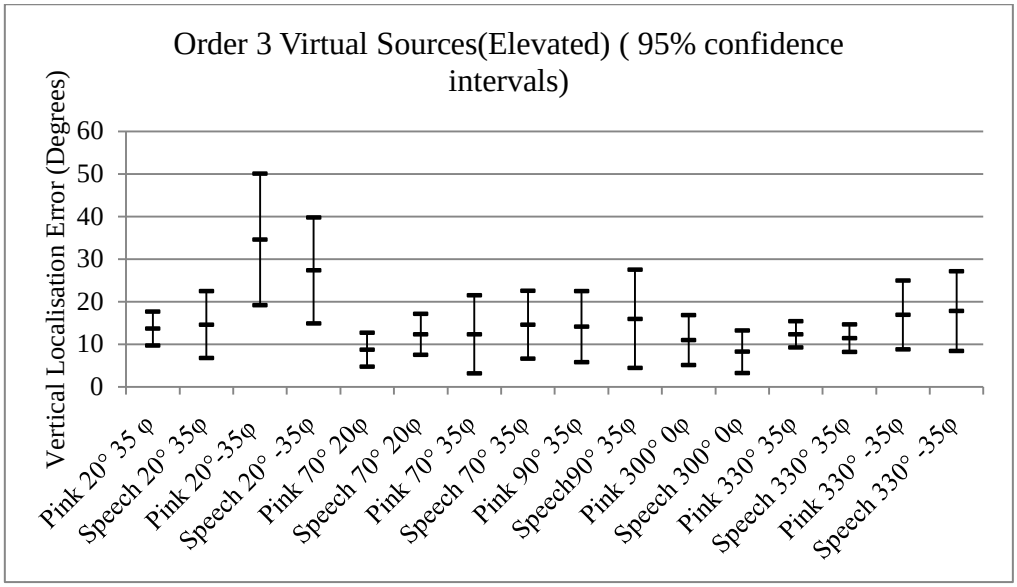


Figure 5.5 Mean and 95% confidence intervals for virtual source positions 3rd order (where ° refers to angle in azimuth and φ angle in elevation)

It can be seen however from the large confidence intervals that the sources placed below the participant especially at $20° - 35\varphi$ caused up/down confusions for all orders, which were also in combination with front back errors. However in terms of 3rd order, the number of confusions for this position was much lower, from the raw data the number of up/down confusions were counted for each order and can be seen in *Figure 5.6.*

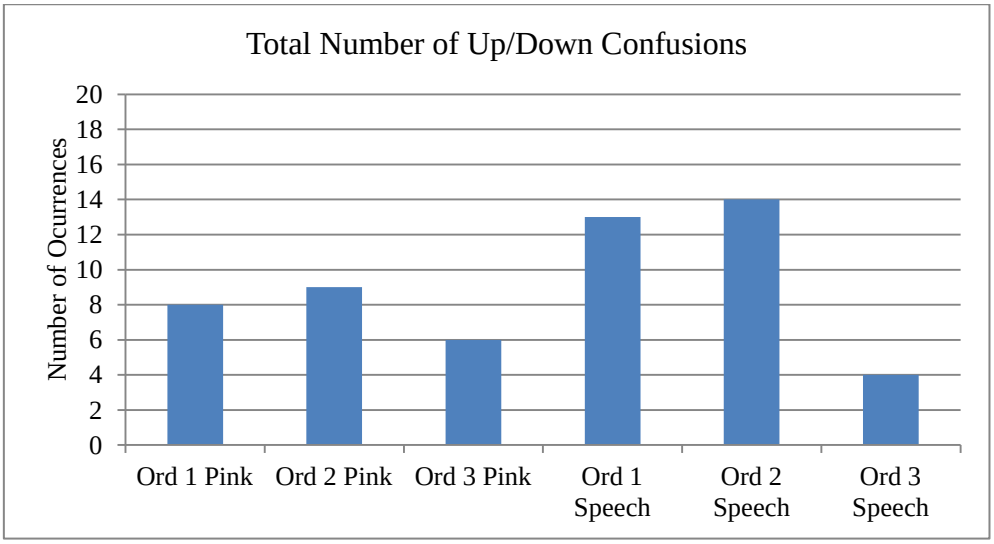


Figure 5.6 Total number of up/down confusions 1st-3rd order

It was also noticed that where participants had up/down confusions a number of front/back confusions were also prevalent, again the number of occurrences were counted from the raw data and are displayed below for each order. *See Figures 5.7 and 5.8.*

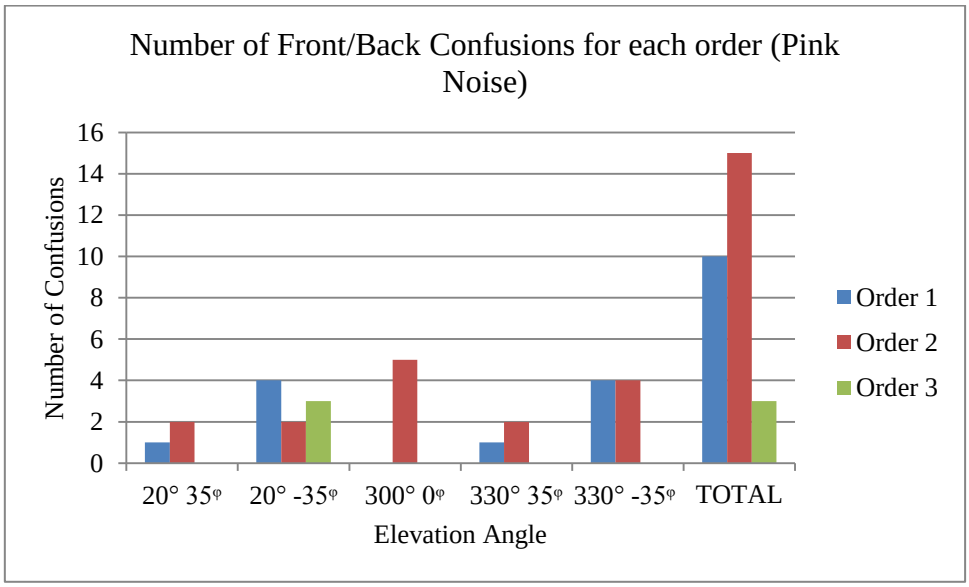


Figure 5.7 Number of Front/Back Confusions 1st, 2nd and 3rd Order for Pink Noise

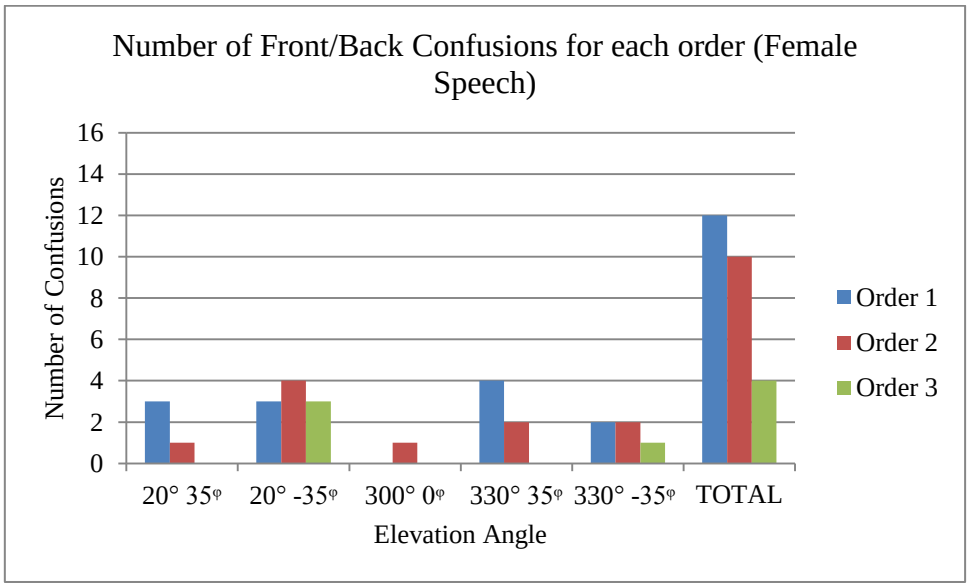


Figure 5.8 Number of Front/Back Confusions 1st, 2nd and 3rd Order for Female Speech

To determine which order performed with the least error the total mean error over all elevated positions was examined. It can be seen that the 3rd order has least total average error

compared to 1st and 2nd order for the speech item. A similar trend was also noticed for the pink noise item. *See Figure 5.9.*

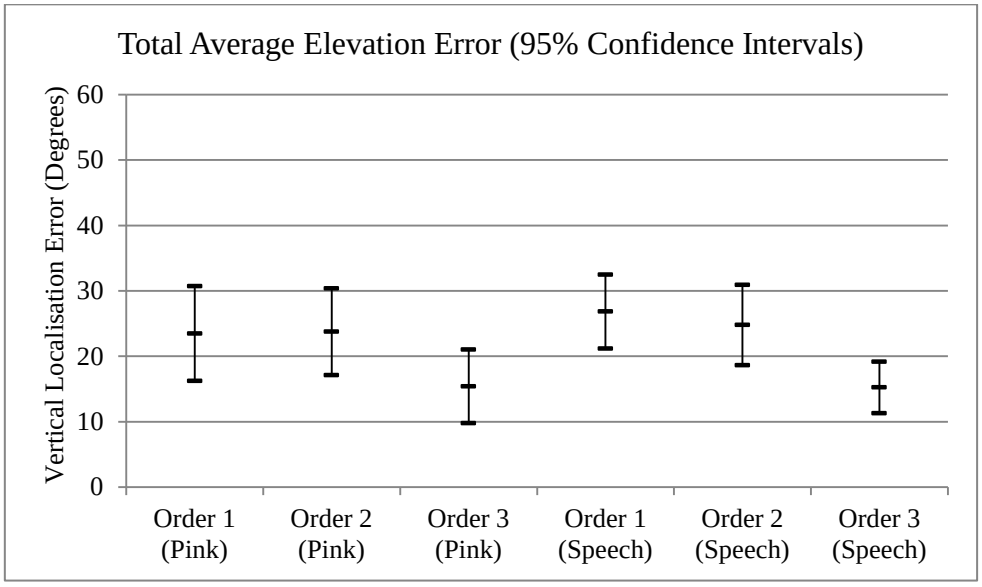


Figure 5.9 Total average error for 1-3rd order

In order to determine if the performance of the 3rd order rendering was significantly different to that of the other renderings for speech or pink noise a two way repeated measures ANOVA was used with the Factor Order and Stimuli at the 0.05 significance level, a Mauchlys test of sphericty was non-significant indicating that the data meet the assumptions of the ANOVA i.e.(the variances are not significantly different) (Field, 2009).

It was found for the factor Order this was significant at the $(p < 0.05, partial\ \eta^2 = .687)$ however the Factor Stimuli was not significant $(p > 0.05, partial\ \eta^2 = .075)$ and the interaction Order*Stimuli was not significant $(p > 0.05, partial\ \eta^2 = .060)$.

To determine where the specific differences were between the orders a Bonferroni post hoc test was carried out, this showed that there was a significant difference between 1st and 3rd and 2nd order and 3rd order in terms of overall mean localisation error, with 3rd order having a significantly smaller error, *See Table 5.4.*

Pairwise Comparisons

Measure: MEASURE_1

(I) Order	(J) Order	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
1	2	.879	2.208	1.000	-6.025	7.784
	3	9.828*	1.489	.001	5.170	14.486
2	1	-.879	2.208	1.000	-7.784	6.025
	3	8.949*	2.108	.011	2.357	15.540
3	1	-9.828*	1.489	.001	-14.486	-5.170
	2	-8.949*	2.108	.011	-15.540	-2.357

Table 5.3 Bonferroni post hoc output table of pairwise comparisons

5.7 Discussion

It was noticeable from the initial inspection of the data the inconsistency of the 1st and 2nd order compared to the 3rd order which placed sources with the least error and with similar localisation to that of real sources, this was found for elevation and azimuth judgements. However elevated sources placed in front of participants using 1st and 2nd order caused a number of front/back confusions as reported, with front/back confusions for 3rd order only being found where sources were placed below participants. Overall larger azimuth errors were found for 1st and 2nd order, a factor that could have caused larger azimuth errors could be due to angular dispersion at higher frequencies as a function of order explained by Daniel (2009). It was also noticed that for the position constrained to the horizontal, participants tended to place this off the horizontal, this was also evident for the real sources, and was also more pronounced for the 1st order rendering. Larger offsets were more evident for the pink noise source than speech, possibly due to more high frequency content, overestimation of elevated sound sources was also found by Liebetrau et al. (2007).

It was thought that the up/down confusions could have been caused by the close proximity of the lower speaker ring to the listener. However the 3rd order material did not suffer badly from this, and up/down confusions for 1st order were also found by Capra et al. (2007). So it is thought that due to the large number of speakers working in the 1st and 2nd order case to pan the virtual source this could have caused confusion for the listeners, contrary to the 3rd

order case where fewer speakers are being used to place the virtual source as explained by Bertet et al. (2006). Further, it has also been stated by Daniel (2009) that for each order the correct reconstruction of the soundfield can only be guaranteed up to a given frequency limit, for 1st order this is 700Hz and second order this is 1300Hz, as a result it has been found by Wenzel et al. (1993) that the degradation of high frequency cues using non individualised HRTF's caused listeners to have up-to-down and down-to-up confusions, in this case it could be that the inability of 1st and 2nd order to accurately reproduce the high frequency content correctly could cause the up-to-down and down-to-up confusions.

It should be acknowledged there were a number of non-ideal situations within the test design. Firstly the way the subject reported their answers could have inflated the localisation errors, because of the reliance on participants to translate where the source was into spherical coordinates. Secondly and most notably the speaker array used was not a regular array. Due to factors previously explained, because of the distribution of speakers eight on the horizontal and only four above and below it could be said that for sources placed on the horizontal there would be an unfair advantage to third order, contrary to this elevated sources would only have four speakers thus disadvantaging second and third order. It is striking that the 3rd order system did perform reasonably in light of the rules regarding ambisonic reproduction and regular 3D rigs, a theoretical analysis of a 3rd order max rE decoder has been carried out by Gorzel et al. (2014) who found that reproduction of the soundfield at the listening position was satisfactory. Regardless 1st order is particularly bad at placing sources, this could also be said for 2nd order in comparison to the 3rd order reproduction, this can be seen in Figures 5.3 and 5.4 where 1st and 2nd order ambisonics have larger confidence intervals compared to 3rd order.

In general the results indicate that 3rd order ambisonics is more effective at accurately localising panned sources in elevation and azimuth than lower order systems, although localisation accuracy remains below the limit set by the human hearing and indicated by the results for the real sources. It is possible that 4th or higher order reproduction may be required to achieve vertical localisation accuracy as good as human hearing. However in terms of requirement for mixed order systems it should be noted that the results presented above were obtained under semi anechoic conditions and in real world reverberant environments the highest Ambisonic order required for the vertical plane might be considerably lower, however work carried out by Pieleanu (2004) into localisation in the median plane found that

the addition of reverberation to simulate natural listening environments did not change the localisation error of virtual sources.

It would seem from these results that the use of 3rd order ambisonics provides the lowest localisation error for the two sound sources used, therefore to provide accurate localisation for elevated sources in the vertical plane 3rd order would provide more accuracy than 1st or 2nd order, this was the case regardless of sound source used. It may be that lower orders like 1st order could be used when accurate localisation of sources is not necessary in order to create broader sound sources using more diffuse content. A clue to this can be found in the work conducted by Baume and Churnside (2010) using 1st order recordings, it was found that for speakers above and below participants there was a clear preference for atmospheric non-directional content, whilst more directional content , like point sources, without room effects the 5.1 surround system was preferred. Possibly due to factors outlined in this investigation that 1st order ambisonics cannot localise sources with accuracy so would be more suited to be used with non-directional content, like reverb and other room effects.

5.8 Summary

The accuracy of localisation with elevated virtual sources has been investigated. Three different orders of ambisonics have been evaluated to determine if using higher order ambisonics can improve the localisation accuracy. It was found that 3rd order ambisonics provides a significantly lower localisation error for the two sound sources used. However it was found that 3rd order ambisonics could not match the localisation accuracy of the real sound sources, therefore in order to match the localisation accuracy of the real sound sources it would be necessary to investigate the extension of ambisonic rendering greater than 3rd order.

5.9 Conclusion

The extension of higher order ambisonics have been investigated to observe whether the inclusion of 3rd order ambisonics can improve localisation for elevated virtual sources. Two different sound sources were used speech and pink noise and it was found that 3rd order ambisonics can significantly reduce the localisation error, however that this is still not as accurate as a real sound source e.g a speaker placed at the virtual source position.

It was identified in the literature review that localisation is not the only important aspect in the perception of multichannel systems. It was described by George et al. (2010) that envelopment is one of the attributes driving multichannel development, therefore the next

chapter will investigate the perception of envelopment using 3D surround systems and also the effect of varying the number and positions of height channels on the perception of envelopment and to answer the questions:

1. How does the addition of height impact on the perception of envelopment?
2. How does the number and positioning of height channels impact on the perception of envelopment?

6 Evaluation of Envelopment in With height and Horizontal Surround Part 1

6.1 Introduction

It was highlighted in the literature review that there seemed to be some disagreement between researchers, regarding the optimal speaker layout that includes height to provide the best sense of envelopment. It was reported by ITU (2012) that an increase in speakers beyond twelve in the horizontal saturated the sensation of envelopment, however an increase in speakers above a listener continued to show improvement in the perception of envelopment. Contrary to this Sawaya et al. (2013) using the same stimulus (decorrelated white noise) found that changes in elevated speaker azimuths made no significant difference to the perception of spatial impression. Therefore this investigation seeks to find if there is an optimal layout in terms of elevated channels to enhance envelopment and to add clarity to the aforementioned contradicting research by ITU (2012), Sawaya et al. (2013)

The first stage involved investigating a number of different 2D and 3D speaker arrays. This was firstly carried out using a simulation in Matlab to measure IACC differences between each of the systems and to investigate how the height channels influence the IACC.

The next stage would be to execute a subjective test with the most promising layouts to find, firstly, the system which is rated the most enveloping which will be used in subsequent tests and as a caveat to investigate how well the objective measure predicts the subjective impression. A similar method has been used by Hiyama et al. (2002) investigating the optimum number of speakers in the horizontal to reproduce the spatial impression of a diffuse soundfield.

6.2 Objective Investigation

The objective measure IACC (Inter aural cross correlation) allows the measurement of the degree of correlation between the left and right ear signals and has been linked to spatial impression (Hiyama et al., 2002). The measurement has been used previously in spatial audio reproduction systems by Tohyama and Suzuki (1989), Hiyama et al. (2002), Mason (2002), Hirst (2006), George et al. (2010), using music and noise signals.

IACC is defined as the maximum value of the normalised cross correlation function for the time interval of ± 1ms. *See Equation 6.1* and *6.2* where Pl and Pr are the impulse responses at

the left and right ears, $\tau$is the offset between ±1ms and is set to account for the maximum inter-aural time difference. t1 and t2 are the time limits of the integration, different variants of IACC exist depending on the time limits. The most general version is IACC for which t1 would be 0 and t2 = inf. This IACC is based on the degree of correlation between the impulse responses at the left and right entrance of the ear. A well-defined source has a sharp maximum of 1, whilst subjective diffuseness has a low value <0.15 (Ando, 1985).

$$IACF_{t1t2}(\tau) = \frac{\left[\int_{t1}^{t2} Pl(t) \cdot Pr(t+\tau)\, dt\right]}{\left[\int_{t1}^{t2} Pl^2(t)\, dt \int_{t1}^{t2} Pr^2(t)\, dt\right]^{\frac{1}{2}}} \quad \text{Equation 6.1}$$

$$IACC_{t1t2} = max\left|IACF_{t1t2}(\tau)\right|, for -1ms < \tau < +1ms \quad \text{Equation 6.2}$$

A large amount of research has been carried out into the IACC and the measurement is not a perfect solution since it does not take into account the effect of low frequencies due to the wavelength of the signal and the distance between the ears (Tohyama and Suzuki, 1989) and as a consequence, it has been explained by Blauert and Lindemann (1986) in terms of concert hall acoustics, that early lateral reflections which contain mainly frequencies up to 3 kHz contribute solely to the expansion in the front back direction, referring to the attribute envelopment. In addition it has also been found by Hidaka et al. (1995) that using the IACC for the measurement of apparent source width mainly IACC values averaged over the 500Hz-2kHz predicted apparent source width.

Bearing this in mind, research has shown that it does to some degree correlate with spatial impression (Nakayama et al., 1971), (Hiyama et al., 2002). Therefore has been used to provide guidance to the expected degree of envelopment for each system investigated.

6.3 Simulation

Sixteen different systems were simulated which included 2D and 3D surround systems, in addition a number of systems also used by NHK in their subjective tests (ITU, 2012) were also simulated, these are highlighted red in *Table 6.1* below.

SYSTEMS (Horizontal)	No Speakers	SYSTEMS (Height) +35°	No Speakers
Mono	1	2x4	8
Stereo	2	2x6 Rings	12
Quad	4	2x8 Rings	16
5.1	5	2x12 Rings	24
Hexagon	6	3x6 Rings	18
Octagon	8	3x8 Rings	24
12 Ring	12	Cube_8	16
24 Ring	24	2x24 Rings	48

Table 6.1 Speaker arrays used in simulation

6.4 HRTF's

In order to simulate the different systems a number of HRTF's (head related transfer functions) of the different speaker positions using the MIT database (Gardener and Martin, 1997) were convolved with de-correlated white noise, these were then summed for the left and right ears, and then the degree of correlation was measured for each system utilising the IACC measurement *See Figure 6.1*.An example of the Matlab script used to carryout this investigation can be seen in *Appendix 4*.

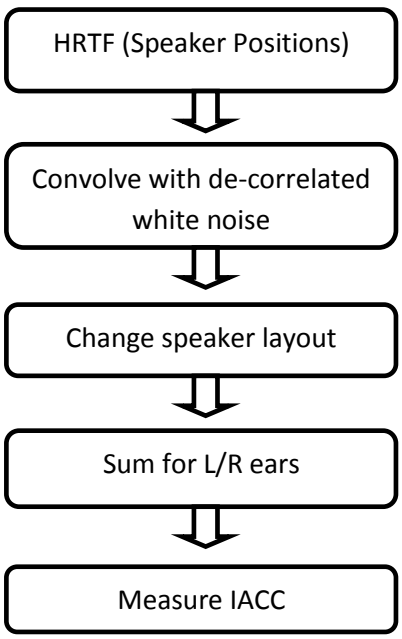


Figure 6.1 Simulation Flow

6.5 Simulation Results

Firstly the correlation for each individual white noise source used to feed to each of the loudspeakers is presented. *See Figure 6.2.*

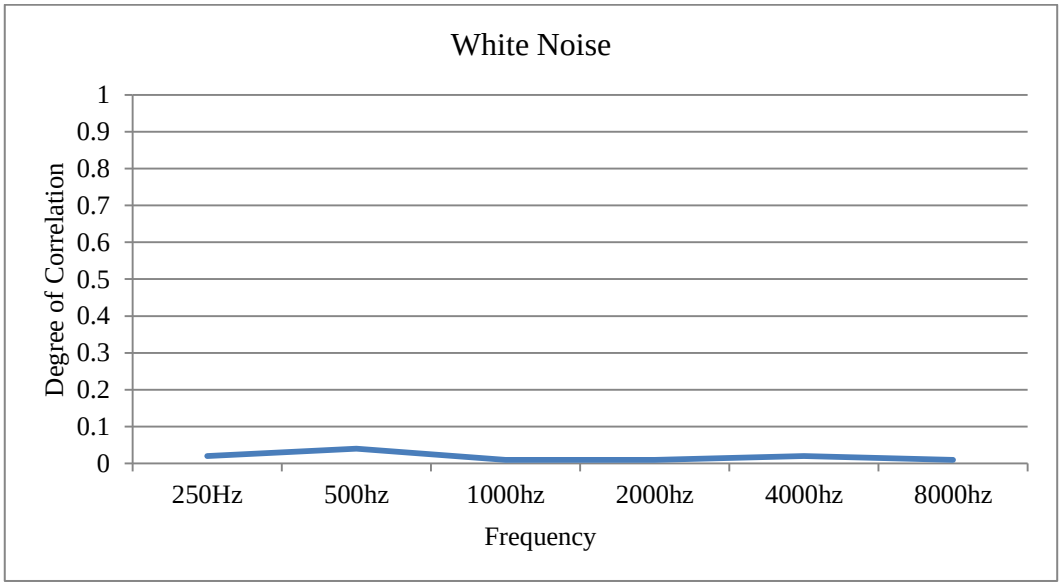


Figure 6.2 IACC of white noise sample

It has previously been stated by Tohyama and Suzuki (1989) that subjective diffuseness can be created by randomness as a statistical property of noise, therefore approximating a reverberant field of a concert hall.

It can be observed that the signals are uncorrelated, therefore reproducing different instances of white noise from each speaker over the different systems, should in theory produce different values of IACC only dependent on the system configuration, since it is stated by Berg and Rumsey (1999b) that from their experience changing the number and distribution of channels gives different sensations of spatial impression.

A number of systems were measured which included horizontal only systems, however since this study is concerned with height systems, only the results for the height systems will be shown with exception of the current 5.1 system.

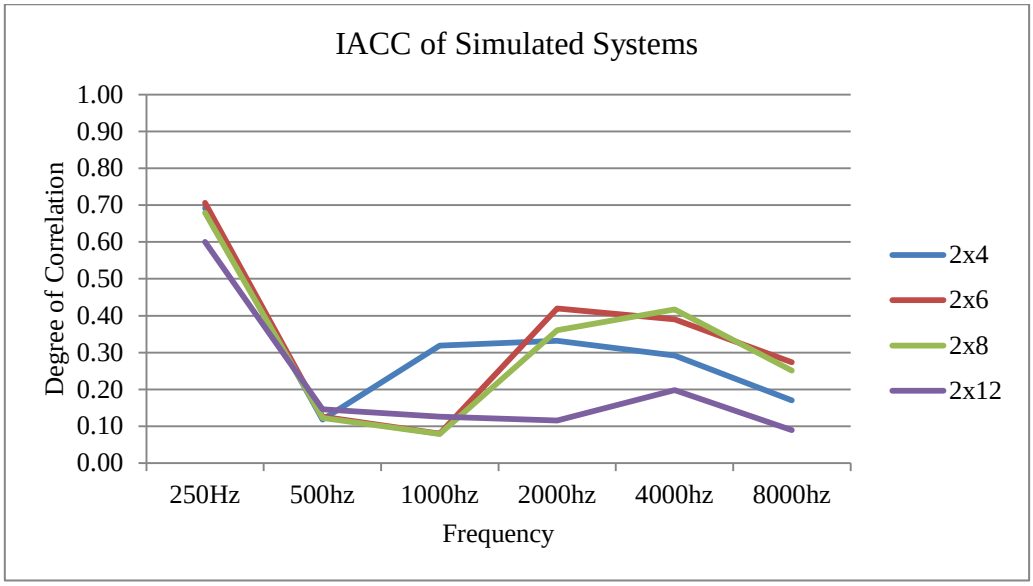


Figure 6.3 IACC of simulated height systems

An expected trend can be observed for the four systems, with the 2x4 channel system showing more correlation towards the values of interest between 500Hz and 2kHz, and the 2x6 channel and 2x8 channel showing a similar performance with near perfect de-correlation between 500Hz and 1kHz. However it can be observed that the 2x12 speaker configuration producing consistent de-correlation although showing a peak around 0.2 which demonstrates that this could be the best that can be achieved. *See Figure 6.3.*

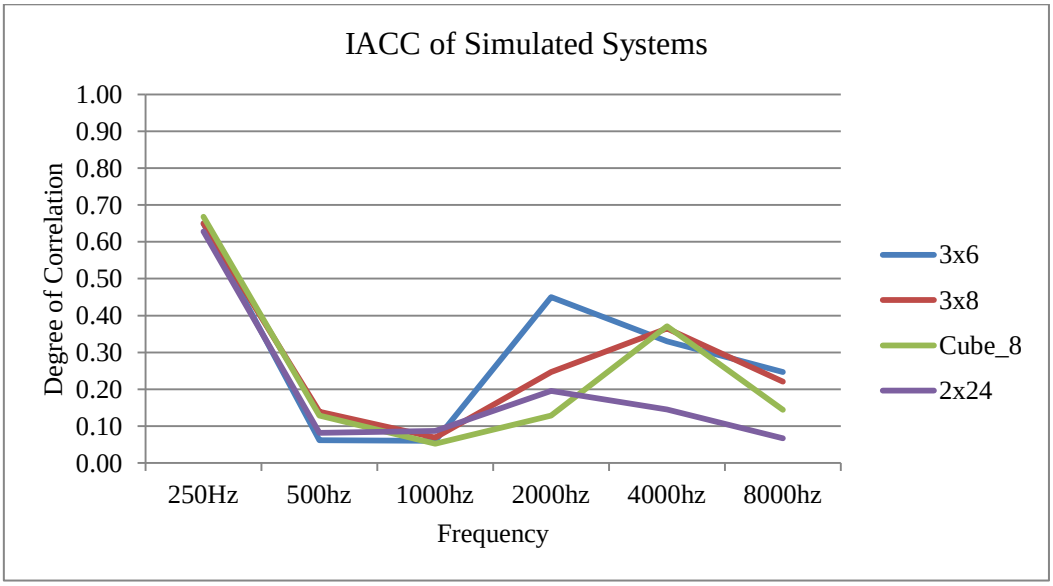


Figure 6.4 IACC of simulated height systems

Adding a further bottom layer which increases the speaker count and provides three rings does not make a marked difference, also it can be seen by doubling the best performing system from the previous figure still only reproduces the same values as before demonstrating the 2x12 configuration provides the optimal de-correlated sound field, however this is contrary to the findings of NHK subjective tests. *See Figure 6.4.*

Taking the best performing system of the 2x12, and plotting this against the current surround standard of the 5.1 system, it can be seen that an almost similar performance is achieved with a much reduced number of speakers *See Figure 6.5*, this confirms a similar result found by Hiyama et al. (2002).

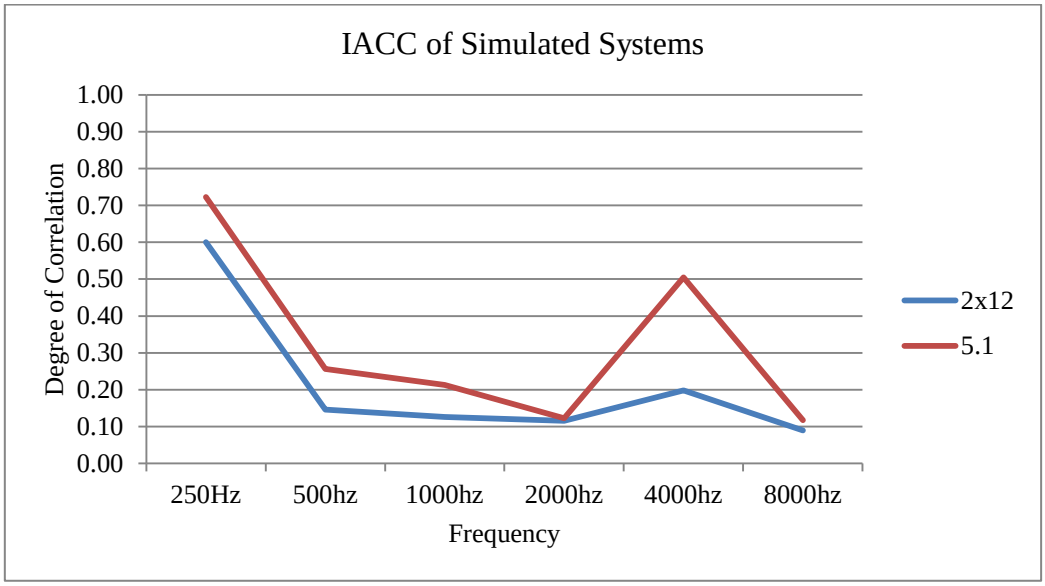


Figure 6.5 IACC of Height system and 5.1 system

6.6 IACC of Real Soundfield

From the initial simulation 8 systems were identified which provided large differences in IACC measures.

Further to the simulated measurements of IACC, measurements of the real soundfield were carried out for each of the systems. A total of twenty four loudspeakers split between twelve at ear level and twelve above the listening position at an elevation of +35° since the stands used would not permit a 40° elevation used in the previous simulation. This would allow the coverage of the 8 systems. For evaluation of the real systems due to resources it was not possible to configure all of the systems assessed in the objective evaluation, therefore only

three of the systems used by NHK could be configured, *See Table 6.2* where the systems used by NHK are highlighted in red.

Array	Number of Channels Horizontal	Number of Channels Vertical
Mono	1	0
Stereo	2	0
ITU 5.1	5	0
Six	6	0
2x6	6	6
9.1	5	4
Twelve	12	0
2xTwelve	12	12

Table 6.2 Speaker arrays used for real soundfield measurement

Mono: Speaker placed 0°

Stereo: Two speakers at ±30°

ITU 5.1: Two speakers at ±30° one at 0° and two at 110°

Six: Six speakers dispersed at even angles around 360° starting at 0°

2x6: Six speakers dispersed at even angles around 360° starting at 0° and another ring above the listener at +35° elevation

9.1: ITU 5.1 layout with four speakers placed above the two stereo frontal channels and the two surround channels

Twelve: twelve speakers dispersed at even angles around 360° starting at 0°

2xTwelve: twelve speakers dispersed at even angles around 360° starting at 0° and another ring of twelve above the listener at +35°

The speaker array was set up in a semi anechoic chamber *See Figure 6.6*, the anechoic chamber conforms to the ISO 3744, ISO 3745, and BS 4196 standards.

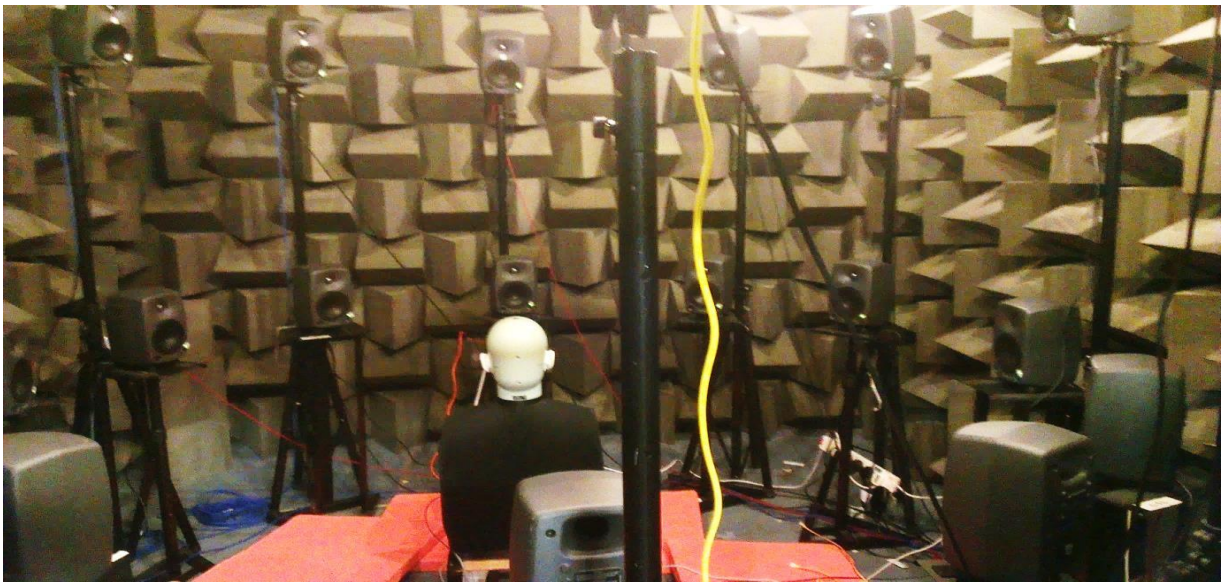


Figure 6.6 Speaker arrays used for IACC measurement

Firstly each of the individual speakers were calibrated by driving them with pink noise of -20dB so that they produced 70dB (A) at the listening position. Then the 8 individual systems were level aligned by measuring the output of each system using a sound level meter at the listening position so that they produced 70.1dB(A). Each of the speakers on the horizontal plane was located with a radius of 1.3m from the listening position. However the elevated channels were at a larger distance of 1.7m, resulting in a difference of .4m, therefore this would cause a difference in time of arrival at the listening position when the two layers were used. This was corrected by introducing a delay in the lower channels of 1.2 ms when the two layers played.

Each of the speakers belonging to the arrays were then driven with de-correlated pink noise of 30 seconds duration which was subsequently recorded using a HATS (Head and Torso Simulator) *See Figure 6.6* this is different than the recommendations of the ISO document (Standards, 2000) where it is recommended for room acoustics to use impulse responses for the measurement, however the IACC has been used by others to measure noise type signals reproduced over loudspeakers (Tohyama and Suzuki, 1989). Further, it was intended to use the same signal for the measurement and the subjective test. Finally, the resultant binaural wav was then analysed using the previously detailed IACC measurement

6.7 Objective Measures Results

Looking at the linear values across all frequencies it can be seen that there are larger than the just noticeable differences for listeners which is 0.075 +- 0.008 as measured by Cox et al.

(1993) *See Table 6.3*. This would mean that utilising these systems in a subjective test should show differences in listener's subjective scores in terms of envelopment.

System	Linear IACC (Simulated)	Linear IACC (Real)
Mono	1	0.9
Stereo	0.39	0.27
I.T.U-5.1	0.17	0.48
Six	0.37	0.29
Twelve	0.1	0.17
2xSix	0.36	0.3
2xTwelve	0.10	0.19
9.1	0.12	0.2

Table 6.3 IACC of real reproduction and simulated reproduction systems

6.8 Subjective Test

The previously described systems in Table 6.3 used for the objective measurements were then used in a subjective test.

6.9 Experimental Design

Twenty four loudspeakers split between twelve at ear level and twelve above the listener at an elevation of 35° were set up this would allow the coverage of the previously detailed systems. Each of the speakers on the horizontal were located with a radius of 1.3m from the listening position, however the elevated channels were at a larger distance of 1.7m, resulting in a difference of .4m, therefore this would cause a difference in time of arrival at the listening position when the two layers were used. This was corrected by introducing a delay in the lower channels of 1.2m/sec when the two layers played.

The same semi anechoic chamber as before was used which conforms to the ISO 3744, ISO 3745, and BS 4196 standards.

An acoustically transparent but visually opaque curtain was hung between the listener and the speaker array to negate any visual clues.

6.10 Stimuli

Pink noise was chosen over white noise, not only to make the test easier on the listener but also because of the low frequency content, which previously stated is said to enhance envelopment. The pink noise samples were made to be 30 seconds in duration and looped.

Listeners were asked to rate the perceived envelopment of the various systems using a test interface which was given to the participants on an I-Pad connected to the replay computer via a wireless network. This had eight sliders for each of the systems based on a modified MUSHRA scheme, with the extremes of each slider being no envelopment and most envelopment, in order to avoid quantization bias, which is caused by a clustering of scores around the tick marks of the slider scale (Zielinski et al., 2007), the tick marks were removed. Listeners could switch instantly between systems without glitches, allowing fast comparisons between height and horizontal systems, this is said to allow small differences to be detected between samples (Bech, 1990). Participants scored each system three times and the test was fully randomized for each participant. Two systems were also used to provide low and high anchors, the low anchor was the mono system and the high anchor was the 5.1 system.

6.11 System Levels

In order to negate any level differences between systems, each systems output was measured at the listening position using a Bruel and Kjaer 2238 sound level meter to produce a level of 71.8dB(A) this was chosen as a comfortable listening level, since it was found by (Airo et al., 1996) that most comfortable listening levels ranged between 52 and 88dB(A) and that 69dB(A) was the average most comfortable listening level for music.

6.11.1 Evaluation of Envelopment

In evaluating the attribute envelopment it can be difficult to ensure that each participant is judging the soundfield in the same way and it has been detailed by Berg (2009) that numerous researchers have provided different definitions of envelopment.

6.11.2 Training

Previous tests by others have utilised different methods to help ensure that each listener is responding to the correct attribute, this has usually taken the form of a short training session to direct the listener's attention. One such method used by George et al. (2010) provided listeners with a statement of the attribute and also provided them with an example of an enveloping sound source and one that was not enveloping, this was achieved by utilising one instance of an applause recording and delaying five instances of it by 500 milliseconds and

distributing it to each of the channels of a surround system, the non-enveloping sample was simply a mono source played from the centre channel.

This method was followed however this did not provide a realistic soundfield. Therefore an alternative method was sought by taking a 10 sec sample of applause and making 5 edits of different sections and discretely feeding each of these different edits to the speakers in the 5.1 system, this would simulate non-correlated applause coming from different directions, with a mono version fed to the centre channel creating a non-enveloping demonstration.

6.11.3 Test Procedure

Firstly before participants took the test they were asked for their understanding of the attribute envelopment, some of these responses were as follows

- *Like you are involved in a sound scene*
- *The sound wraps around you*
- *Immersed in the sound scene*
- *The feeling of being surrounded or involved in the sound scene*
- *Hearing sound coming from all around you and not being able to localise it in any one direction*

It would seem that there was some consensus on the understanding of envelopment between participants.

After this part a statement from George et al. (2010) was used explaining the term to each listener to reinforce what was being asked of them:

'*Envelopment is a subjective attribute of audio quality that accounts for the enveloping nature of the sound, a sound is said to be enveloping if it wraps around the listener*'

Finally after this stage the participants were then led into the semi anechoic chamber where before starting the test proper they were played the previously prepared two samples of applause, finally they were then presented with the test interface with the eight different pink noise samples to evaluate.

Fourteen listeners from a pool of professional acoustic and audio researchers took part that could be considered experienced listeners. The subjective test instructions given to the listener can be seen in *Appendix 5*.

6.12 Subjective Test Results

Participants commented that the majority of the samples had very small differences making it hard to separate them. This was also found by Bech et al. (2008) in their test listeners could not tell the difference in envelopment between the 5 and 4 channel surround systems. Interestingly a common statement was that for some of the samples it felt like you were standing under a waterfall, referring to the systems with elevated channels, this was probably due to the fact that the listeners had no prior awareness of the 3D system. Comments were also made regarding large spectral differences between the systems, Martin (2002) states that an increase in speakers around a listener can cause detrimental effects on timbral characteristics compared to stereo, because of constructive and destructive interference caused by multiple coherent sources with similar amplitudes and the reduced head shadowing. Further, Silzle et al. (2004) also points out that comb filtering is much more audible in the vertical plane than the horizontal plane due to the binaural processing.

It is stated in the MUSHRA test standard (ITU, 2001) that where no anchor points are used each participants data should be normalised, since anchor points were used the data were not normalised prior to statistical analysis.

Initial inspection of the data indicated that as the speaker count increased so did the subjective score, however it can be seen in *Figure 6.7* that the scores did start to plateau above the 5.1 surround system. Looking closer at the spread of the confidence intervals this does show that the systems with the greatest number of channels e.g 2x12, 9.1 and the twelve on the horizontal did receive consistent scores because of the smaller confidence intervals.

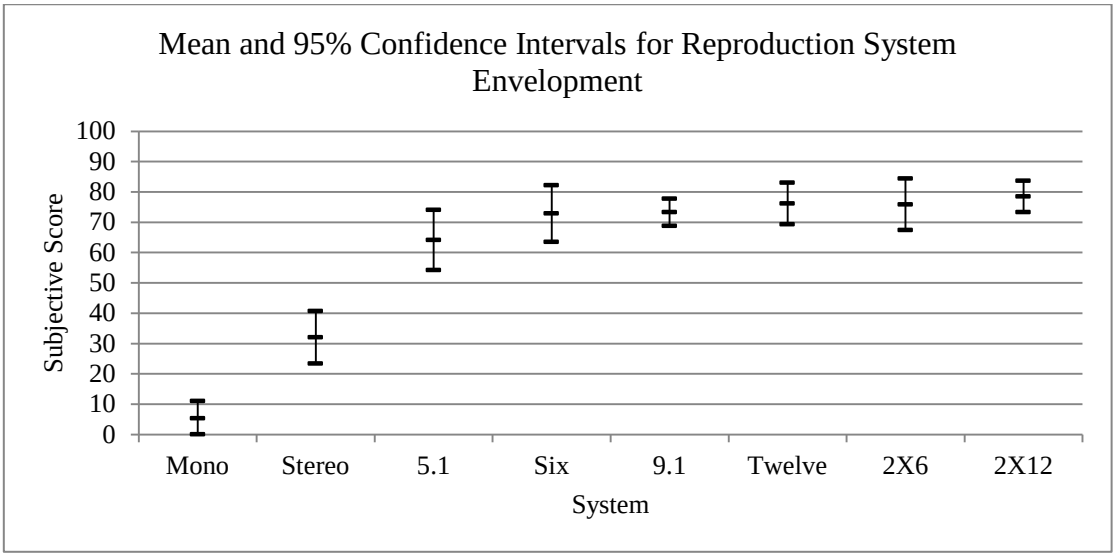


Figure 6.7 Mean and 95% confidence intervals of surround horizontal and with height Systems

Before carrying out an ANOVA to check for significance it was first necessary to check that the data meet the assumptions of the ANOVA, A Mauchly's test of sphericity (Field, 2009) was carried out and this was found to be non-significant indicating the data meet the assumptions of the ANOVA i.e.(the variances are not significantly different).

A one way repeated measures ANOVA was carried out with factor System (Field, 2009), This found significant differences between systems ($p = < .05, \eta^2 = .875$)From the plot of confidence intervals this is suspected to be between the mono and stereo, and surround systems with and without height.

A Bonferoni post hoc test (Field, 2009) revealed that there were significant differences between the mono and stereo systems <0.05 significance and also the mono system and the surround systems 2D and 3D <0.05 significance level. However looking exclusively between the 2D and 3D surround systems there were no significant differences >95% significance level.

To determine how well the IACC predicted the subjective perception of envelopment the linear IACC was plotted against the averaged subjective scores for each system. It can be observed that the as the degree of correlation goes down the subjective score increases, which is also related to the number and positioning of the speakers in each. Further, it can be observed that the trend displayed plateaus around the systems with the greatest number of channels in the horizontal and with height, which are highlighted *See Figure 6.8.*

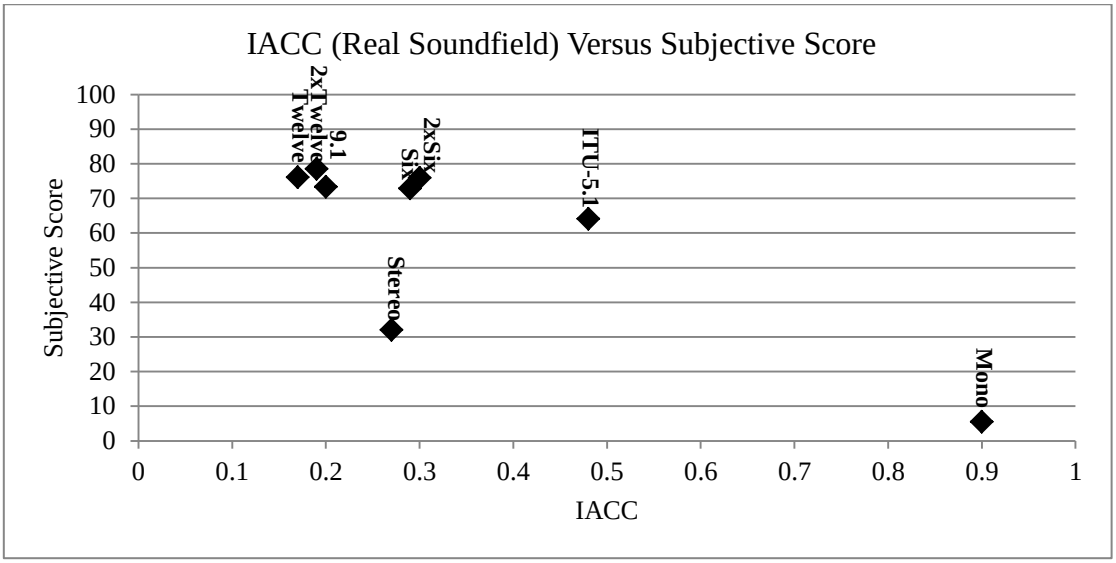


Figure 6.8 Scatter plot of IACC versus subjective score

However there does seem to be the outlier which was the stereo system, this was judged as not being enveloping, however the low IACC could be pointing to the fact that it had a broad soundstage or ASW (apparent source width). This also highlights the fact that participants were focussing on envelopment and not confusing ASW with envelopment.

In order to establish the strength of the relationship between the IACC and the subjective scores a Pearsons correlation (Field, 2009) was carried out for the real and simulated measures, As expected this showed a strong negative correlation *See Table 6.4.* However it can be seen that the simulated soundfield provided a stronger correlation.

Subjective Score Vs Real Soundfield (IACC)	Correlation Coeff -0.81
Subjective Score Vs Simulated Soundfield (IACC)	Correlation Coeff -0.87

Table 6.4 Pearsons correlation coefficient of subjective score versus IACC for real and simulated soundfields

6.13 Discussion

From the outset the objective measures of the simulated soundfield and of the real soundfield showed that three systems, the 9.1, the twelve speakers on the horizontal and the system with twelve speakers on the horizontal and twelve above the listener produced the lowest IACC. Linking this to the subjective test it was observed that these systems received the most consistent scores observing the smaller confidence intervals *See Figure 6.7.* However looking at the mean score for all of the height systems this was similar.

A basic IACC measure was used to asses a number of simulated and real surround systems. It was not the objective in this investigation to evaluate a number of variants of IACC, however higher correlations between the IACC and the subjective measures may have been found with more refined variants of IACC like the perceptually grouped IACC developed by Mason (2002) which separated the measurement of musical signals into the source related and environment related parts of the signal based on perceptual streaming. Another variant of IACC was that described by Conetta (2011) entitled IACC0 where the IACC was calculated based on the dummy head facing forward and the mean IACC was calculated across 22 frequency bands from 150Hz-10kHz, this measure was also paired with the IACC90, which was the same calculation as the previous but with the dummy head facing 90°, the product of IACC0*IACC90 was designed to give a measure of IACC in the median and frontal plane.

In comparing the 2D surround systems with the 3D systems no significant differences were found and it was interesting to see that in the case of the system with six on the horizontal plane and then adding a further ring of six above the listener the scores were identical, this could also be seen for the twelve speaker array. Contrary to this the 9.1 system which utilises the 5.0 array on the horizontal and an additional four channels above the listener showed a bigger improvement than the system with six on the horizontal and six above the listener pointing to the fact that the actual positioning of the elevated speakers is important to reproduce low correlation between channels, as found by Hiyama et al. (2002) for horizontal layouts. It is also interesting to note that NHK in their studies found that an increase in speakers in the horizontal beyond 12 saturated the perception of envelopment causing the scores to plateau, however that an increase in speakers in the vertical continued to improve scores. Yet in this study it was found that an increase in speakers in the vertical did not continue to improve the scores for envelopment.

When considering the training given to listeners the decision was made to use the ITU 5.1 system instead of the 3D system. In retrospect it may have been better to have given listeners training using the 3D system instead, since this is a system which they would not have been familiar with. On the other hand it was interesting to get the listeners reaction to this reproduction format without their prior knowledge, had training been provided using the 3D system this might have influenced the results.

Taking into account the limited aspects of this investigation, which include only the assessment of envelopment, the limited elevation of the height channels and the small number

of systems. Overall in this investigation the improvements height brings over horizontal in terms of the current 5.1 system the improvements are small and more importantly not significantly different.

Some of the systems tested in this evaluation feature large amounts of speakers, therefore it is necessary to appreciate that if wider adoption of 3D systems is to happen a smaller number of speakers is more desirable, especially in the case of domestic use. Therefore in light of this investigation it may be possible when using diffuse sound fields to use a small amount of speakers above a listener. Although, because the minimum number of elevated channels in this test was four it is hard to say where the threshold is before the reproduction is affected.

6.14 Summary

An initial computer simulation of a number of 2D and 3D surround systems was carried out using a decorrelated noise signal as the input in order to measure the IACC of each system. From the initial simulation it was found that the addition height lowers the measured IACC value compared to 2D systems, comparable results were also found for IACC measured from the real soundfield using a similar signal. From this outcome a subjective test was initiated where it was found that there was no significant difference between the 2D and 3D surround systems in terms of the attribute envelopment. Comparable performance in terms of envelopment was found between the 3D system with four height channels and that of the 3D system with twelve height channels.

6.15 Conclusion

A subjective and objective evaluation of 2D and 3D surround systems has been carried out utilising a decorrelated noise signal. Objective measurements showed that the value of IACC was lowered with the addition of height channels. A subjective test was carried out to investigate the perception of envelopment utilising a number of selected 2D and 3D systems. The outcome of the subjective test showed that there was no significant difference between surround systems in terms of the attribute envelopment utilising a decorrelated noise signal.

It could be argued that a decorrelated noise signal does not have a large amount of external validity. The next experiment will investigate the perception of envelopment utilising soundfields more representative of the kind of soundfields encountered in everyday listening situations. The sound scenes will be rendered using 1st order and higher order ambisonics to answer the question:

1. How does the type of sound scene reproduced over a surround with height system affect envelopment?
2. Does higher order ambisonics add to the perception of envelopment?

7 Evaluation of Envelopment in With height and Horizontal Surround Part 2

7.1 Introduction

Previous work to investigate the required number of channels in the vertical to enhance envelopment found that there was no significant difference between the different configurations tested. Further, that comparable performance was found between the system with four height channels and the system with twelve height channels. Since the system with four height channels provided the best balance between number of channels and subjective score, this system has been used in this investigation. In addition this system had also received the most consistent scoring suggesting that listeners were in agreement on the perceived envelopment. The previous study in *Chapter 6* used four height channels as a minimum, therefore it was uncertain if the number of height channels were reduced further this would have an effect on the subjective scores. To investigate further a system using only three height channels was also added this would allow an insight as to whether this would reduce ratings of envelopment.

The previous test had used pink noise as the stimulus, although this being an ideal stimulus with which to assess envelopment, since it is not a complex signal, it is not representative of the type of stimulus which is commonly listened to by the general population. It is also stated by Rumsey (2002) that is more appropriate to use the sort of program material for which the product will be used. Therefore it was decided to extend the previous testing by utilising audio material more representative of the type of material commonly encountered.

7.2 Stimuli

It was desirable to have a large selection of sound scenes for evaluation, however the availability of suitable with height material was difficult to obtain therefore this restricted the choice of sound scenes. In addition it was desirable to use higher order ambisonics although the availability of the higher order microphone or Eigenmike is even less due to its cost, however Technicolor supplied a recording made as part of the FASCINATE project using the Eigenmike which allowed the reproduction of higher order ambisonic sound scenes.

Two of the sound scenes had been recorded in two different indoor spaces. One a small reverberant church and the other the Royal Albert Hall. It was of interest to see if the height

channels added to the perception of the room, since both microphones provide the elevation component therefore capturing the ceiling reflections which would through normal recordings not be picked up separately.

Sound Scene	Description
Football Sound scene	Sound of crowd from all around with tannoy announcements from top front left
Concert Sound scene	Sound of applause from all around with ambience and reflections from hall in height and rear channels
Jazz Ensemble	Musicians spread from left to right in front of the soundfield microphone, room sound from behind and above.

Table 7.1 Sound scenes used in the experiment

7.3 Ambisonics

In order to render the sound scenes over the different systems ambisonics was used. Two of the sound scenes had been recorded using the Soundfield microphone, whilst the applause recording had been made using the Eigenmike which is detailed in *Chapter 3, Section 3.5.4.*

All of the sound scenes were rendered in 1st order, however the Eigenmike allows higher order recording upto 4th order therefore in addition some systems also allowed 3rd order renderings.

7.4 Speaker Arrays

MONO- Centre channel at 0°

STEREO- Speakers displaced at ±30°

5.1- Three front channels at 0°, ±30° and two rear channels at ± 110°

OCTAGON- Eight speakers equally dispersed at 45° intervals starting at 0°

3_OCT_3- Eight speakers equally dispersed at 45° intervals starting at 0° with three channels above and below starting at 60° and dispersed at 120° intervals and an elevation of ±35°

4_OCT_4- Eight speakers equally dispersed at 45° intervals starting at 0° with four channels above and below starting at 45° and dispersed at 90° intervals and an elevation of ±35°

7.5 Decoding

The mono system was fed with the W component, whilst the stimuli for the octagon configuration was decoded using shelf filtering above and below 400Hz.

The remaining three systems were not regular with respect to ambisonic theory.

These were:

- 5.1
- 4_OCT_4
- 3_OCT_3

Therefore optimised decoders had to be used to ensure correct soundfield reconstruction. For the 5.1 system an optimised matrix was used called ITU-5, which is detailed in Heller et al. (2010). Of the decoders created it was detailed by Heller et al. (2010) that the ITU-5 decode provided the best sense of envelopment. For the two height systems the T designs method proposed by Kaiser (2011) was used which is included in the ambisonic toolkit by Heller et al. (2012). Again these systems were also decoded using dual band shelf filtering above and below 400Hz. *Figures 7.1, 7.2 and 7.3* show the uniformity of the velocity and energy vectors of each decode for the 3D arrays.

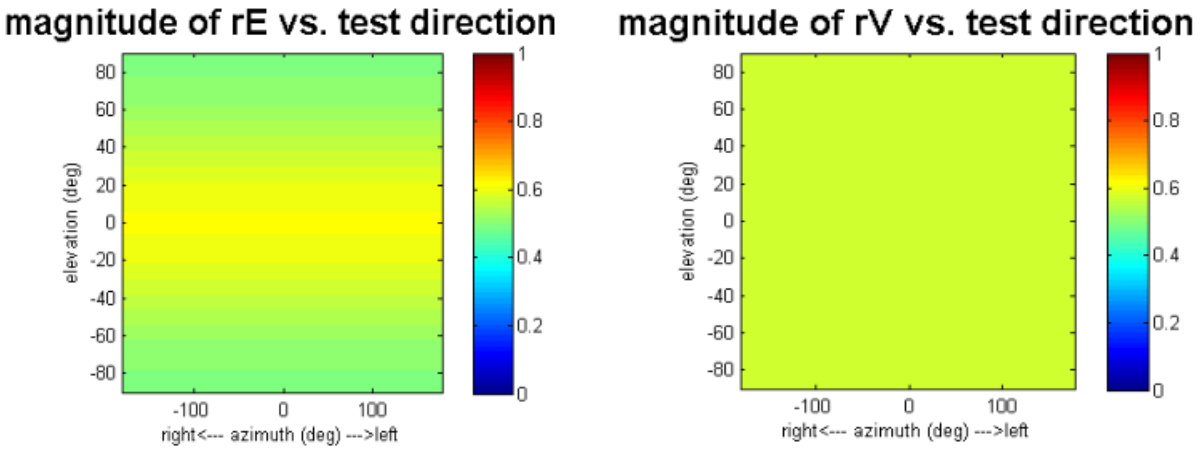


Figure 7.1 Uniformity of Energy and Velocity for 1st Order Decode 4_OCT_4

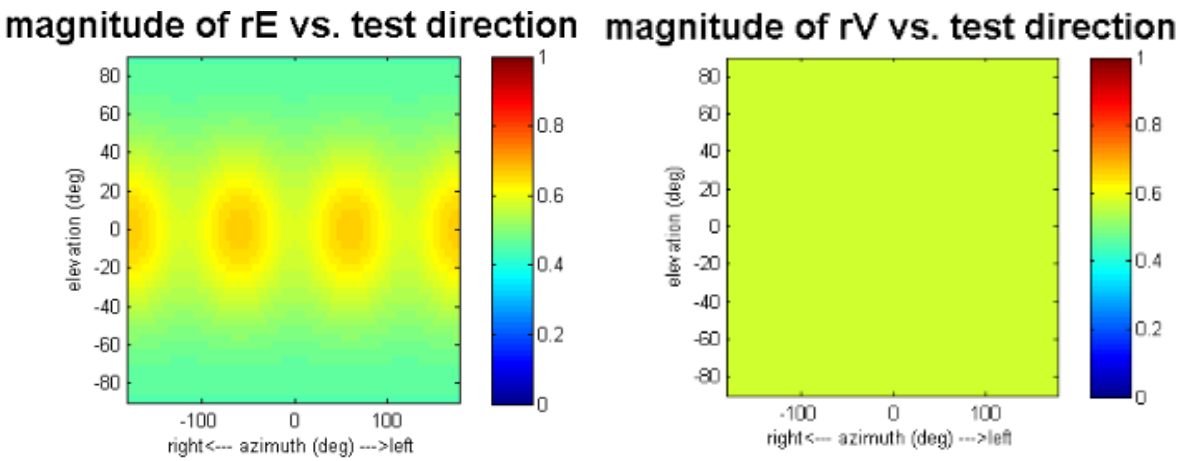


Figure 7.2 Uniformity of Energy and Velocity for 1st Order Decode 3_OCT_3

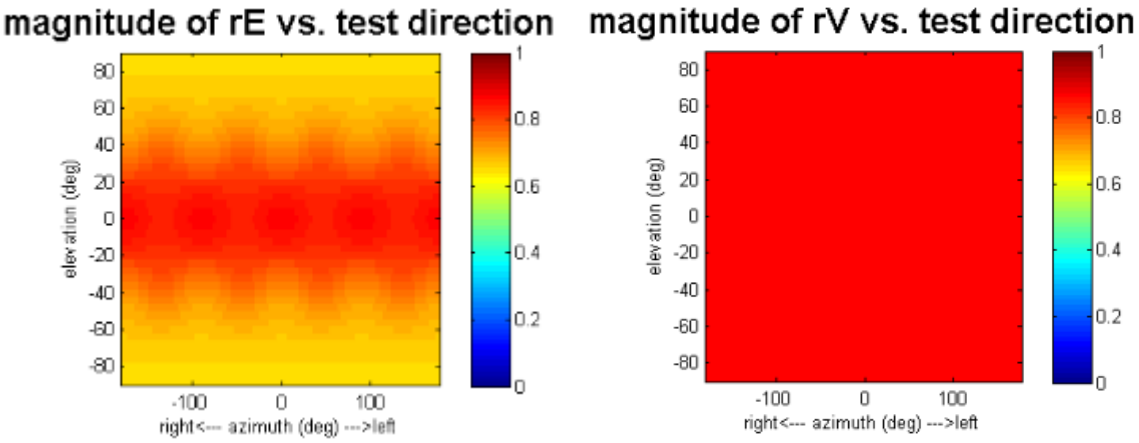


Figure 7.3Uniformity of Energy and Velocity for 3rd Order horizontal 2nd Order Height Order Decode 4_OCT_4

A total of twenty seven speakers were used which covered the desired systems for assessment.

One of the height systems used a mixed order decode (Travis, 2009) where the horizontal order was 3 whilst the vertical order was 2. The *Table 7.2.* below shows the breakdown of the systems.

System	Stimulus Music	Stimulus Atmos (Football)	Stimulus Applause (Proms)
Mono	W component	W component	W component
Stereo	Blumlein Pair	Blumlein Pair	Blumlein Pair
5.1	1st order	1st order	1st order
8 Speaker Pantophonic	1st Order	1st Order	1st Order and 3rd Order
14 Speaker Periphonic	1st Order	1st Order	1st Order
16 Speaker Periphonic	1st Order	1st Order	1st Order and 3rdH/2ndV

Table 7.2 Speaker configurations and ambisonic orders used in experiment

7.6 Test Environment

The test was conducted in the semi anechoic chamber used which conforms to the ISO 3744, ISO 3745, and BS 4196 standards.

7.7 Array Dimensions

The radius of the horizontal arrays was 1.35m, however the height stands for the elevated speakers cannot occupy the same position so the height speakers was at a slightly larger radius 1.7m. Therefore a delay of 1.2m/secs was inserted into the horizontal channels in order to insure that the sounds arrive at the listening position at the same time.

7.8 Calibration of individual Speakers

Each of the individual speakers was calibrated to a level of 71.8dB(A) by using a Bruel and Kjaer 2238 sound level meter, similar to the previous envelopment test by driving each speaker with pink noise of -20dB and measuring the resultant output at the listening position. to produce a level of 71.8dB(A). This was chosen as a comfortable listening level, since it was found by Airo et al. (1996) that most comfortable listening levels ranged between 52 and 88dB(A) and that 69dB(A) was the average most comfortable listening level for music.

7.9 Level Alignment between Systems

Subjective level alignment was carried out to minimise level differences between systems, this method has been used before by Zacharov (2000) and Hirst (2006), it has also been found by Zacharov (2000) that listeners are capable of very accurate level alignment with low error, in addition subjective loudness data has been used previously to validate the accuracy of loudness meters (Crockett et al., 2004). The subjective level alignment was carried out by

utilising a pre-test where two listeners were asked to compare the level of the mono centre channel with a signal input at (0db) to each of the multichannel systems in turn and adjusting the multichannel system to be of the same relative level to that of the mono centre channel. This is crucial since envelopment can be influenced by the playback level (Soulodre et al.,

2003) The input level to each system can be seen below in *Table7.3.*

Mono	0dB
Stereo	-1.18dB
5.1	-8.32dB
Oct	-10.32dB
4_Oct_4	-15.6dB
3_Oct_3	-13.2dB

Table 7.3 Input levels of the systems used relative to the centre channel at 0dB

7.10 Objective Measure

Using a Bruel & Kjaer head and torso simulator set up at the listening position and connected to a laptop via an RME audio interface, recordings were made of each of the different stimuli reproduced over the six different systems in turn. No time windowing or filtering was applied to the IACC.

It can be seen from the review of measurements *See Table 7.4* that the different sound scenes provide a wide range of values, which is also true looking across the different systems. The differences are greater than the just noticeable difference limen found by Cox et al. (1993) and so differences may be audible to listeners. To investigate this a subjective test was carried out.

System	Linear IACC (Proms)	Linear IACC (Football)	Linear IACC (Music)
Mono	0.95	0.97	0.97
Stereo	0.29	0.95	0.33
5.1 1st order	0.32	0.8	0.36
8 Horizontal 1st order	0.22	0.72	0.33
8 Horizontal 3rd Order	0.24		
16 3D 1st order	0.2	0.44	0.37
16 3D 3rd Order	0.2		
14 3D 1st order	0.3	0.64	0.72

Table 7.4 IACC Values for Each System and Sound-Scene

7.11 Evaluating Envelopment

Outlined in the previous envelopment test it is difficult to ensure that each participant is judging the soundfield in the same way. Therefore as before some training steps were provided to each listener which followed the same steps which were:

1. Asking each listener their understanding of the term envelopment
2. A written statement of the attribute
3. Two aural examples demonstrating a non-enveloping source and an enveloping source

However the previous aural examples used an applause signal as the aural training source. Since this test featured an applause signal as one of the test stimuli, the training example for this current test utilised a pink noise source. A different version of the pink noise signal was fed to each of the speakers in the 5.1 system creating an enveloping soundfield. For the non-enveloping source pink noise signal was fed to the single mono channel.

The same test procedure was used as outlined in *Chapter 6*, *Section 6.11.3* where each participant was asked for their understanding of the attribute 'envelopment'. Then each participant was given written instructions on the task ahead and again also provided with a written statement of envelopment from George et al. (2010)

After this listeners were then led into the semi anechoic chamber where they were played the previously described aural examples. A total of 16 listeners took part. The test instructions given to the listener can be seen in *Appendix 6.*

The test protocol was based on a modified MUSHRA scheme, two low and high anchors were included, the low anchor was the mono system and the high anchor was the 5.1 system. Listeners were provided with an ipad which was connected to the playback computer using a wireless network, each sample had a playback button and slider for scoring which started at 0 for no envelopment and 100 for most envelopment. This allowed them to score and playback audio samples, allowing switching between samples synchronously, therefore allowing small differences to be detected and scored (Bech, 1990).

7.12 Results

It is stated in the MUSHRA test standard (ITU, 2001) that where no anchor points are used each participants data should be normalised, since anchor points were used the data were not normalised prior to statistical analysis.

Most listeners commented that the test was quite challenging and that the differences were subtle. Before conducting analysis an error was calculated between the listeners first and second answer and because of the difficulty of the test participants with a higher than 20% error were rejected, this amounted in two listeners being excluded.

7.12.1 Football Atmos

It can be seen that for the two height systems they have a higher mean score than that of the two horizontal surround systems. However, certainly compared to the horizontal surround systems the overlapping confidence intervals suggest there is not a significant difference. *See Figure 7.4.*

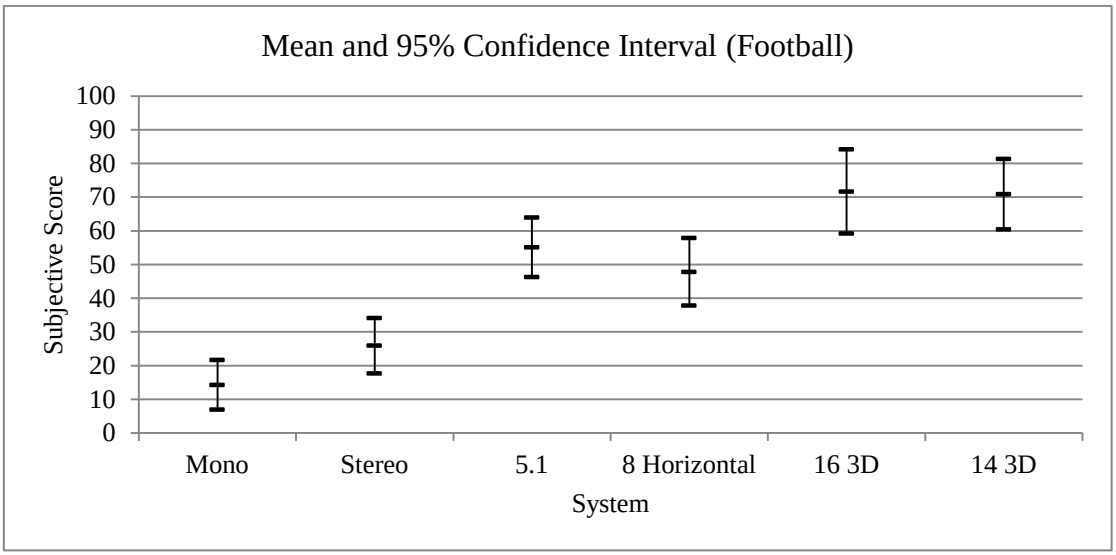


Figure 7.4 Mean and 95% confidence intervals for football sample

7.12.2 Proms Applause

Looking at the results for the proms applause, there seems to be an increase in the subjective score until the eight speaker octagon system where the scores seem to plateau. Further, it can be seen that the eight speaker octagon which included a 3rd order rendering of the applause received the highest mean score. *See Figure 7.5.*

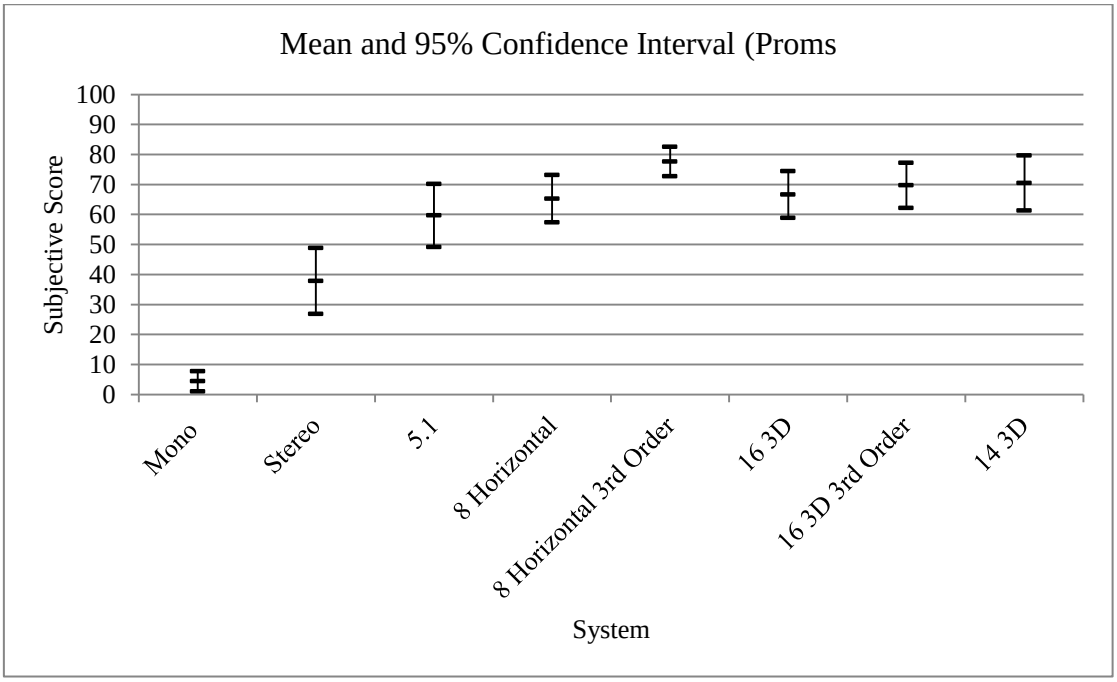


Figure 7.5 Mean and 95% confidence intervals for proms applause sample

7.12.3 Music

A similar trend is also noticed for the music stimulus, although the height system with only three height channels has received a much lower mean score. Further, the height system with four channels has the highest mean score and the least variance. *See Figure 7.6.*

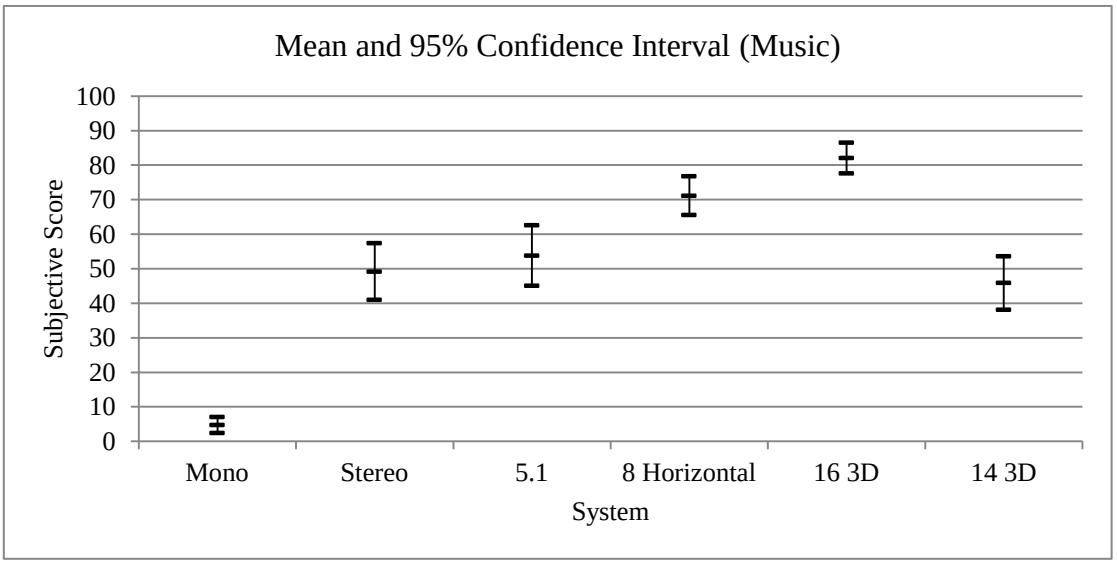


Figure 7.6 Mean and 95% confidence intervals for music sample

To investigate if any of the results were significant a one way repeated measures ANOVA has been carried out with the dependent variable ‘envelopment’ and the independent variable ‘system’.

7.13 ANOVA Analysis

A two way ANOVA was carried out for the factors System and Stimuli. To increase the power of the ANOVA analysis the scores were pooled across all of the sample and systems. To check that the data meet the assumptions of the ANOVA, Mauchlys test of sphericity was carried out and was found to be non-significant, indicating that the variances between the groups were not different.

It was found that the factor Stimuli was not significant($\rho => 0.05, partial\ \eta^2 = .052$) however the factor System was significant($\rho =< 0.05, partial\ \eta^2 = .89$) and the interaction System*Stimuli was significant($\rho =< 0.05, partial\ \eta^2 = .42$).

Following up the ANOVA with a post-hoc test using a Bonferroni correction it was found that the 16_3D system was significantly different to that of all the other systems with the exception of the 8 horizontal system.

Investigating the interaction between System*Stimuli, an ANOVA was run for each sound scene. Firstly to check the data meet the assumptions of the ANOVA Mauchlys test of sphericity was carried out for each of the samples with a significant value indicating the data are not normally distributed. For the football sample the data meet the assumptions of the ANOVA as the test of sphericity was non-significant for this sample. However for the football atmos and the music sample this was not the case so a Greenhouse Geisser correction was applied (Field, 2009).

A Type III sum of squares was carried out between systems (Field, 2009), This found significant differences between systems F=63.775 P=0.000 for the football sample. This was assumed to be between the periphonic systems and the horizontal systems and as would be expected between all systems and the mono presentation. A follow up Bonferroni post hoc test was carried out to find where the differences lay. As expected there was a significant difference between the horizontal surround systems and the two height systems <0.05 significance.

For the proms sample as stated sphericity was violated $x^2\ (27) = 69.4 < 0.05$ therefore the degrees of freedom was corrected using the Greenhouse Geisser estimate of sphericity at .43. A Bonferroni post hoc test showed that the periphonic systems were not significantly different in terms of envelopment at the >0.05 significance level to that of the horizontal surround systems.

For the music sample again sphericity was not met $x^2\ (14) = 26.3 < 0.05$ therefore the degrees of freedom was corrected using the Greenhouse Geisser estimate of sphericity at .67. A Bonferroni post hoc test showed that the periphonic system with four elevated channels was significantly different to the horizontal surround systems at the <0.05 significance level. However, the periphonic system with three elevated channels was not significantly different to the stereo or 5.1 system at >0.05 significance. But the 8 speaker pantophonic system was significantly more enveloping than the periphonic system with three height channels at<0.05 significance.

7.14 Discussion

A number of horizontal 2D and 3D surround systems have been assessed. Firstly, looking at the results for the objective measures has shown that height systems do provide slightly more de-correlated signals, this can be seen for the football stimuli where for the horizontal surround systems the IACC values are high >.72 but that the addition of height channels has decreased the IACC .The IACC has shown some success in predicting listener's scores, although care has to be taken when evaluating the IACC when considering envelopment, since the values for two of the sound scenes using the stereo set up have provided similar IACC values as the surround systems. One reason for this could be because the stereo system utilized a Blumlein coincident technique synthesized from the Soundfield microphone recordings and it has been explained by Wells (2010) that the rear lobes of the figure of eight patterns in a Blumlein technique are out of phase, which can cause a spread of the captured room response (reverberant information) outside of the speakers, providing a greater source width, since it is not possible to envelop the listener with only two channels in a semi anechoic environment.

However, looking at the subjective score between the periphonic system with three height channels and that of the system with four height channels, the mean scores are nearly identical.

The result for the proms excerpt was also interesting since there was no significant difference between the horizontal and with height systems. One reason for this could be because of the type of signal, since it was a random applause signal with a low IACC, it may have been harder for the listeners to discern between the reproduction systems, since it has been found by Santala and Pulkki (2009) that the finest detail a human can perceive is 15° to 30° for sources on the horizontal when the sound source is de-correlated and reproduced over many speakers. Further the increase in resolution did provide more consistent scoring from the listeners as can be seen from the confidence intervals for the 3rd order 8 channel ambisonic system, although this was not significantly different to the other surround systems 2D and 3D surround.

Again the music stimuli showed a trend of increasing the number of reproduction channels which in turn showed the subjective score increased. It was also observed that the periphonic system with four elevated channels received the most consistent scores, which can be seen from the tight confidence intervals. Further, this system was also significantly more enveloping than that of the other surround systems 2D and 3D. Yet the periphonic system with the three elevated height channels performed poorly with the music excerpt, this could have been caused by the irregularity of the speaker layout, the plots in *Figure 7.2* of the uniformity of the energy around the soundfield, show that the energy is not uniform around the array, this would cause instability of virtual images. Some subjects had stated that images were unstable when rotating their head and hard to locate, but interestingly for the other excerpts, football sample and proms, listeners were not as critical with these sources for that system, possibly because of previously outlined factors by Santala and Pulkki (2009), or because the music stimulus had clearly distributed foreground sources which were affected with the irregular layout. This would indicate that the listeners were confusing other factors when deciding on the attribute ‘envelopment’.

7.15 Summary

A number of reproduction systems which includes 2D and 3D surround systems have been utilised in order to investigate if the addition of height enhances the attribute envelopment. Three different recorded sound scenes were used in a subjective test where it was found that using two of the sound scenes the football atmos and the music recording the addition of height significantly enhances the perception of envelopment. However, using an applause sample it was found that there was no significant difference between the 2D and 3D systems. In addition the system with three height channels had the lowest mean score of all the

surround systems when using the music sample, although this was not the case for the football atmos and the applause recording, which suggests that for this type of sound source the impairment brought by only using three height channels was not perceived by the listeners. The objective measure IACC generally illustrated that the addition of height reduces the value of IACC.

7.16 Conclusion

Three different sound scenes rendered using 1st order and higher order ambisonics have been used to assess the perception of envelopment of a number of 2D and 3D surround systems. It has been found that utilising two of the sound scenes, the football atmos sample and the jazz band sample the addition of height enhances the perception of envelopment. However, using the applause sample it was found that there was no significant between the 2D and 3D surround systems.

It was documented in the literature review by Bech (1999) and Zacharov (2000) that spatial quality is a multidimensional attribute, therefore the evaluation of 3D surround systems carried out in this work so far using attributes derived from previous research may have limited the information obtained. In order to alleviate this the next chapter details the initialisation of a descriptive analysis which allows a group of listeners to answer in a less restrictive way in order to answer the question:

1. Are there any unique descriptors related to with height surround?

8 Exploring Spatial Perception in 3D Surround Systems

8.1 Introduction

In the previous two *chapters 6* and *7,* the attribute envelopment has been investigated, since it was explained that the perception of envelopment is one of the attributes driving multichannel development (George et al., 2010) . Although the attribute envelopment is an important part of the multichannel listening experience it is considered that utilising only one attribute for the assesment of 3D surround systems limits the information obtained, therefore a more exploratory method was needed.

The question also arises as to whether envelopment is a suitable attribute for use in the assessment of with height surround, since it has been defined as a preferntial attribute for spatial audio reproduction utilising horizontal surround systems.This sentiment is echoed by Oode et al. (2011) in that the terms 'envelopment' and 'width' apply to horrizontal surround and that Oode *et al* proposed that 'spatial uniformity' be used for 3D surround systems.

Further it has also been proposed by Sazdov et al. (2007) that the attribute ‘engulfment’ be used as a unique descriptor for the perception of sound from 3D speaker systems, using the analogy of the sound covering over the listener similar to being engulfed by the sea. Although this descriptor was defined by the experimenters therefore providing a limited and possibly biased result.

It was therefore decided to utilise a more exploratory method called descriptive analysis, this would allow a less restricted response from a group of qualified listeners which would provide a greater understanding of the benefits of 3D surround systems.

8.2 Descriptive analysis

Previous research by others into spatial audio which has included the ITU 5.1 system used today have identified attributes like envelopment, apparent source width and timbre to be important in terms of listener preference. In addition global asessments of the spatial audio quality have also been considered, however global methods of assesment hinder uncovering

any other relevant parts of the listening experience and it is stated by Eisler (1966) that an indication of preference from a listener tells nothing of the importance the listener attaches to the property investigated. Since new 3D reproduction systems are becoming available do the attributes used in 2D surround reproduction allow the capture of information about these new systems or are there any new attributes? Furthermore it is stated in the ITU-R BS 2159-4 (BS.2159-4, 2012) multichannel sound technology in home and broadcasting that the six requirements are:

1. *The directional stability of the frontal sound image should be maintained over the entire higher resolution imagery area. Coincidence of position between sound image and video image also should be maintained over the wide imagery area.*

2. *The sound image should be reproduced in all directions around the listener, including elevation.*

3. *The sensation of three-dimensional spatial impression that augments a sense of reality should be significantly enhanced. This may be achieved by the use of side and/or back, top and/or bottom loudspeakers.*

4. *Exceptional sound quality should be maintained over wider listening area than that*

provided by current 5.1 channel sound system.

5. *Compatibility with the current 5.1 channel sound system specified in Recommendation ITU-R BS.775 should be ensured to an acceptable degree*.

6. *Live recording, mixing and transmission should be possible.*

The above guidelines are very general and therfore do not allow for a more detailed assesment of with height systems.

A small number of researchers have utilised other methods which are not so restrictive to uncover more information about the listening experience.

These methods are known as descriptive analysis and are a family of techniques for the generation of attributes to adequately describe a product succinctly. These methods have their

roots in the food industry and a number of techniques exist to elicit descriptors from participants. These are repertory grid technique, free choice profiling and consensus vocabulary development (Campo et al., 2010). Further more it is stated by Murray et al. (2001) that the descriptive analysis method allows for the determination of the relationship between descriptive terms and consumer preference, allowing the determination of the 'desired composition' allowing for product optimisation.

The repertory grid technique, consensus vocabulary and free choice profiling have been used in audio research previously by Berg and Rumsey (1999b), (Zacharov and Koivuniemi, 2001, Berg and Rumsey, 2000), Berg and Rumsey (2006) and more recently by Lokki et al. (2011), Wankling et al. (2012).

The proposed method to be used in this work utilises descriptive analysis method. This involves several distinct stages.

1. *Development of descriptive attributes*: This involves gathering a group of experienced listeners and using a number of stimuli to elicit responses from them based on their impression of the stimuli.

2. *Analysis of Descriptive Terms*: In the initial stage the experimenter gathers each of the terms from the listeners and counts the frequency of occurance of each word, previous work in audio using elicitation have retained descriptors occuring ≥ 4 times (Wankling et al., 2012) Further redundancy of words can be explored by removing similes, this happens at the focus group stage, without the interaction of the experimenter, where the listening panel decide on which words are similar in order to reduce the number of descriptors.

3. *Focus Group*: Once the frequency of the descriptors has been logged the terms are then taken to a focus group of the original listeners where the listeners task is to reduce the number of terms by exploring redundancy. This is done by grouping words which they decide as being similar with the main objective of reducing the number of descriptors and through this process develop a consensus vocabulary. Finally the

listening panel must reduce the terms to derive scales and descriptions of each scale which can then be used in a subsequent listening test.

8.3 Generation of Descriptive Terms

A number of reproduction systems were utilised from mono to a 3D surround system in order to elicit a full range of descriptors from the listening panel. The 3D system that was chosen is based on the previous experiments, since it has shown the most consistent rating from listeners.

8.3.1 Reproduction Systems

The systems used were:

- **Mono:** speaker at 0°
- **Stereo**: two speakers at ±30°
- **5.0:** two speakers at ±30°, one at 0° and two rear channels at ±110°
- **Octagon**: eight speakers equally dispersed at 45° intervals starting at 0°
- **Oct_Cube**: eight speakers equally dispersed on the horizontal at 45° intervals starting at 0° with four channels above and below starting at 45° and dispersed at 90° intervals and an elevation of ±35°

8.3.2 Test Environment

The test was conducted in a semi anechoic chamber which conforms to the ISO 3744, ISO 3745, and BS 4196 standards.

8.3.3 Array Dimensions

The radius of the horizontal arrays was 1.35m, however the height stands for the elevated speakers cannot occupy the same position so the height speakers were at a slightly larger radius 1.7m. Therefore a delay of 1.2ms was inserted into the horizontal channels in order to insure that the sounds arrive at the listening position at the same time.

8.3.4 Calibration of individual Speakers

Each of the individual speakers was calibrated to a level of 71.8dB(A) by using a Bruel and Kjaer 2238 sound level meter, similar to the previous envelopment test by driving each speaker with pink noise of -20dB and measuring the resultant output at the listening position. 71.8dB(A) was chosen as a comfortable listening level, since it was found by (Airo et al., 1996) that most comfortable listening levels ranged between 52 and 88dB(A) and that 69dB(A) was the average most comfortable listening level for music.

8.3.5 Stimuli

The stimuli below were all recorded in live spaces, the jazz ensemble were recorded in a small reverberant church using a soundfield microphone with the players assembled around the microphone in a semi-circular arc. The remaining two other recordings were recorded using an Eigemike, which has been described previously in *Chapter 3, Section 3.5.4* the applause was recorded during the Proms festival in the Royal Albert hall as part of the FASCINATE project (Thomas et al., 2011), whilst the choir were recorded in the orchestra hall at the University of Detmold[2]. *See table 8.1*. As previously mentioned the availability of 3D surround scenes limited the test material, therefore this is not an exhaustive number of stimuli, but it is hoped by reproducing these over various systems will elicit a full range of descriptors from the listening panel.

Sound Scene	**Description**
Jazz Ensemble	Sound sources around the frontal arc, room ambience from rear and height channels
Applause	Sound of applause from all around with ambience and reflections from hall in height and rear channels (microphone placed among the many standing spectators)
Choir	Choir distributed across the frontal sound stage with ambience from hall in height and rear channels

Table 8.1 Sound sources used in elicitation

8.3.6 Ambisonic Decoding

To decode each of the sound scenes to the relevant systems the optimised decoding scheme described in *Chapter 3, Section 3.5.14* was used. The jazz band sample was recorded with a soundfield microphone therefore could only be reproduced in 1st order. However the applause excerpt and the choir excerpt was recorded using the Eigenmike therefore could be reproduced in 1st and 3rd order where applicable. *See Table 8.2.*

[2] Both of these recordings were provided by Technicolor Hanover.

System	Ambisonic Order	Stimulus 1	Stimulus 2	Stimulus 3
Mono	W Component	Jazz Band	Applause	Choir
Stereo	Blumlein Pair	Jazz Band	Applause	Choir
5.1	1st Order	Jazz Band	Applause	Choir
Octagon	1st Order	Jazz Band	Applause	Choir
Octagon	3rd Order		Applause	Choir
Octagon_Cube	1st Order	Jazz Band	Applause	Choir
Octagon_Cube	3rd Horizontal/ 2nd Vertical		Applause	Choir

Table 8.2 Systems and Decoding Order Used

8.3.7 Level Alignment between Systems

In order to minimise level differences between systems a subjective level alignment was carried out, the same method described in *Chapter 7*, *Section 7.9*. This was carried out by utilising a pre-test where two listeners were asked to compare the level of the mono centre channel with a signal input at 0dB to each of the multichannel systems in turn and adjusting the multichannel system to be of the same relative level to that of the mono centre channel. *See Table 8.3.*

Stereo	-2.8dB
5.0	-7.1dB
Oct_1	-6.9dB
Oct_3	-3dB
Oct_Cube_1st Order	-6.9dB
Oct_Cube_3rd order	-3.8dB

Table 8.3 Gains of respective systems relative to mono centre channel

As before all speakers were screened from view so listeners had no visual cues as to the loudspeaker configurations, using an acoustically transparent but visually opaque curtain.

8.3.8 Collection of Individual terms

In order to collect each of the listeners descriptive terms an interface created in MAX/MSP was used. This was given to the listeners on an iPad which was connected to the playback computer using a wireless network. This allowed listeners to audition and enter descriptive

words for all of the samples. Further, this would mean that there were no acoustic obstructions between the listener and the speakers.

8.3.9 Participants

Ten listeners who were either acoustic researchers or lecturers in the acoustic department took part in the elicitation. These are listeners who have taken part in listening tests before so can be classed as selected assesors (Bech and Zacharov, 2006). In the selection of listeners it is desireable to have listeners who are experienced and can describe what they hear, since naïve listeners may not be able to communicate adequately what they hear. Only some of the listening panel have previously worked in spatial audio or would be expected to be knowledgeable in the spatial audio literature.

8.3.10 Instructions to Participants

Each listener was given a printed instruction sheet describing the nature of the experiment and their task *See Appendix 7*. In this it was described that they should firstly focus on describing the spatial aspects of each sound scene, but also if necessary describe any timbral aspects also. Participants could also ask questions about anything they did not understand.

8.4 Descriptors elicited

A total of 380 descriptors were provided by ten participants. The minimum provided by a single listener across the three different sound samples and systems was 21 with the maximum being 107. However there was a large amount of redundancy in that many terms were used multiple times. In order to reduce these terms only terms that were mentioned with a frequency of four or more times were kept, this is similar to the criteria used by Wankling et al. (2012). The initial set of descriptors can be seen below in *Table 8.4*.

Attribute	Frequency	Attribute	Frequency	Attribute	Frequency	Attribute	Frequency
enveloping	54	distant	10	realistic	6	coloured	4
spacious	29	focussed	10	spread	6	confusing	4
wide	27	natural	9	surrounding	6	depth	4
narrow	23	reverberant	9	closed	5	detached	4
frontal	19	small	8	distributed	5	flat	4
close	14	bright	7	presence	5	forward	4
dull	12	open	7	surrounded	5	front	4
unnatural	12	behind	6	thin	5	pleasant	4
immersive	11	noisy	6	blurred	4	separated	4
diffuse	10	phased	6	boxy	4	unfocussed	4

Table 8.4 Frequency of Attributes

It is interesting to note the most frequent descriptors were *envelopment ,spacious, wide, narrow*. These descriptors have appeared in the work of Berg and Rumsey (1999b), Zacharov and Koivuniemi (2001), Mason (2002), Berg and Rumsey (2006), Choisel and Wickelmaier (2006a), Kim and Martens (2007) it is also stated by Rumsey (2002) that the most useful terms that arise repeatedly in spatial audio are envelopment and source width. In this evaluation the next most common descriptors are related to the listener and the positoning of the source in terms of the use *frontal , close* e.t.c. Other terms are related to the resolution of the reproduced image with descriptors like *focussed, blurred* and other terms related to timbre like *colouration, boxy* and *bright*.

Another of point of interest is that some of the descriptors naturally create opposite constructs, for example *open-closed* and *focussed-unfocussed, bright-dull, narrow-wide* and *unnatural-natural*.

Some terms were used less frequently and were unique to the system with height, these terms were *roomy* or *sense of room*. This suggests that the addition of height conveys to the listener a greater sense of the room or space the recording was made in. Interestingly for the horizontal surround system less frequent terms were *defined, circular* indicating that listeners could percieve a clear boundary around them.

However only tentative conclusions can be drawn at this stage regarding the use of these constructs since some of the uses are unclear, especially when trying to find out if any are unique to the 3D system. Especially with the use of terms like *wide* because are listeners referring to *wide* in the horizontal or vertical?

8.4.1 Horizontal versus Surround with Height

It is of interest to inspect the differences between the attributes elicited between the horizontal surround systems and that of the systems with height. Below in *Table 8.5* the attributes elicited for the horizontal surround systems can be seen.

Attribute	Frequency	Attribute	Frequency
enveloping	28	immersive	5
wide	12	reverberant	5
spacious	11	bright	4
unnatural	8	closed	4
close	7	distant	4
behind	6	focussed	4
diffuse	6	noisy	4
dull	6	phasey	4

Table 8.5 Attributes of >4 given for horizontal surround systems

Below in *Table 8.6* the attribute elicited for the surround system with height can be seen. Which show similar attributes as that found for the horizontal surround systems.

Attribute	Frequency
enveloping	21
spacious	15
wide	10
immersive	6
natural	6
diffuse	4
open	4
realistic	4
surrounding	4

Table 8.6 Attributes of >4 given for surround with height

It is interesting to note that the descriptors used for the horizontal systems were replicated for the system with height, except a few unique descriptors which were *open*, *natural*, *realisitc* and *surrounding*.

It was also noted that *unnatural* was used to describe the horizontal surround system whilst this term was not used for the system with height. Further, the term *natural* occurred as many times for the system with height as *unatural* for the horizontal system. It can also be seen that the descriptors elicited for the height system contained no negative descriptors.

A comparison is given between the descriptors elicited in this study and those elicited by Zacharov and Koivuniemi (2001) and Berg and Rumsey (2002). *See Table 8.7.*

Author	Zacharov and Koivuniemi,2001	Berg and Rumsey,2002	Current Study
	Naturalness	naturalness	Naturalness
	Broadness	Source Width	Width
	Sense of Depth	Source Envelopment	Envelopment
	Sense of Space	Presence	Presence
		Ensemble Width	Spacious
		Room Envelopment	
		Room Width	

Table 8.7 comparison of spatial attributes from other studies with the current study

8.5 Focus Groups

In total three focus groups were held, unfortunately only 6 of the original 10 listeners were able to take part in these meetings not including the panel leader. This panel encompassed teaching staff and post-doctoral researchers involved in acoustics and sound engineering. The focus groups concentrated solely on the spatial aspects due to time constraints.

8.5.1 Focus Group 1

The first focus group meeting lasted for approximatley 2 hours during which the aim of the meeting was to reduce the number of descriptors by exploring terms which had a similar meaning. The list of descriptors =>4 shown in *Table 8.4* were displayed upon a white board for all the group members to see.

The focus was on spatial attributes and it was noted by the group that most terms in the initial word list were in fact relating to spatial attributes, however there were a few terms which related to timbral atributes and also a few miscellaneous words *See Table 8.8 and 8.9* so the intial stage of the first meeting was spent assigning each word to the agreed category. There was some discussion around the descriptor *flat* since it was described that this could relate to the systems that do not include height, however it was finally agreed that in the traditional sense this decriptor would usually be understood as pertaining to the spectral content of a sound source and not to the spatial dimensions.

Other
phased
noisy

Table 8.8 Miscellaneous Descriptors

Timbral
dull
bright
thin
boxy
coloured
flat

Table 8.9 Timbral Attributes

The first set of words that related to spatial atributes that were picked out were frontal, front and forward *See Table 8.4* and it was decided that front and forward were discarded in favour of frontal.

There was much discussion about the term *envelopment* and also that this term had been defined in the realms of concert hall acoustics and could be refering to a two dimensional attribute only.

Attention turned to other similar words like *surrounded* and *surrounding* which it was discussed could be suggesting something subtly different, it was discussed that *surrounded* could refer to being *surrounded* by people for example individual sources and *surrounding* could be referring to the perception of a room, the room is *surrounding* you. In addition the descriptors *blurred, unfocussed* and *confusing* were discussed and could also indirectly be related to *surrounded, surrounding*, since the terms *blurred* and *unfocussed* could be refering to sources coming from every direction hence lead to the term *confusing*. Further, the conversation then led to the terms *frontal, behind, spread, wide* and *narrow* and that the descriptors previously described such as *surrounded, surrounding*, were also related to what the group termed as spatially biased and that these terms could be separated into *foreground* sources and also *background* sources, based on whether the listener was listening to the direct sound or the reverberant or background sound (Griesinger, 1996). Further to this an elicitation carried out by Guastavino et al. (2007) also found that listeners used descriptors based on the spatial distibution of sound. Finally it was also decided by the group that *confusing* was not a helpful descriptor as it was ambigious therefore was eliminated from the list.

It was finally decided after much debate that there was a high level term called *Event Distribution* which was related to how spatially biased the soundfield was in relation to the listener which was realised from the descriptors shown in *Table 8.10.*

Event Distribution
wide
narrow
frontal
behind
spread
surrounding
surrounded
distributed
enveloping

Table 8.10 Words classed as event distribution

Event distribution would also be separated into two scales, one for *foreground distribution* and another for *background distribution*. Since it was discussed previously that listeners could use two methods of analysis, one where they could hone in and listen to clearly percieved sources in the foreground and another where they could just percieve the background sounds.

It was thought that the scale descriptor *event distribution* may not be the most appropriate descriptor for the scale as it may be confused with temporal aspects of the sound scene, however this was not communicated to the focus group so as not to influence the process.

8.5.2 Remaining Terms

After the first meeting the group had been able to reduce the 40 descriptors by redundancy or by eliminating ambigious terms or by assigning terms to the timbral or other categories.

8.5.3 Summary

By the end of the first focus group it had been decided that there was a descriptor entitled *Event distribution* that would be split into two scales one for the *background distribution* and one for the *foreground distribution*

It was agreed that the remaining terms were related to whether the listener is involved in the sound scene or not and in addition how realistic it was based on expectation. These are shown in *Table 8.11*.

Remaining Descriptors
spacious
close
realistic
immersive
distant
focussed
small
presence
depth
detached
Open
Closed

Table 8.11 the remaining words left after the first focus group session

8.5.4 Focus Group Session No2

The second focus group lasted for 1hour and 30 minutes and the start of the meeting was spent reviewing the previous meetings outcome and that everyone was still in agreement.

Disscusion was then turned to the remaining descriptors. *See Table 8.11.* It was discussed by the group that two further categories could be derivied from the remaining set of words the first category was *realism* and the second was *physical size*.

There was much debate over certain descriptors and which category they belonged to, for example the word *detached*, but after much debate it was decided that this belonged in the *realism* category because it was agreed by the group that this referred to being *detached* from the sound scene similar to other words *immersed* and *presence*, with *detached* being the antonym of *immersed* and *presence*, therefore a category entitled *realism* was agreed on and contained the words shown in *Table 8.12.* But there was much debate over the descriptor *realism* and another term was put forward to cover this called *plausibility*, however it was argued that *realism* was a better descriptor. The example was given that people would think

well I am in a semi anechoic chamber with a curtain surrounding me therefore this is not *plausible*. Therefore it was decided that *realism* was more suitable.

Realism
realistic
unnatural
natural

Table 8.12 Words classed as relating to realism

The second classification related to the physical size of the reproduced sound scene and included the terms shown in *Table 8.13.*

Physical Size
spacious
close
distant
small
depth
open
closed

Table 8.13 Words classed as relating to physical size of the sound scene

8.5.5 Final Focus Group Session No3

The final focus group lasted for 1hr, the objective of this meeting was to finalise the descriptors to be used in the subsequent listening test.

8.5.6 Event Distribution

The initial stage of the final focus group was used to make sure that everyone was still in agreement with the previous meetings outcome. However there were some concerns raised over the term *event distribution* since this could be confused with temporal aspects of the sound scene, one example was given relating to the timings of events happening within the soundfield like a bell rings at 10am, a person walks across the front at 10.01am, therefore it was agreed that event distribution be changed to *Spatial distribution* to avoid confusion.

It was agreed that the category *Realism* itself would be a scale and would have ends of scale *unrealistic* and *realistic See Table 8.12*. Further it was also agreed that there should also be a scale of *Presence* which would have end points *detached* and *immersed*. The question was also raised as to whether there was much independence between *Presence* and *Realism*, however it was discussed that there was not enough information to make that judgement. *See Table 8.14* for descriptors relating to presence.

Presence
Immersed
Detached

Table 8.14 Descriptors Relating to Presence

There was also further discussion on the terms *surrounded, surrounding* and *enveloping* and if there was any relationship to the term *presence*, although it was argued that you could be *surrounded* or *enveloped* but still not feel *immersed* or a sense of *presence*.

The category of physical size would have a scale called *size* and would have ends of scale *big* and *small*.

Finally as previously mentioned the category *Event distribution* was changed to *Spatial distribution* and had two scales one for the *foreground distribution* and one for the *background distribution*, with ends of scale being *biased* and *uniform*. Where the term *biased* meant that the sound field was coming from a particular direction, for example a mono source directly in front of the listener. Whilst *uniform* meant that the soundfield was equally spread all around the listener.

8.6 Summary

It was highlighted by the group that there had been no obvious terms related to the ‘height’ aspect, which was one of the objectives of the descriptive analysis, but it was discussed that the reproduction of height could indirectly influence already known attributes in surround sound reproduction.

In the end five scales were agreed on which related to spatial aspects of the reproduction system, these scales were derived from the initial 574 descriptors gathered at the elicitation stage.

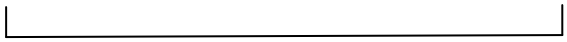
Each of the descriptors and the question for the each listeners are shown below.

8.6.1 Foreground Spatial Distribution

Foreground sources are described as being: mainly close and clearly percieved audio sources) (Zielinski et al., 2003)

Question: How evenly is the foreground sound distributed spatially?

Ends of Scale: Biased Uniform



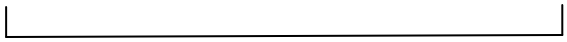
Where the term biased meant that the sound field was comming from a particular direction, for example a mono source directly in front of the listener. Whilst uniform meant that the soundfield was equally spread all around the listener.

8.6.2 Background Spatial Distribution

Background sources are described as being: room response, reverberant sounds, unclear, "foggy".(Zielinski et al., 2003)

Question: How evenly is the background sound distributed spatially?

Endsof scale: Biased Uniform

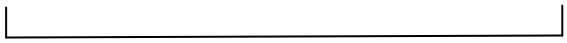


Where the term biased meant that the sound field was comming from a particular direction, for example a mono source directly in front of the listener. Whilst uniform meant that the soundfield was equally spread all around the listener.

8.6.3 Presence

Question: Do you feel part of the sound scene?

Ends of scale: Detached Immersed



8.6.4 Realism

Question: How well does the reproduction match your expectations

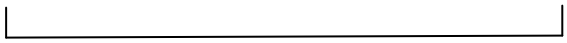
Ends of Scale: Unrealistic Realistic

8.6.5 Size

Question: Whats your impression of the physical size of the soundscene? Are there any of the reproductions which allows you to interpret better the size of the reproduced space.

Ends of Scale:

Small Large



Using the descriptors agreed upon by the panel the reproduction systems could be explored further. This would allow the evaluation of a number of different systems and also the ability to evaluate whether non panel members could use and understand the descriptors.

8.7 Summary

A number of reproduction systems which included 2D and 3D surround systems have been utilised to reproduce three different types of sound scene, this was in order to elicit a number of descriptors from a panel of experienced listeners. A total of 380 descriptors were realised from the elicitation stage. Following the elicitation three separate focus groups reduced the 380 descriptors to five attribute scales relating to the spatial aspects of multichannel reproduction.

9 Perception of 3D Surround Systems

9.1 Introduction

In the previous chapter it was highlighted that current research into the perception of spatial audio has until now been concerned with 2D systems and the attributes derived have been using these systems. In an attempt to discover if there were any unique attributes in surround systems with height listening experience an elicitation followed by three focus group discussions was carried out. A large number of descriptors were reduced to five attribute scales. In this chapter a subjective evaluation is detailed which used the descriptors elicited. The objective of this was to:

- Uncover if there were any unique attributes which are particular to with height surround system listening experience using the derived scales.
- Allow the perceptual measurement of these systems.
- Determine if listeners could understand and use the scales to assess a number of reproduction systems.

9.1.1 Systems

The systems outlined in *Section 8.3.1* were used and the ambisonic decoding outlined in *Section 8.3.6* was used. However in addition four systems were also added which used VBAP rendering, these were not used in the elicitation stage so would provide a further comparison of spatial rendering methods. These were mono, stereo, 5.0 (the ITU arrangement without a sub) and also a 9.0 system this included a 5.0 system with an additional four speakers above the two front channels left and right and above the two rear channels at an elevation of 60°.

9.1.2 Test Environment

The same test environment was used as described in *Section 8.3.2.*

9.1.3 Array Dimensions

The radius of the horizontal arrays was 1.35m, however the height stands for the elevated speakers cannot occupy the same position so the height speakers were at a slightly larger radius 1.7m. Therefore a delay of 1.2ms was inserted into the horizontal channels in order to align them with the height channels.

For the 9.0 system the height stands were at a higher elevation of 60° which meant these stands were at a greater distance of 2.2m from the listening position. A delay of 2.8ms was inserted into the horizontal channels to align them with the height channels to align them with the height channels.

9.1.4 Calibration of Individual Speakers

Each of the individual speakers was calibrated by driving each speaker with pink noise of -20dB and measuring the resultant output at the listening position to produce a level of 71.8dB(A). This was chosen as a comfortable listening level, since it was found by Airo et al. (1996) that most comfortable listening levels ranged between 52 and 88dB(A) and that 69dB(A) was the average most comfortable listening level for music.

9.1.5 Stimuli

Two of the sound scenes outlined in *Chapter 8*, *Section 8.3.5* were used, these were the applause recording and the choir recording which were reproduced using ambisonics. The sound-scenes and ambisonic decoding used for each system can be seen in *Table 8.2* in *Chapter 8.*

In addition two pieces of audio were provided by the BBC R&D department, these sound-scenes were created using VBAP rendering. The first was a radio drama production Pinocchio (Churnside, 2012) and the second was a live event recorded using a multi microphone set up in Manchester cathedral of the group Elbow (Churnside, 2011). *See Table 9.1* for the description of each of the sound-scenes.

Stimuli	Description
Drama	30 sec section from radio drama Pinocchio, foreground speech sources in front centre channel with sounds of seagulls overhead and sea in background around listener, voice over of radio presenter in centre channel towards the end of clip.
Music	30 sec section of Elbow live concert with foreground vocals in centre channel and background instruments surrounding the listener, crowd noise and reverberation around the listener

Table 9.1 Additional Sound Scenes Used in Subjective Test

It was desirable to introduce sound scenes which were not part of the elicitation process especially sound scenes which were produced and mixed by engineers to take advantage of 3D surround systems. This would allow further points of comparison with the previous sound scenes used in the elicitation, since these were not produced for any specific playback format and were not complex sound scenes. This would allow an observation of the generalizability of the attribute scales.

9.1.6 Down mixing

A slightly different method was used to derive the sound scenes for the audio stimulus supplied by the BBC R&D department. Initially the sound scenes were mixed for the 9.0 system and then down-mixing was then used to derive the 5.0, stereo and mono soundfields. *Equation 9.1* shows the down mix coefficients for the 5.0 system, where $L_{5.0}$, $R_{5.0}$ stands for left surround right surround etc and $L_{9.0}, R_{9.0}$ stands for left, right for the 9.0 system and $TFL_{9.0}, TFR_{9.0}, TRL_{9.0}, TRR_{9.0}$, donates top front left, top front right, top rear left and top rear right.

Auditioning the different versions of the down-mixes there were no audible artefacts or movement of sources. What was apparent was the reduced perception of space when switching from the 3D surround system to the 2D surround system, which was even more noticeable when switching from the 3D system to the stereo system.

Equation 9.2 shows the down mix coefficients for the stereo system and finally *Equation 9.3* shows the down mix coefficients for the mono system.

$$L_{5.0} = L_{9.0} + \frac{1}{\sqrt{2}} TFL_{9.0}$$

$$R_{5.0} = R_{9.0} + \frac{1}{\sqrt{2}} TFR_{9.0}$$

$$C_{5.0} = C_{9.0}$$

$$L_{s,5.0} = L_{s,9.0} + \frac{1}{\sqrt{2}} TRL_{9.0}$$

$$R_{s,5.0} = R_{s,9.0} + \frac{1}{\sqrt{2}} TRR_{9.0}$$

Equation 9.1

$$L = L_{5.0} + \frac{1}{\sqrt{2}} C_{5.0} + \frac{1}{\sqrt{2}} L_{s,5.0}$$

$$R = R_{5.0} + \frac{1}{\sqrt{2}} C_{5.0} + \frac{1}{\sqrt{2}} R_{s,5.0}$$

Equation 9.2

$$M = L + R$$

Equation 9.3

9.1.7 Level Alignment between Systems

In order to minimise any level differences between the different stimuli a subjective level alignment was carried out by two listeners, described previously in *Chapter 7, Section 7.9.* This involved comparing the multichannel systems with the mono system fed with an input signal of 0dB and adjusting the non mono systems to be of the same loudness. The levels for the ambisonic systems used are shown in *Chapter 7, Table 7.3*, however for the sound scenes supplied by the BBC R&D department these were level aligned separately, to align these systems the same method was used where the non mono systems were aligned with reference to the centre channel at input of 0dB. The levels for each system can be seen below in *Table 9.2.*

System	Level
Stereo	-5.6
5.0	-7
9.0	-7.8

Table 9.2 Input Levels of Each System Relative to the Centre Channel of Input 0dB

9.1.8 Test Procedure

Each listener was presented with a written instruction sheet detailing their task and an explanation of each of the attributes which were to be rated See *Appendix 8,* the attributes scales used can be seen in *Chapter 8, Section 8.6.1-8.6.5* To score each of the systems the listeners were presented with an I-Pad which was connected to the replay computer using a wireless network. Each of the systems for a given sound-scene and attribute were presented simultaneously with sliders and playback buttons, this allowed the listener to switch between systems synchronously and compare systems at will and provide a score, this is said to allow the discrimination of small differences between samples (Bech, 1990), in addition two low and high anchors were included, the low anchor was the mono system whilst the high anchor was the 5.1 system. Sixteen listeners took part, four were from the original panel whilst the remaining twelve are acoustic researchers who had taken part in listening tests before and had an interest in audio research.

9.2 Results

The subjective test utilised two spatial audio rendering methods VBAP and ambisonics. Throughout this thesis ambisonics has been used, therefore the analysis of the ambisonic systems will be carried out first. Direct comparisons cannot be made between the two different rendering methods since each method used a different set of stimuli.

9.2.1 Ambisonic and VBAP Systems Results and Analysis

The mean and 95% confidence intervals have been calculated for each stimuli and each system. The spatial characteristics of each stimulus were different, therefore presenting the system results separated by sample allows an insight into the performance of each system for a given attribute.

9.2.2 Foreground Attribute

The mean and 95% confidence intervals have been calculated for each stimuli and each system and each rendering method, starting with the *foreground attribute*, this method of data

analysis has been used previously by (George et al., 2010), (Conetta, 2011). *See Figure 9.1, 9.2.*

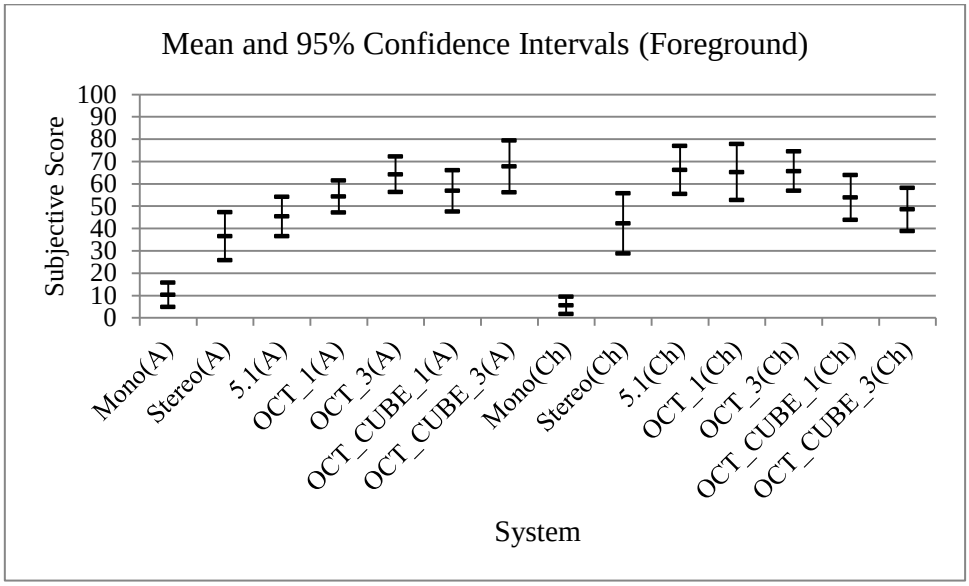


Figure 9.1 Mean and 95% Confidence intervals for foreground attribute (A donates applause sample, Ch donates Choir sample)

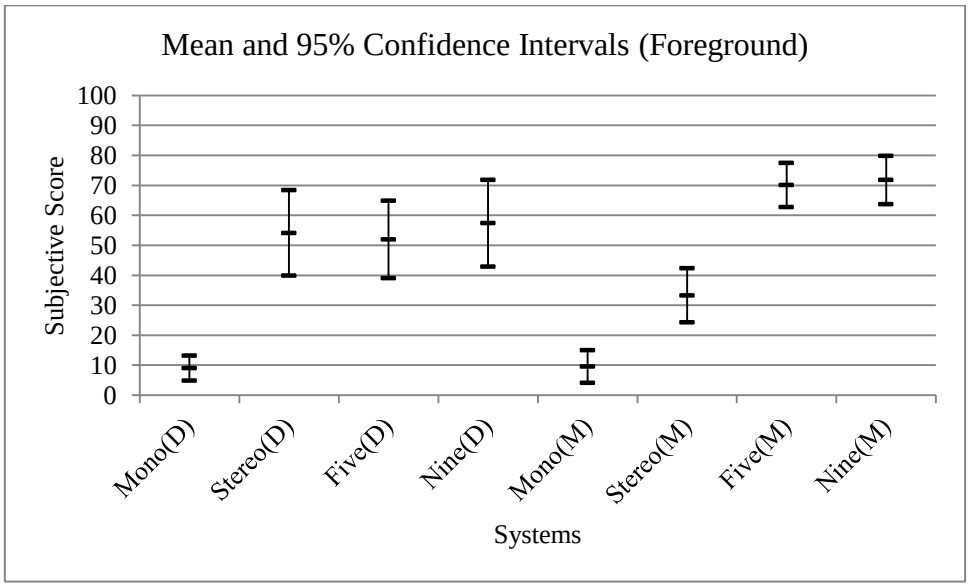


Figure 9.2 Mean and 95% Confidence Intervals for VBAP Systems (D) Donates drama Sample (M) Donates Music Sample

The *foreground* attribute scale was derived by the listeners to scale how evenly distributed the *foreground* sound was, with a low score indicating that the sound scene was biased to a particular direction and a high score indicating that it was uniformly spread around the

listener, the definition of the *foreground* attribute can be seen in *Chapter 8, Section 8.6.1*. It can be observed for all systems and samples that as expected the mono system had the lowest mean score due to the fact that the sound source was only coming from directly in front of the listener. For the other existing systems these have been ranked in order of number of channels especially for the surround systems, with the exception of the drama sample which was predominantly a frontal sound scene using the stereo channels, therefore these system having the same mean score, this would suggest that the listeners were using the scales as expected for this attribute.

9.2.3 Background Attribute

The *background* attribute scale was derived by the listeners to scale how evenly distributed the background sound was, where a low score would indicate that the background sound was biased to a particular direction and a high score would indicate that it was evenly spread around the listener, the formal definition of the *background* attribute can be seen in *Chapter 8, Section 8.6.2*. The mean and 95% confidence intervals for the background attribute are shown below. *See Figure 9.3, 9.4*.

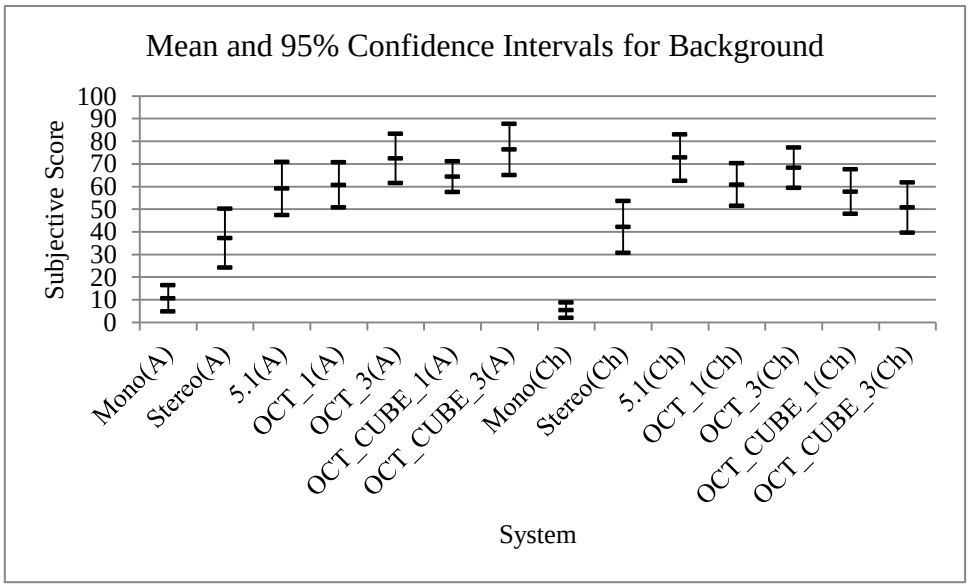


Figure 9.3 Mean and 95% Confidence intervals for background attribute (A donates applause sample, Ch donates Choir sample)

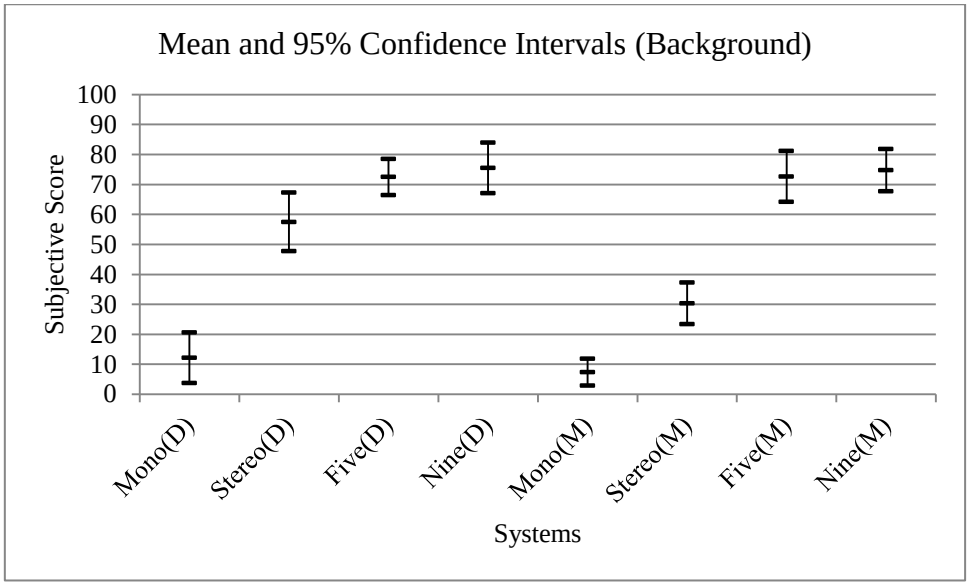


Figure 9.4 Figure 24 Mean and 95% Confidence Intervals for VBAP Systems (D) Donates drama Sample (M) Donates Music Sample

As for the *foreground* attribute a similar trend has been noticed for the background attribute where the mono system has received the lowest mean score and the score has increased as the number of channels have increased. It is interesting to note the overlapping confidence intervals between the stereo system and the surround system for the drama sample which indicates that listeners have not used the scales as expected, since the surround systems had background sound equally distributed around the listener, although the high mean score for the stereo sound scene may be due to a broader soundstage.

9.2.4 Presence Attribute

The *presence* scale was derived by the listeners to scale the degree of involvement in the sound scene, where a low score would indicate that the listener was detached from the sound scene and a high score would indicate that they were immersed in the sound scene. *See Figure 9.5, 9.6.*

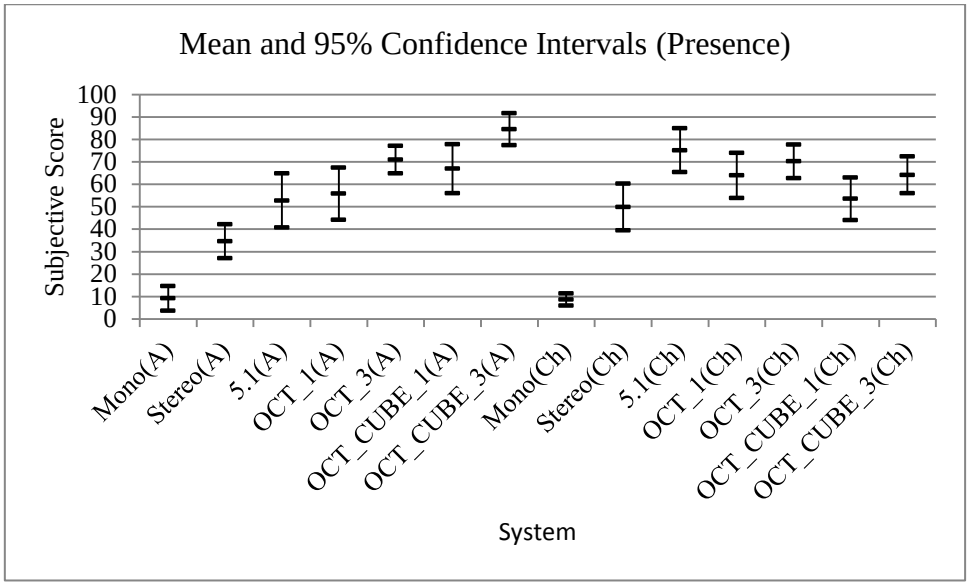


Figure 9.5 Mean and 95% Confidence intervals for presence attribute (A donates applause sample, Ch donates Choir sample)

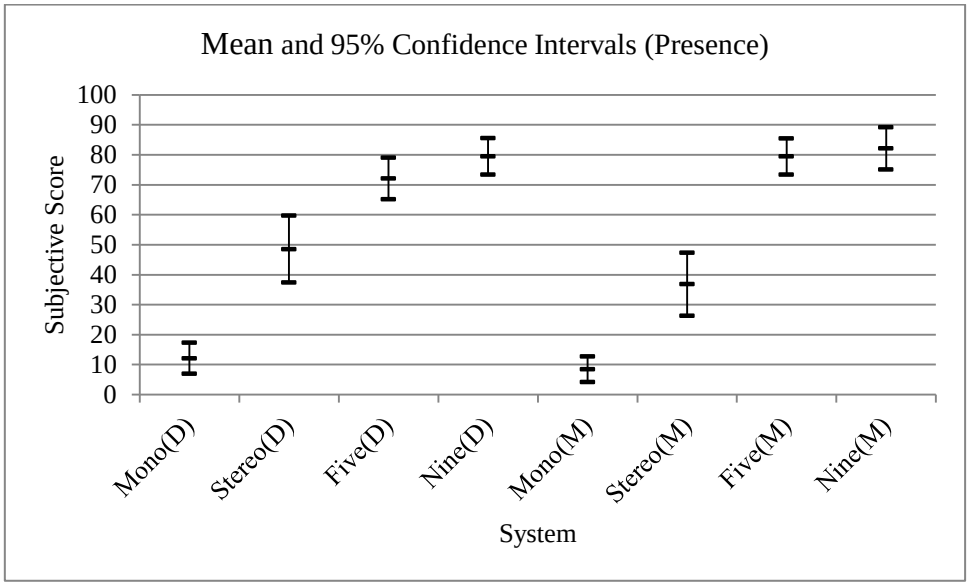


Figure 9.6 Figure 24 Mean and 95% Confidence Intervals for VBAP Systems (D) Donates drama Sample (M) Donates Music Sample

It is interesting to note the similar trend for the mean scores for the VBAP and ambisonic systems, where the score for the perceived presence seems to increase as the number of channels increase. For the ambisonic systems using the applause samples the higher order

systems received the highest mean scores and had the smallest confidence intervals, whilst this is not true for the choir sample as the 5.1 system had the highest mean score.

9.2.5 Size Attribute

The *size* scale was derived by the listeners to scale the perceived size of the reproduced space, a low score would indicate a small space whilst a high score would indicate a large space. A similar trend was observed for the size attribute as previously observed for the presence attribute, the mean and 95% confidence intervals can be seen below in *Figure 9.7, 9.8* for the two audio samples.

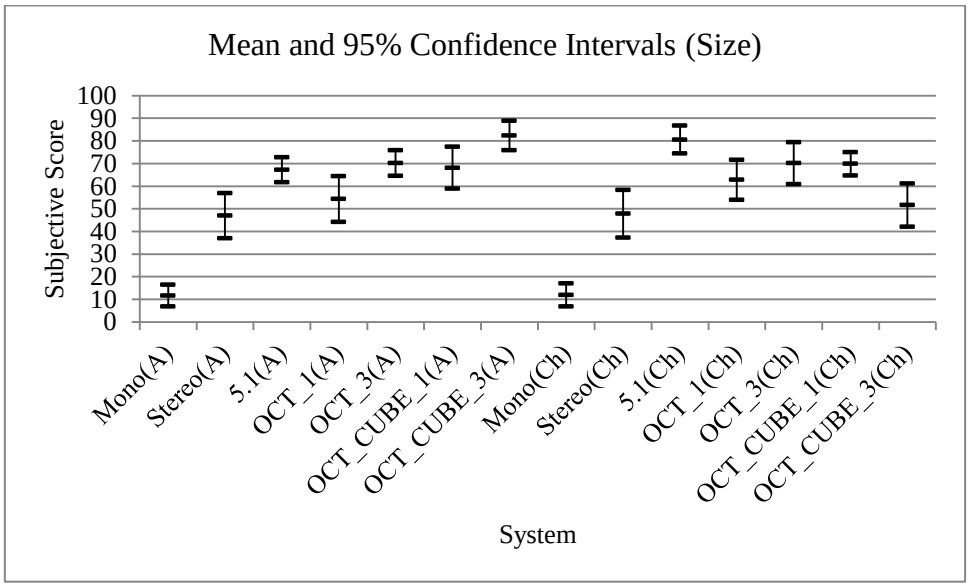


Figure 9.7 Mean and 95% Confidence intervals for size attribute (A donates applause sample, Ch donates Choir sample)

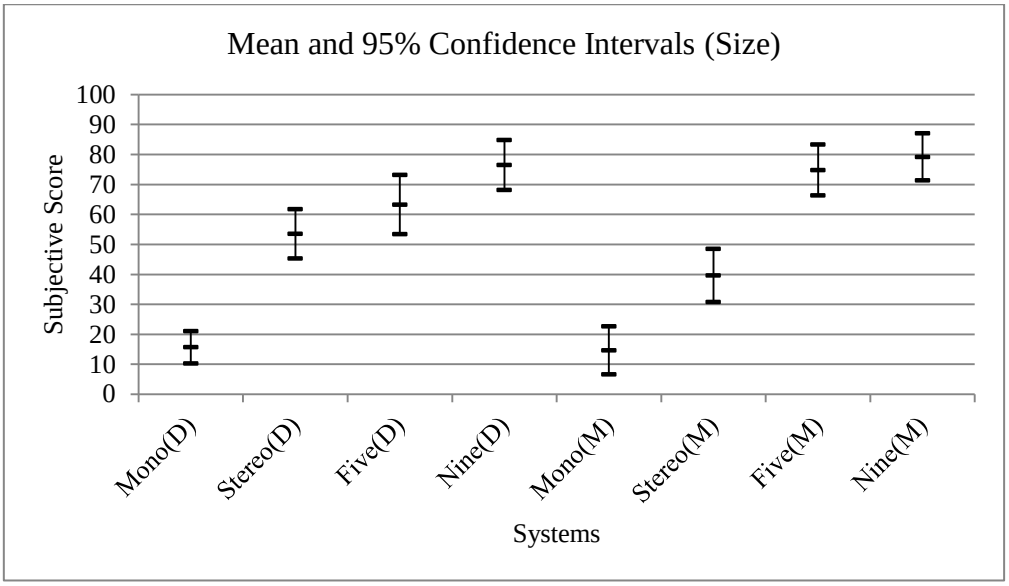


Figure 9.8 Mean and 95% Confidence Intervals for VBAP Systems (D) Donates drama Sample (M) Donates Music Sample

It appears for the ambisonic systems the higher order system with height using the applause sample received the highest mean score. Conversely for the choir sample it can be observed that the 5.1 received the highest mean score. For the VBAP rendered systems using the drama sample the highest mean score was given to the 9 channel 3D system, whilst for the music sample the 2D and 3D systems received a similar mean score.

9.2.6 Realism

The *realism* scale was derived by the listeners to rate how realistic the sound scene was, with a low score indicating that the sound scene was unrealistic and a high score indicating the sound scene was realistic. *See Figure 9.9, 9.10,* for mean and 95% confidence intervals for ambisonic and VBAP systems.

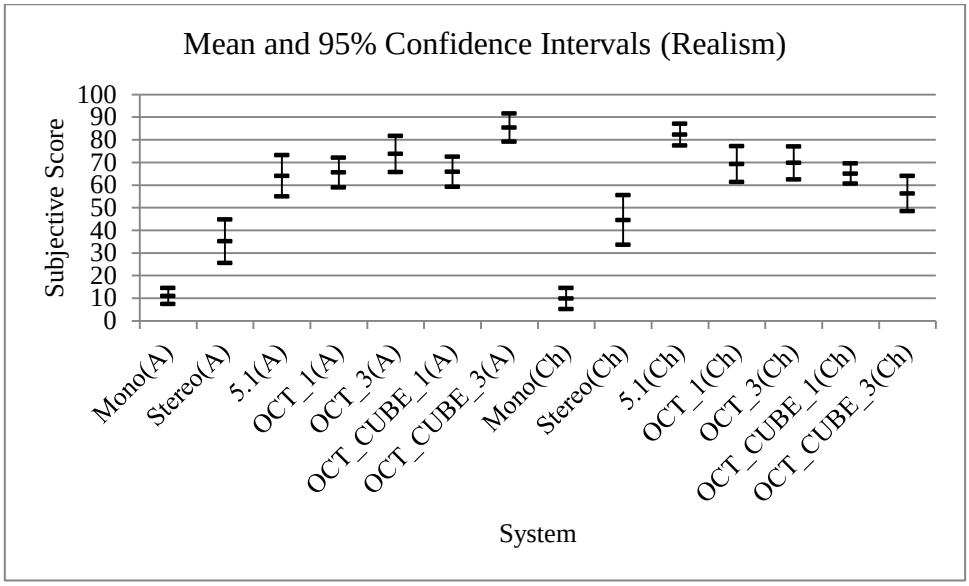


Figure 9.9 Mean and 95% Confidence intervals for realism attribute (A donates applause sample, Ch donates Choir sample)

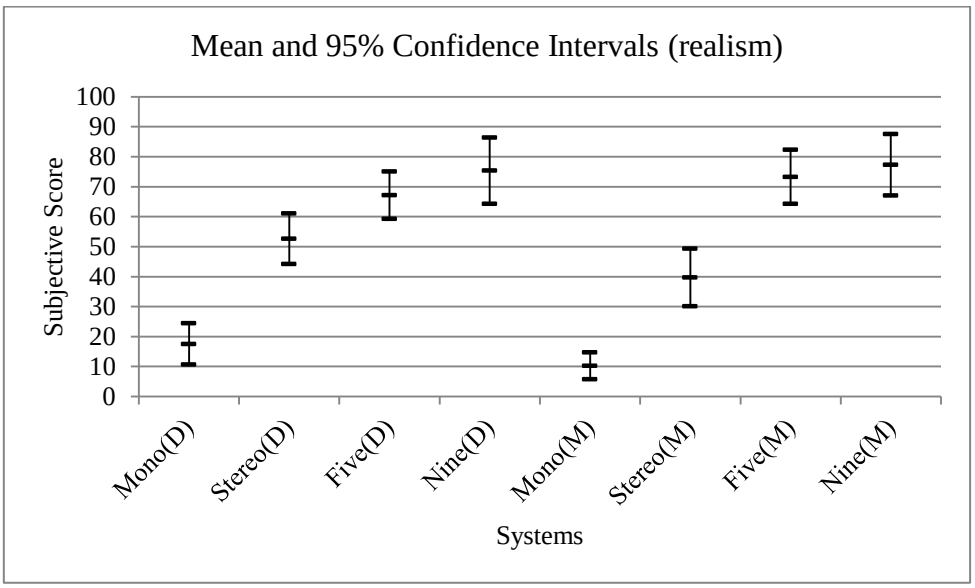


Figure 9.10 Mean and 95% Confidence Intervals for VBAP Systems (D) Donates drama Sample (M) Donates Music Sample

A similar trend is noticed for the *realism* attribute compared to the previous *size* attribute where the highest mean score for the ambisonic systems has been for the higher order 3D system using the applause sample and that the highest mean score for the choir sample was for the 5.1 system. For the VBAP systems a slightly higher mean score was observed for the

3D system using the drama sample, however for the music sample similar mean score were given to the 2D and 3D systems.

9.3 Significance of Results

A two way repeated measures analysis of variance (ANOVA) was carried out with two factors, the first was stimulus with two levels and the second was system with seven levels.

Before running the ANOVA a Mauchlys test of sphericty was carried out for each of the attributes factors, System, Stimulus and the interaction, where sphericity has been violated (the data do not support the assumption that variances are equal across conditions) a Greenhouse-Geisser correction was used (Field, 2009). Where a Greenhouse Geisser correction has been used the cells in the *Table 9.3* below has been marked with an asterisk.

		Foreground	**Background**	**Presence**	**Realism**	**Size**
Ambisonic	System	ρ= <0.05, partial η^2= .71*	p=<0.05, partial η^2= .77*	p= <0.05, partial η^2= .80	p=<0.05, partial η^2= .87	p= < 0.05, partial η^2= .81
	Stimulus	p=>0.05, partial η^2=.011	p=>0.05, partial η^2 = .04	p=>0.05, partial η^2= .03	p=>0.05, partial η^2= .005	p=> 0.05, partial η^2= .009
	Sys*Stim	p=<0.05, partial η^2 =.23	p=<0.05, partial η^2= .26	p= <0.05, partial η^2= .35*	p= <0.05, partial η^2= .40*	p= < 0.05, partial η^2= .37*
VBAP	System	p= <0.05, partial η^2= .82	p=<0.05, partial η^2= .86	p=<0.05, partial η^2= .92	p= <0.05, partial η^2= .84	p=<0.05, partial η^2= .94
	Stimulus	p = >0.05, partial η^2= .042	p=<0.05, partial η^2= .44	p=>0.05, partial η^2= .02	p= >0.05, partial η^2= .06	p=>0.05, partial η^2= .000
	Sys*Stim	p =< 0.05, partial η^2= .44	p=<0.05, partial η^2= .36	p=<0.05, partial η^2= .19	p= >0.05, partial η^2= .14	p=<0.05, partial η^2= .51

Table 9.3 Significance of Two Way ANOVA for the Five Attributes used, (*) Sphericity Violated Donates Greenhouse Geisser Correction

In order to investigate the differences between systems and for the interaction, for each attribute and stimulus Bonferroni corrections were used in order to make pairwise comparisons starting with the foreground attribute.

9.3.1 Foreground Attribute

The tables below show the significance between each of the pairwise comparisons, the cells highlighted indicate significant differences at p=<0.05 starting with the factor System See Table *9.4, 9.5*, then a breakdown of the interaction System*Stimuli. *See Tables 9.6 and 9.7.*

	Mono	**Stereo**	**5.1**	**Oct_1**	**Oct_3**	**Oct_Cube_1**	**Oct_Cube_3**
Mono		0	0	0	0	0	0
Stereo	0		0.21	0.04	0.036	0.181	0.382
5.1	0	0.21		1	0.947	1	1
Oct_1	0	0.04	1		1	1	1
Oct_3	0	0.036	0.947	1		1	0.992
Oct_Cube_1	0	0.181	1	1	1		1
Oct_Cube_3	0	0.382	1	1	0.992	1	

Table 9.4 Pairwise Comparisons for Factor System Using Bonferroni Correction Ambisonic

	Mono	**Stereo**	**5**	**9**
Mono		0	0	0
Stereo	0		0.017	0.006
5	0	0.017		1
9	0	0.006	1	

Table 9.5 Pairwise Comparisons for Factor System Using Bonferroni Correction VBAP

		Mono	Stereo	5.1	Oct_1	Oct_3	Oct_Cube_1	Oct_Cube_3
Applause	**Mono**		0	0	0	0	0	0
	Stereo	0		0.21	0.04	0.036	0.181	0.382
	5.1	0	0.21		1	0.947	1	1
	Oct_1	0	0.163	1		1	1	1
	Oct_3	0	0.135	0.108	1		1	0.992
	Oct_Cube_1	0	0.153	1	1	1		1
	Oct_Cube_3	0	0.171	0.116	1	1	1	
Choir	**Mono**		0.001	0	0	0	0	0
	Stereo	0.001		0.132	0.117	0.04	1	1
	5.1	0	0.132		1	1	1	0.84
	Oct_1	0	0.117	1		1	1	0.199
	Oct_3	0	0.04	1	1		1	0.03
	Oct_Cube_1	0	1	1		1		1
	Oct_Cube_3	0	1	0.084	0.199	0.03	1	

Table 9.6 Pairwise Comparisons for Interaction System*Stimuli Using Bonferroni Corrections for Foreground Attribute Ambisonics

		Mono	Stereo	5	9
Drama	**Mono**		0	0	0
	Stereo	0		1	1
	5	0	1		1
	9	0	1	1	
Elbow	**Mono**		0	0	0
	Stereo	0		0	0
	5	0	0		1
	9	0	0	1	

Table 9.7 Pairwise Comparisons for Interaction System*Stimuli Using Bonferroni Correction for Foreground Attribute VBAP

9.3.2 Background Attribute

The tables below show the significance between each of the pairwise comparisons, the cells highlighted indicate significant differences at the p=<0.05, the significance tables start with pairwise comparisons for the factor System *See Tables 9.8, 9.9,* then a breakdown of the interaction System*Stimuli. *See Tables 9.10 and 9.11*.

	Mono	Stereo	5.1	Oct_1	Oct_3	Oct_Cube_1	Oct_Cube_3
Mono		0	0	0	0	0	0
Stereo	0		0.35	0.003	0.003	0.52	0.09
5.1	0	0		1	1	1	1
Oct_1	0	0.003	1		0.086	1	1
Oct_3	0	0.003	1	0.086		0.041	1
Oct_Cube_1	0	0.052	1	1	0.041		1
Oct_Cube_3	0	0.09	1	1	1	1	

Table 9.8 Pairwise Comparisons for Factor System Using Bonferroni Correction Ambisonic

	Mono	Stereo	5	9
Mono		0	0	0
Stereo	0		0	0
5	0	0		1
9	0	0	1	

Table 9.9 Pairwise Comparisons for Factor System Using Bonferroni Correction VBAP

		Mono	Stereo	5.1	Oct_1	Oct_3	Oct_Cube_1	Oct_Cube_3
Applause	**Mono**		0.011	0	0	0	0	0
	Stereo	0.011		0.34	0.043	0.027	0.04	0.005
	5.1	0	0.34		1	0.571	1	0.88
	Oct_1	0	0.043	1		0.445	1	1
	Oct_3	0	0.027	0.571	0.445		1	1
	Oct_Cube_1	0	0.04	1	1	1		0.341
	Oct_Cube_3	0	0.005	0.88	1	1	0.341	
Choir	**Mono**		0	0	0	0	0	0
	Stereo	0		0.025	0.025	0.005	0.576	1
	5.1	0	0.025		0.569	1	0.597	0.035
	Oct_1	0	0.025	0.569		0.861	1	1
	Oct_3	0	0.005	1	0.861		1	1
	Oct_Cube_1	0	0.576	0.597	1	0.841		1
	Oct_Cube_3	0	1	0.035	1	0.124	1	

Table 9.10 Pairwise Comparisons for Interaction System*Stimuli Using Bonferroni Correction for Background Attribute Ambisonics

		Mono	Stereo	5	9
Drama	**Mono**		0	0	0
	Stereo	0		0.209	0.159
	5	0	0.209		1
	9	0	0.159	1	
Elbow	**Mono**		0	0	0
	Stereo	0		0	0
	5	0	0		1
	9	0	0	1	

Table 9.11Pairwise Comparisons for Interaction System*Stimuli Using Bonferroni Correction for Background Attribute VBAP

9.3.3 Presence

The tables below show the significance between each of the pairwise comparisons, the cells highlighted indicate significant differences at the p=<0.05, the significance tables start with pairwise comparisons for the factor System *See Tables 9.12, 9.13,* then a breakdown of the interaction System*Stimuli. *See Tables 9.14 and 9.15*.

	Mono	Stereo	5.1	Oct_1	Oct_3	Oct_Cube_1	Oct_Cube_3
Mono		0	0	0	0	0	0
Stereo	0		0.38	0.095	0	0.073	0
5.1	0	0.38		1	1	1	1
Oct_1	0	0.095	1		0.538	1	0.09
Oct_3	0	0	1	0.086		0.259	1
Oct_Cube_1	0	0.073	1	1	0.259		0.042
Oct_Cube_3	0	0	1	0.09	1	0.042	

Table 9.12 Pairwise Comparisons for Factor System Using Bonferroni Correction Ambisonic

	Mono	Stereo	5	9
Mono		0	0	0
Stereo	0		0	0
5	0	0		0.458
9	0	0	0.458	

Table 9.13 Pairwise Comparisons for Factor System Using Bonferroni Correction VBAP

		Mono	Stereo	5.1	Oct_1	Oct_3	Oct_Cube_1	Oct_Cube_3
Applause	**Mono**		0	0	0	0	0	0
	Stereo	0		0.259	0.116	0	0.002	0
	5.1	0	0.259		1	0.273	0.986	0.1
	Oct_1	0	0.116	1		0.614	1	0.001
	Oct_3	0	0	0.273	0.614		1	0.213
	Oct_Cube_1	0	0.002	0.986	1	1		0.161
	Oct_Cube_3	0	0	0.01	0.001	0.213	0.161	
Choir	**Mono**		0	0	0	0	0	0
	Stereo	0		0.199	1	0.235	1	1
	5.1	0	0.199		0.564	1	0.001	1
	Oct_1	0	1	0.564		1	1	1
	Oct_3	0	0.235	1	1		0.006	1
	Oct_Cube_1	0	1	0.001	1	0.006		1
	Oct_Cube_3	0	1	1	1	1	1	

Table 9.14 Pairwise Comparisons for Interaction System*Stimuli Using Bonferroni Correction for Presence Attribute Ambisonics

		Mono	Stereo	5	9
Drama	**Mono**		0	0	0
	Stereo	0		0.001	0
	5	0	0.001		0.52
	9	0	0	0.52	
Elbow	**Mono**		0	0	0
	Stereo	0		0	0
	5	0	0		0.436
	9	0	0	0.436	

Table 9.15 Pairwise Comparisons for Interaction System*Stimuli Using Bonferroni Correction for Presence Attribute VBAP

9.3.4 Size

The tables below show the significance between each of the pairwise comparisons, the cells highlighted indicate significant differences at p=<0.05, the significance tables start with pairwise comparisons for the factor System *See Tables 9.16, 9.17,* then a breakdown of the interaction System*Stimuli. *See Tables 9.18 and 9.19.*

	Mono	Stereo	5.1	Oct_1	Oct_3	Oct_Cube_1	Oct_Cube_3
Mono		0	0	0	0	0	0
Stereo	0		0	0.489	0.008	0.003	0.01
5.1	0	0		0.041	1	1	0.889
Oct_1	0	0.489	0.041		0.102	1	0.112
Oct_3	0	0.008	1	0.102		1	1
Oct_Cube_1	0	0.003	1	1	1		1
Oct_Cube_3	0	0.01	0.889	0.112	1	1	

Table 9.16 Pairwise Comparisons for Factor System Using Bonferroni Correction Ambisonic

	Mono	Stereo	5	9
Mono		0	0	0
Stereo	0		0.003	0
5	0	0.003		0.004
9	0	0	0.004	

Table 9.17 Pairwise Comparisons for Factor System Using Bonferroni Correction VBAP

		Mono	Stereo	5.1	Oct_1	Oct_3	Oct_Cube_1	Oct_Cube_3
Applause	**Mono**		0	0	0	0	0	0
	Stereo	0		0.015	1	0.063	0.029	0
	5.1	0	0.015		0.693	1	1	0.019
	Oct_1	0	1	0.693		0.178	1	0.001
	Oct_3	0	0.063	1	0.178		1	0.009
	Oct_Cube_1	0	0.29	1	1	1		0.412
	Oct_Cube_3	0	0	0.019	0.001	0.009	0.412	
Choir	**Mono**		0	0	0	0	0	0
	Stereo	0		0	0.636	0.116	0.008	1
	5.1	0	0		0.184	1	0.149	0.001
	Oct_1	0	0.636	0.184		1	1	1
	Oct_3	0	0.116	1	1		1	0.293
	Oct_Cube_1	0	0.008	0.149	1	1		0.04
	Oct_Cube_3	0	1	0.001	1	0.293	0.04	

Table 9.18 Pairwise Comparisons for Interaction System*Stimuli Using Bonferroni Correction for Size Attribute Ambisonics

		Mono	Stereo	5	9
Drama	**Mono**		0	0	0
	Stereo	0		0.161	0.002
	5	0	0.161		0.003
	9	0	0.002	0.003	
Elbow	**Mono**		0	0	0
	Stereo	0		0	0
	5	0	0		0.116
	9	0	0	0.116	

Table 9.19 Pairwise Comparisons for Interaction System*Stimuli Using Bonferroni Correction for Size Attribute VBAP

9.3.5 Realism

The tables below show the significance between each of the pairwise comparisons, the cells highlighted indicate significant differences the significance tables start with pairwise comparisons for the factor System *See Tables 9.20, 9.21* then a breakdown of the interaction System*Stimuli *See Tables 9.22 and 9.23*.

	Mono	Stereo	5.1	Oct_1	Oct_3	Oct_Cube_1	Oct_Cube_3
Mono		0	0	0	0	0	0
Stereo	0		0	0	0	0	0.01
5.1	0	0		1	1	0.736	1
Oct_1	0	0	1		1	1	1
Oct_3	0	0	1	1		0.682	1
Oct_Cube_1	0	0	0.736	1	0.682		1
Oct_Cube_3	0	0.001	1	1	1	1	

Table 9.20 Pairwise Comparisons for Factor System Using Bonferroni Correction Ambisonic

	Mono	Stereo	5	9
Mono		0	0	0
Stereo	0		0	0
5	0	0		0.268
9	0	0	0.268	

Table 9.21 Pairwise Comparisons for Factor System Using Bonferroni Correction VBAP

		Mono	Stereo	5.1	Oct_1	Oct_3	Oct_Cube_1	Oct_Cube_3
Applause	**Mono**		0.001	0	0	0	0	0
	Stereo	0.001		0.005	0.011	0	0.001	0
	5.1	0	0.005		1	0.583	1	0.29
	Oct_1	0	0.11	1		1	1	0.017
	Oct_3	0	0	0.583	1		1	0.345
	Oct_Cube_1	0	0.001	1	1	1		0.014
	Oct_Cube_3	0	0	0.029	0.17	0.345	0.014	
Choir	**Mono**		0	0	0	0	0	0
	Stereo	0		0	0.025	0.007	0.055	1
	5.1	0	0		0.04	0.156	0	0.001
	Oct_1	0	0.007	0.156		1	1	0.762
	Oct_3	0	0.007	0.156	1		1	0.31
	Oct_Cube_1	0	0.055	0	1	1		1
	Oct_Cube_3	0	1	0.001	0.762	0.31	1	

Table 9.22 Pairwise Comparisons for Interaction System*Stimuli Using Bonferroni Correction Realism Attribute Ambisonics

		Mono	Stereo	5	9
Drama	**Mono**		0	0	0
	Stereo	0		0.009	0.004
	5	0	0.009		0.061
	9	0	0.004	0.061	
Elbow	**Mono**		0	0	0
	Stereo	0		0.001	0
	5	0	0.001		0.238
	9	0	0	0.238	

Table 9.23 Pairwise Comparisons for Interaction System*Stimuli Using Bonferroni Correction Realism Attribute VBAP

9.3.6 Significance of Results Summary

For the *foreground* and *background* attributes the results seem to suggest that there is no significant difference to the perception of these attributes with the addition of height. This result has been found for the ambisonic rendered sound scenes and also for the sound scenes that were rendered using VBAP. Looking at the results for the presence attribute it appears

that *presence* is significantly enhanced over the current 5.1 configuration with the use of height channels using higher order ambisonics, this was found for the applause sample. Whilst for the choir sample it appears that this is not the case. For the VBAP rendered sound scenes it appears that the 2D surround system is not significantly different to that of the height systems for the *presence* attribute, however for the drama sample the level of significance between the 2D and 3D system is borderline.

It also appears that for the attribute *realism* using the applause sample the addition of height using 3rd order ambisonics adds to the *realism* over the current 5.1 configuration. Whilst for the choir sample the addition of height in either 1st or 3rd ambisonics with height enhances the *realism* over the current 5.1 configuration. For the sound scenes rendered using VBAP there is no significant enhancement of *realism* with the addition of height using either the elbow recording or the drama sound scene.

When looking at the significance values for the *size* attribute, it can be seen that for the applause sample the addition of height using 3rd order ambisonics provides a better sense of the size of the reproduced space. This is not the case when using the choir sample and the results show that increasing ambisonics to 3rd order there is no improvement over the stereo system, contrary to this it appears that the 5.1 system does provide a significant enhancement over the stereo system. Therefore as before for the choir sample which is predominantly a frontal sound scene with subtle room tone it appears that the 5.1 system is not significantly different to the other 2D surround systems, which suggests that increasing the number of channels in the horizontal in terms of the choir sound scene does not improve the perception of the *size* of the reproduced space. For the VBAP rendered sound scenes a similar result has been found where depending on the sound scene used the addition of height makes a significant difference to the perceived *size* of the reproduced sound scene, yet for the Elbow music sample again the 5 channel surround system is significantly different to that of the stereo system but that the inclusion of height does not provide significant enhancement.

9.4 Subjective Evaluation Summary

Overall it has been observed that depending on the stimulus used has had an effect on the scores provided by the listeners. The use of the applause sample has shown that the inclusion of height has enhanced the perceived size of the reproduced space and for the attribute *presence* it has shown that the increase in ambisonic order gives the same perceived presence as utilising the addition of height. It is also discussed by Shirley et al. (2013) that the

inclusion of the high resolution Eigenmike recordings means that the feeling of presence is increased. This would be expected from the theory of ambisonic decoding, since higher order decoding provides greater resolution in the reproduced sound scene (Bertet et al., 2013). For the two other attributes it appears that *foreground* and *background* are not significantly enhanced with the use of height channels.

A different set of results were obtained for the choir sample, since overall it seems that for the choir sample there is no clear advantage in utilising height, since it was found that for most of the attributes the 5.1 system had a significantly higher mean score. This suggests that for reproductions which only utilise the frontal channels for the direct sound sources and subtle room tone in the surround or height channels, the 5.1 system would provide a similar experience.

It could be said that these results are only valid for the ambisonics rendered system, yet utilising the five attribute scales with a different set of stimuli and reproduction systems it has been found that there is one attribute which separates the 3D system from the 2D surround systems and that is the attribute *Size*. This seems to suggest that the extension of height in surround systems provides the listener additional information which helps in perceiving the *size* of the reproduced space in comparison to 2D surround systems. However, as found for the ambisonic systems this seems to be affected by the input signal used.

Since the *size* attribute was one of five scales, it is necessary to investigate as to whether the scales are in fact independent. This would establish whether the scales are correlated which would suggest that they are measuring a similar underlying percept, which may suggest that there is further redundancy between scales therefore allowing a further reduction of attribute scales.

9.5 Independence between Attributes

The use of the descriptive analysis technique aims to achieve independence between attribute scales i.e. that the scales are in fact measuring independent sensations, furthermore it allows an insight into the underlying perceptual space of a number of reproduction systems. Correlation matrices have been produced for the 2D and 3D systems used, since this research is primarily concerned with the surround systems attention has been focused on these systems starting with the data combined for the ambisonic and VBAP 2D and 3D systems. In order to guide the judgement of the strength of the correlations the definitions given by Dancey and Reidy (2004) have been used. *See Table 9.24.*

Value of The Correlation Coefficient	Strength of Correlation
1	Perfect
0.7-0.9	Strong
0.4-0.6	Moderate
0.1-0.3	Weak
0	Zero

Table 9.24 Strength of the Relationship for Correlation Coefficients (Dancey and Reidy, 2004)

It is desirable to have attribute scales with a wide application which can be used with 2D and 3D surround systems rendered with various methods. In this case the correlation between the five scales used for the ambisonic and VBAP systems combined can be seen below in Figure 9.25.

Correlation Matrix

		Foreground	Background	Presence	Size	Realism
Correlation	Foreground	1.000	.011	.279	.269	.200
	Background	.011	1.000	.229	.011	.086
	Presence	.279	.229	1.000	.296	.342
	Size	.269	.011	.296	1.000	.356
	Realism	.200	.086	.342	.356	1.000

Table 9.25 Correlation Coefficients between Attributes for the Ambisonic 2D and 3D Systems and VBAP 2D and 3D Systems Combined

It can be observed that the correlations between attribute scales for all of the ambisonic and VBAP 2D and 3D systems combined shows weak correlations.

To explore the correlations further, the surround systems 2D and 3D are then separated based on the rendering method. *Table 9.26* below shows the correlations for the ambisonic surround systems 2D and 3D combined.

Correlation Matrix

		Foreground	Background	Presence	Size	Realism
Correlation	Foreground	1.000	.280	.288	.262	.350
	Background	.280	1.000	.484	.270	.512
	Presence	.288	.484	1.000	.229	.425
	Size	.262	.270	.229	1.000	.363
	Realism	.350	.512	.425	.363	1.000

Table 9.26 Correlation Coefficients between Attributes for the Ambisonic 2D and 3D Systems Combined (Moderate Correlations Marked in Red)

It can be seen that there are moderate correlations between attributes for the ambisonic systems. *Table 9.27* below shows the correlations between the attributes for the VBAP 2D and 3D systems combined.

Correlation Matrix

		Foreground	Background	Presence	Size	Realism
Correlation	Foreground	1.000	.097	.231	.258	-.091
	Background	.097	1.000	.020	.201	.647
	Presence	.231	.020	1.000	.453	.075
	Size	.258	.201	.453	1.000	.321
	Realism	-.091	.647	.075	.321	1.000

Table 9.27 Correlation Coefficients between Attributes for the VBAP 2D and 3D Systems Combined (Moderate Correlations Marked in Red)

The initial correlations for 2D and 3D systems with the ambisonic and VBAP systems combined shows weak correlations between attribute scales. However the correlation tables analyzed separately for the ambisonic surround systems and the VBAP surround systems indicate moderate correlations (highlighted) between attribute scales. For the ambisonic surround systems this is between the attributes *background*, *presence* and *realism*. On the other hand for the VBAP surround systems, this is between the attributes *size* and *presence* and similar to the ambisonic systems, between *background* and *realism* attributes.

It would seem that in the ambisonic systems case the correlations between attributes *background*, *presence* and *realism* would suggest that these scales could be measuring the

same underlying percept. Although caution is needed since this is not the case for the VBAP surround systems, since the only similar correlation between attributes is with *background* and *realism*.

It is also of interest to investigate the attribute scales use regarding 2D and 3D surround systems separately for each rendering method, since it has been observed that the correlations for the combination of 2D and 3D systems has shown slightly different correlations between attribute scales, suggesting that there are slight differences in scale usage for the ambisonic and VBAP systems. The correlation matrix for the ambisonic 2D and 3D systems can be seen below in *Tables 9.28* and *9.29* respectively.

Correlation Matrix

		Foreground	Background	Presence	Size	Realism
Correlation	Foreground	1.000	.233	.301	.244	.295
	Background	.233	1.000	.757	.688	.557
	Presence	.301	.757	1.000	.512	.442
	Size	.244	.688	.512	1.000	.499
	Realism	.295	.557	.442	.499	1.000

Table 9.28 Correlation Coefficients between Attributes for the Ambisonic 3D Systems (moderate-strong Correlations Marked in Red)

Correlation Matrix

		Foreground	Background	Presence	Size	Realism
Correlation	Foreground	1.000	.287	.183	.278	.375
	Background	.287	1.000	.822	.650	.474
	Presence	.183	.822	1.000	.636	.372
	Size	.278	.650	.636	1.000	.333
	Realism	.375	.474	.372	.333	1.000

Table 9.29 Correlation Coefficients between Attributes for the Ambisonic 2D Surround Systems (Moderate- Strong Correlations Marked in Red)

It can be seen that there are moderate/strong correlation values between several attributes including *background*, *presence*, *size* and *realism* for both 2D and 3D systems. The

correlations between the attributes were also carried out for the surround systems using the VBAP rendering. *See Tables 9.30 and 9.31.*

Correlation Matrix

		Foreground	Background	Presence	Size	Realism
Correlation	Foreground	1.000	.188	.101	.105	-.332
	Background	.188	1.000	-.107	.166	.320
	Presence	.101	-.107	1.000	.417	.197
	Size	.105	.166	.417	1.000	.448
	Realism	-.332	.320	.197	.448	1.000

Table 9.30 Correlation Coefficients between Attributes for the VBAP 3D Surround System (Moderate Correlations Marked in Red)

		Foreground	Background	Presence	Size	Realism
Correlation	Foreground	1.000	.317	.170	.141	.196
	Background	.317	1.000	.120	.231	.558
	Presence	.170	.120	1.000	.122	.139
	Size	.141	.231	.122	1.000	.253
	Realism	.196	.558	.139	.253	1.000

Table 9.31 Correlation Coefficients between Attributes for the VBAP 2D Surround System (Moderate Correlations Marked in Red)

It can be observed that there are only two moderate correlations for the 3D VBAP system, this is between the attributes *presence* and *size* and also *realism* and *size.* For the 2D system there is only one moderate correlation which is between *background* and *realism*.

9.5.1 Summary

The initial investigation of correlations between attribute scales with the ambisonic and VBAP 2D and 3D systems combined showed that there were weak correlations between attribute scales. Carrying out further investigation of correlations between attribute scales by separating the analysis firstly by rendering method and then by reproduction system two situations have arisen. The correlations between attribute scales for the 2D and 3D ambisonic surround systems point to the fact that there is redundancy between scales and suggests that the scales are measuring the same underlying percepts. However in light of the results for the

VBAP 2D and 3D systems caution is needed because in general the correlations between attributes for the VBAP 2D and 3D systems are weak. In addition to this it has been observed across the ambisonic and VBAP 2D and 3D systems the *foreground* and *background* attributes have

remained weakly correlated. Reasons for the differences between the correlations for the ambisonic systems and the VBAP systems will be explored in the Discussion section.

9.6 Principal Components Analysis- Focus on Surround Systems

From the previously investigated correlation coefficients it can be observed that for ambisonics and VBAP systems combined there was observed weak correlations between attribute scales. Separating the systems by rendering method for the ambisonics surround systems there were moderate correlations between the attributes *background*, *presence* and *realism*. For the VBAP surround systems it was found that there were moderate correlations between size and presence and similar to the ambisonic surround 2D and 3D systems between *background* and *realism*. To explore the relationship between the attribute scales further a Principal Component Analysis has been carried out. As before the 2D and 3D surround systems data has been combined for the ambisonic and VBAP systems for the initial analysis.

A principal component analysis (PCA) is used to find the underlying dimensions of a data set and also in the case of large data sets to reduce the data by finding redundancy. Firstly the PCA uses eigenvalues to determine the importance of a factor, it is reported by Jolliffe (2005) that a factor should be retained if it has an eigenvalue of >.7, this criterion was used. In addition to this Cattell's scree test is also used (Cattell, 1966) where the point of inflexion in the plot determines the number of factors or components to include. Usually the component at the point of inflexion is also discarded and everything to the left of that point on the scree plot is kept (Field, 2009). However it is not always appropriate to discard the component at the point of inflexion since it will sometimes be >.7, it is also true that there may be occasions when the component after the point of inflexion can also be ≥ to .7 and therefore according to (Jolliffe, 2005) this is an important component.

9.6.1 Ambisonic and VBAP Surround Systems

For the ambisonic and VBAP 2D and 3D surround systems an initial PCA analysis was run, in order to determine the number of components to keep the scree plot was used.

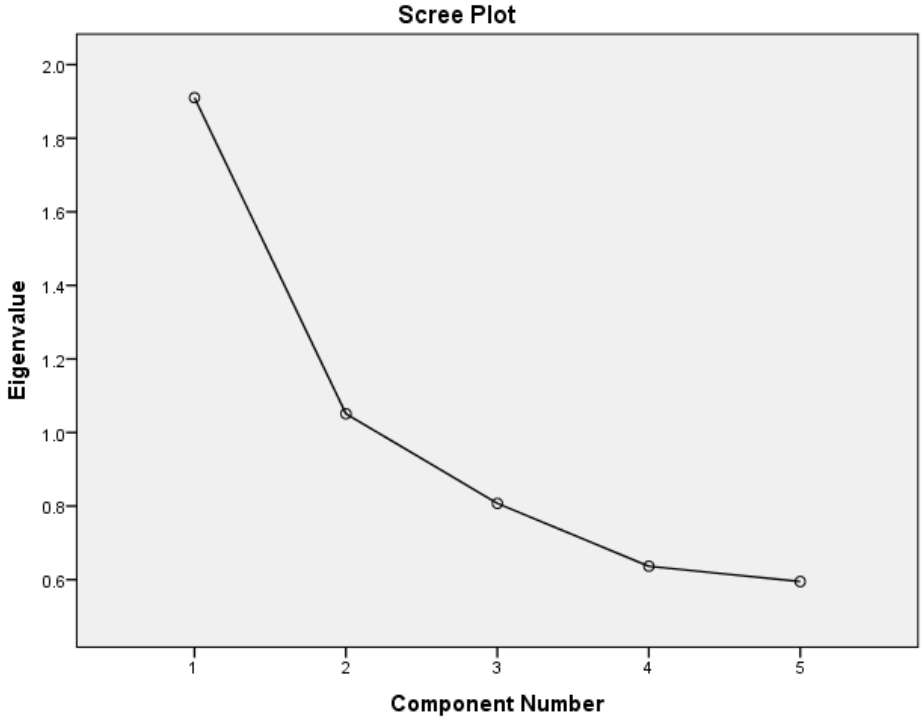


Figure 9.11 Scree Plot of Component Number versus Eigenvalue for Ambisonic and VBAP 2D & 3D Surround Systems Combined

It can be observed from the scree plot that there is a point of inflexion at the second component, however there is another inflexion point at the third component which is >.7 therefore the PCA was rerun to extract three factors. The three factors extracted explained a total of 75% of the variance, the first factor explaining 38% of the variance, the second 21% and the third 16%.

To get a clearer loading of the factors the three factors were subjected to a varimax rotation, this helps in the interpretation of the principal components by maximising the loadings onto each component (Field, 2009). This showed that *size* and *realism* loaded onto the first component, *background* and *presence* onto the second and finally *foreground* onto the third component. The component plots can be seen in *Figure 9.12*.

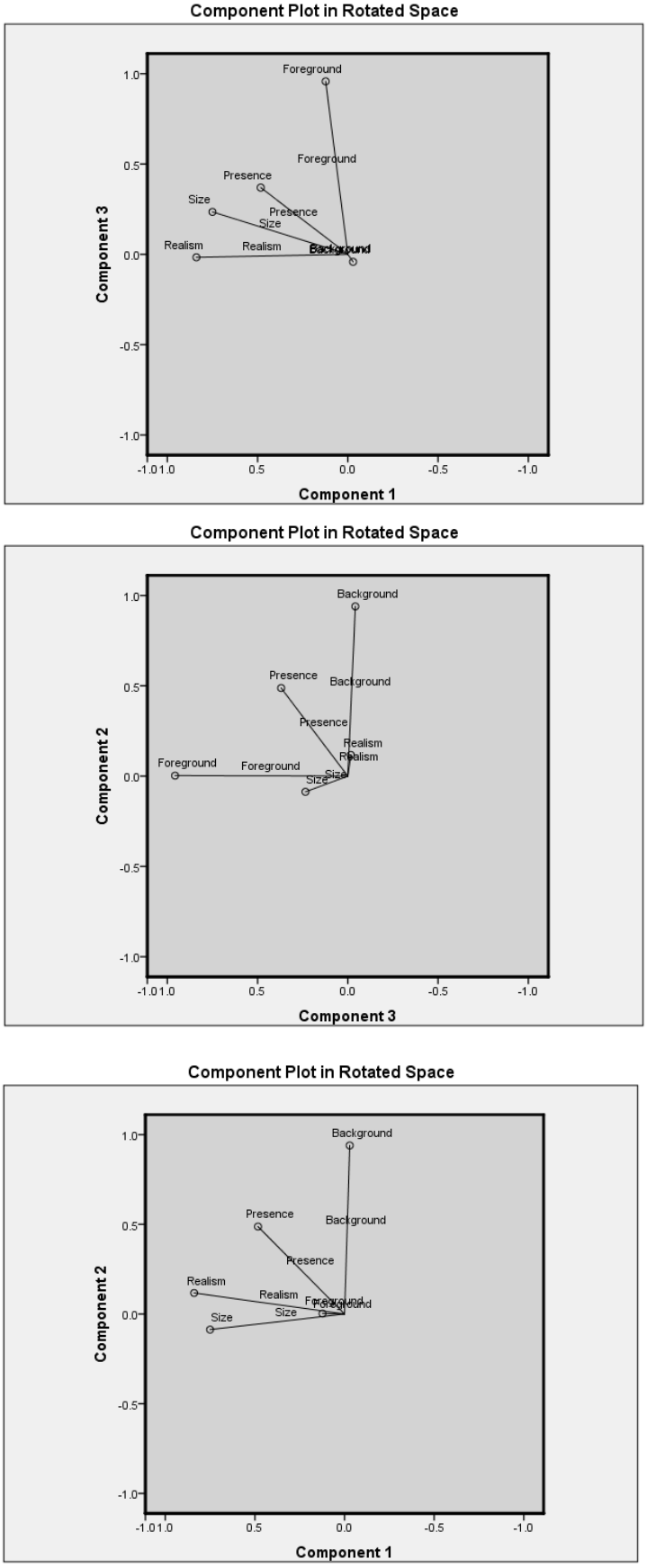


Figure 9.12 A, B, C Principal Component Plots of Attributes for Ambisonic and VBAP 2D and 3D Systems

It was of interest to observe the underlying dimensions separated by rendering method, therefore a PCA was firstly conducted for the ambisonic surround systems.

9.6.2 Ambisonic Surround Systems

For the ambisonic surround systems an initial PCA was run, the resultant scree plot can be seen below in *Figure 9.13*.

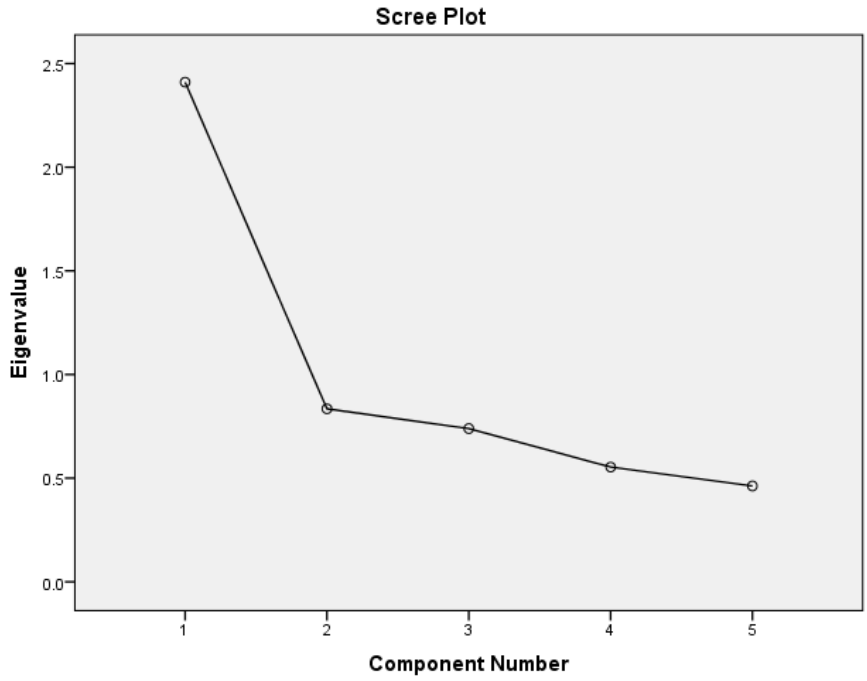
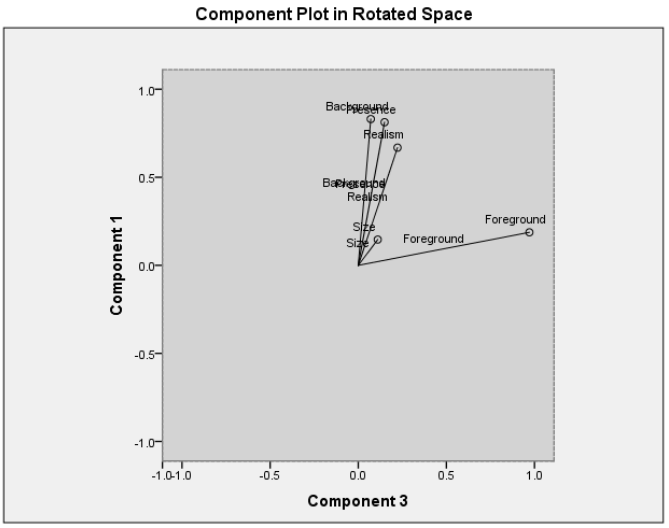
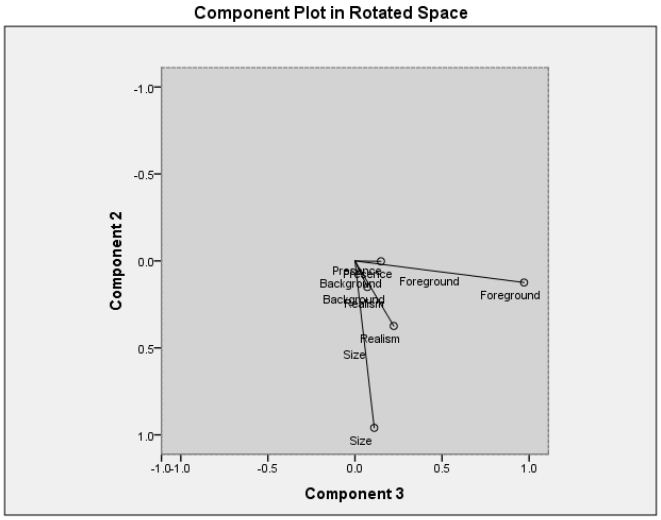


Figure 9.13 Scree Plot of Component Number versus Eigenvalue (Ambisonic Surround Systems)

It can be observed that there is a point of inflexion at the second component, however the third component was >.7, therefore the PCA was rerun this time to extract 3 factors. It was found that three factors explained 79% of the variance, the first factor explained 48% of the variance, the second 16% of the variance and the third 14%.

Again a varimax rotation was used to help in the interpretation of the principal components by maximising the loadings onto each component (Field, 2009). This showed that the attributes *background*, *presence* and *realism* loaded onto the first factor, with *size* loading onto the second component and finally *foreground* onto the third component. The component plots can be seen below in *Figure 9.14.*





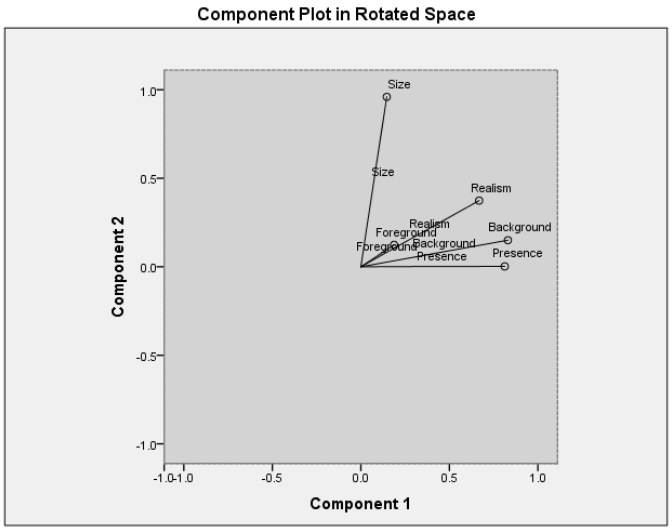


Figure 9.14 A, B, C Principal Component Plots of Attributes for Ambisonic Surround Systems

9.6.3 VBAP Surround Systems

The same approach was used as before for the VBAP surround systems, the intial PCA was run in order to identify the number of components to extract. The scree plot from the initial PCA can be seen below in *Figure 9.15.*

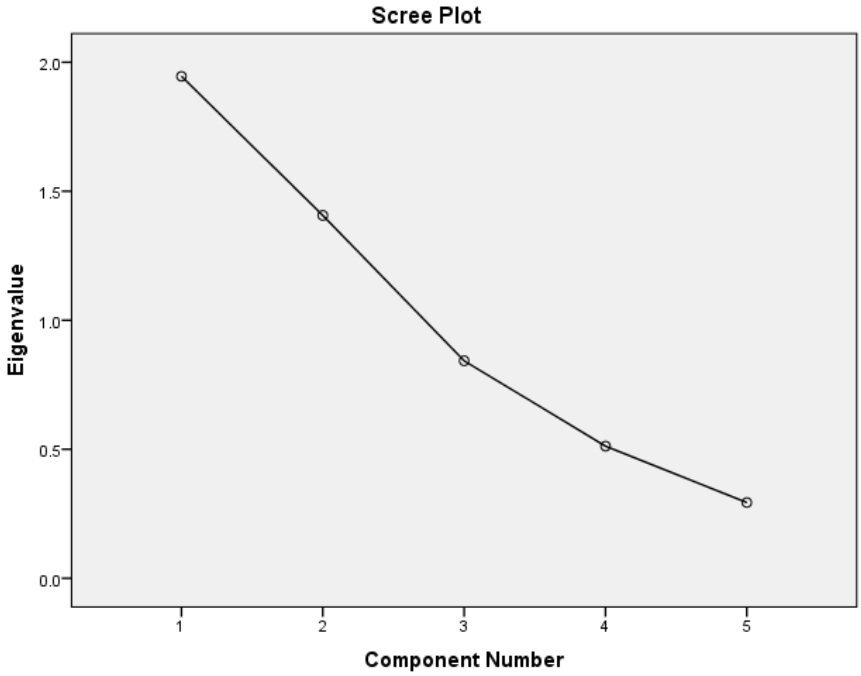


Figure 9.15 Scree Plot of Component Number versus Eigenvalue (VBAP Surround Systems)

It can be seen that there is not a definite point of inflexion, however there are again three components with a >.7 eigenvalue and in fact two with a >1 eigenvalue, therefore the PCA was rerun to extract 3 factors and it was found that the three factors explained a total of 83% of the variance. The first factor explaining 38% of the variance, the second 28% and the last 16%.

A varimax rotation showed that *realism* and *background* loaded onto the first component, the second component was *size* and *presence* and finally the third component was *foreground. See Figure 9.16.*

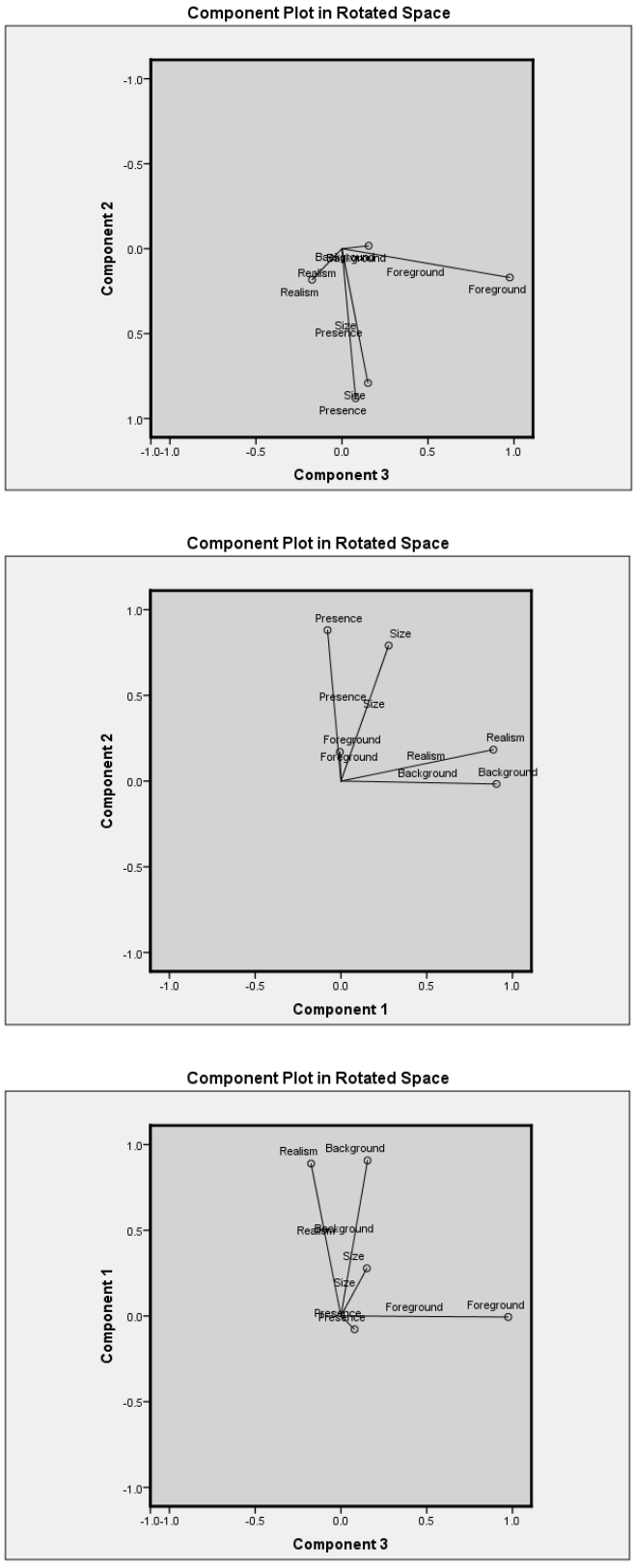


Figure 9.16 A, B, C Principal Component Plot of Attributes for VBAP Surround Systems

It would seem from the PCA of the ambisonic surround systems and the VBAP rendered surround systems that there is a slightly different grouping of attributes. This was expected since there were different correlations between attribute scales for ambisonic and VBAP surround systems. Common to both rendering methods is the fact that for the VBAP systems *realism* and *background* load onto the first factor and the same is observed for the ambisonic systems with the addition of *presence*.

Further analysis has been carried out using the PCA analysis, this time the 2D surround systems and 3D surround systems have been separated in order to investigate the underlying differences between the 2D and 3D systems.

9.6.4 Ambisonic 3D Surround System

The data for the ambisonic 3D surround systems was combined for both stimuli, this procedure was also used for the horizontal only surround systems.

For the ambisonic 3D surround systems an initial PCA was carried out, the resultant scree plot can be seen in *Figure 9.17*.

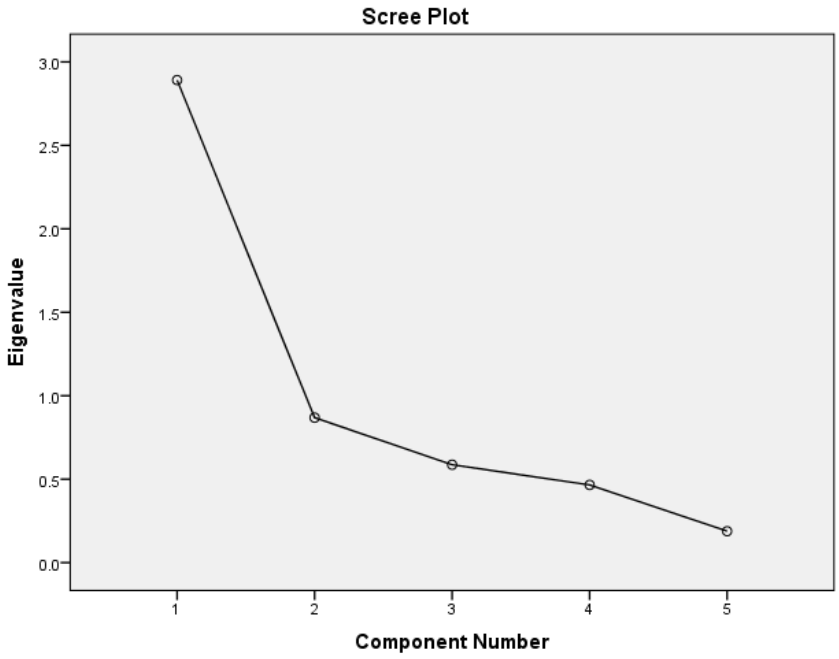


Figure 9.17 Scree Plot Showing Eigenvalue versus Component Number

Observing the scree plot it was not so clear as to the number of components to extract, since there was a point of inflexion at two components and also at four, however the eigenvalue of the third component was <.7 therefore the PCA was rerun to extract two factors. It was found that the two components which had eigenvalues of >.7 explained a total of 75% of the variance, with the first factor explaining 57% of the variance and the second 17% of the variance.

The varimax rotation showed that the factors which loaded onto the first component were *realism*, *presence*, *size* and *background* with the second factor being *foreground*. This can be seen in the component plot below in *Figure 9.18*.

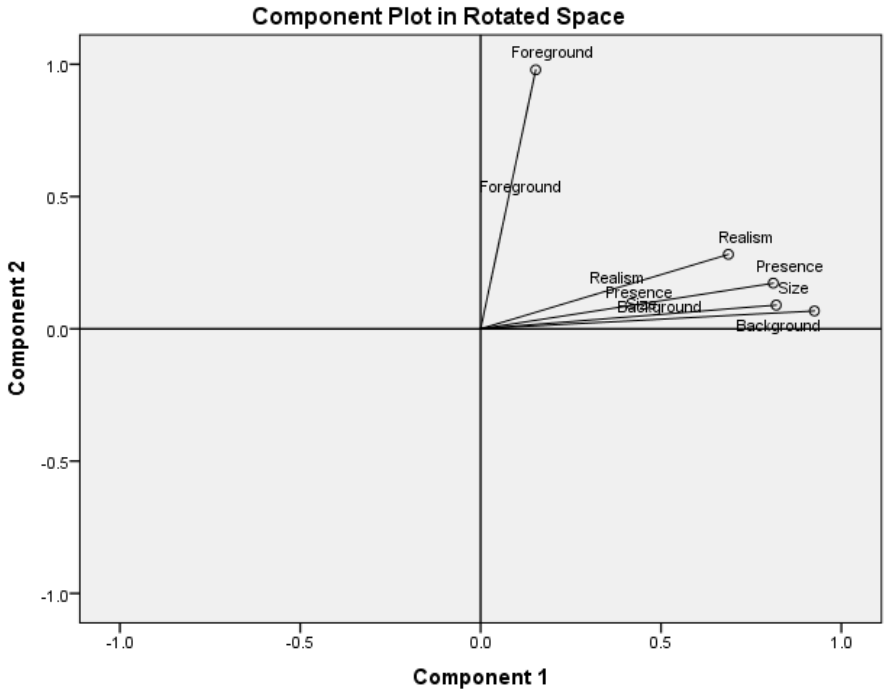


Figure 9.18 Principal Component Plot for Ambisonic Height Systems using Varimax Rotation

9.6.5 Ambisonic 2D Surround Systems

As before an initial PCA analysis was run, the resultant scree plot can be seen below in *Figure 9.19* which guided the decision as to the number of factors to extract.

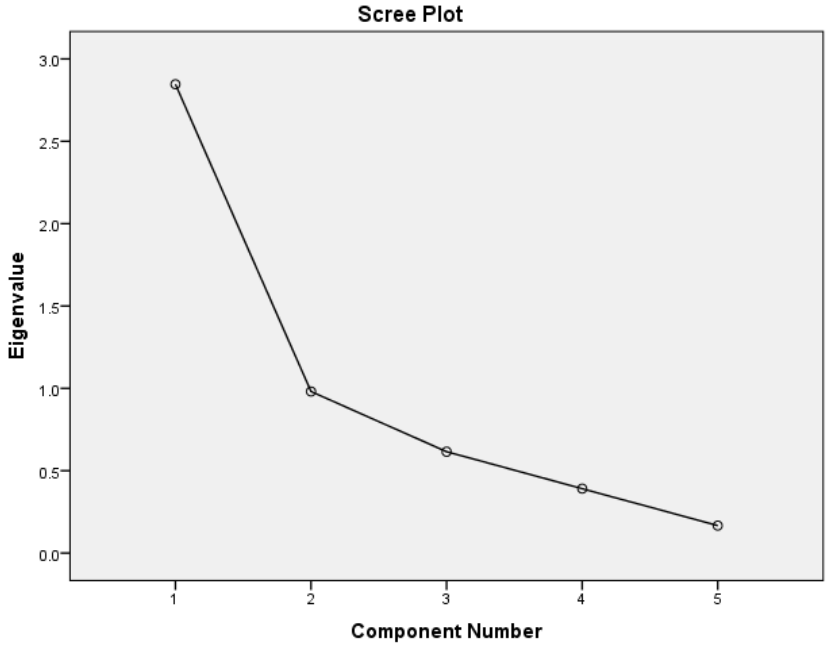


Figure 9.19 Scree Plot Showing Eigenvalue versus Component Number

Similar to the previous outcome the initial scree plot was not so clear to interpret regarding the number of components to extract, however a decision was made two extract the first two factors since these had eigenvalues >.7 (Jolliffe, 2005).

The PCA analyses for the horizontal ambisonic surround systems was run again to extract two factors. The two factors explained a total variance of 75%. The first factor explaining 57% of the variance and the second factor explaining 17% of the variance.

The varimax rotation showed that the first factor which loaded highly onto the first component was *size*, *presence* and *background* and again *foreground* on the second component. It appears that slightly different to the ambisonic with height system the attribute *realism* appears to be a combination of the two factors. *See Figure 9.20.*

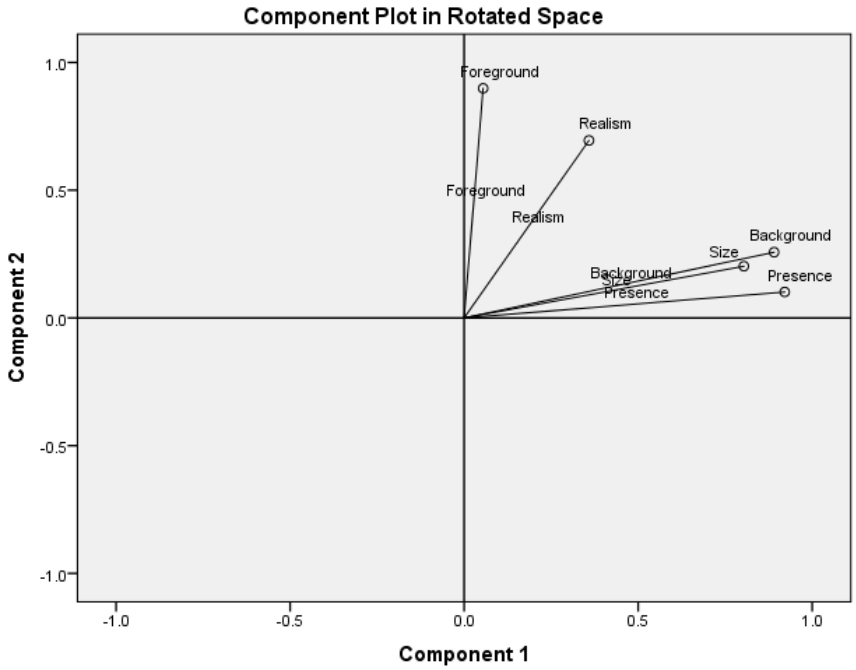


Figure 9.20 Principal Component Plot for Ambisonic Horizontal Surround Systems using Varimax Rotation

The principal component analysis in *Figure 9.18* for the ambisonic 3D surround systems has revealed that four factors load onto the first component, it is interesting that for the horizontal surround systems realism is shifted from the second principal component and as suggested by the component plot is a combination of the two components.

9.7 VBAP Systems

Firstly the PCA analysis is presented for the nine channel with height VBAP system.

9.7.1 VBAP 3D Surround System

As before an initial PCA was run and the resultant scree plot can be seen below in *Figure 9.21*.

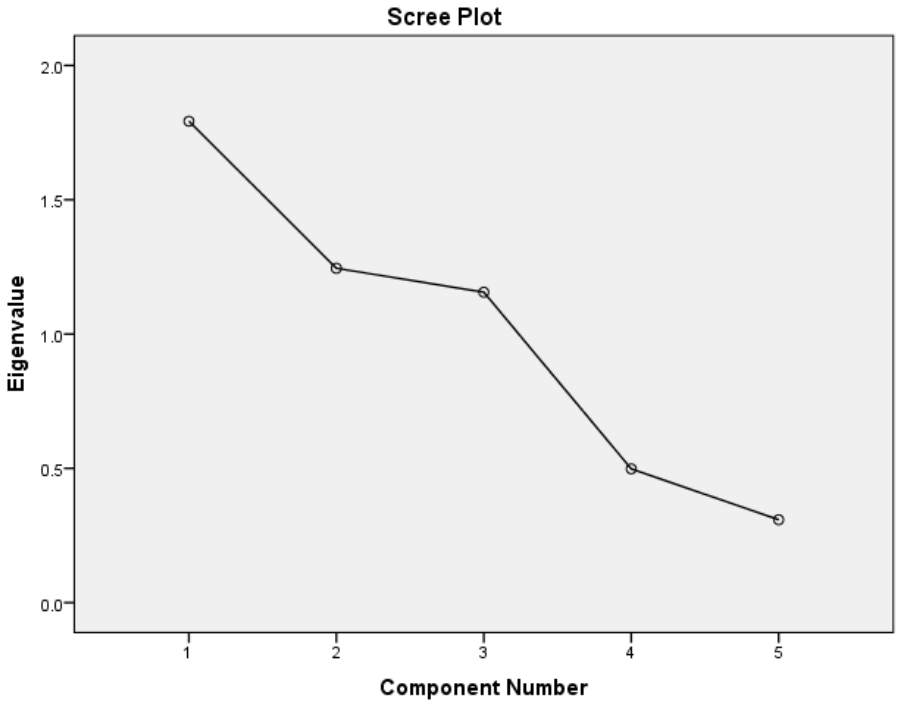
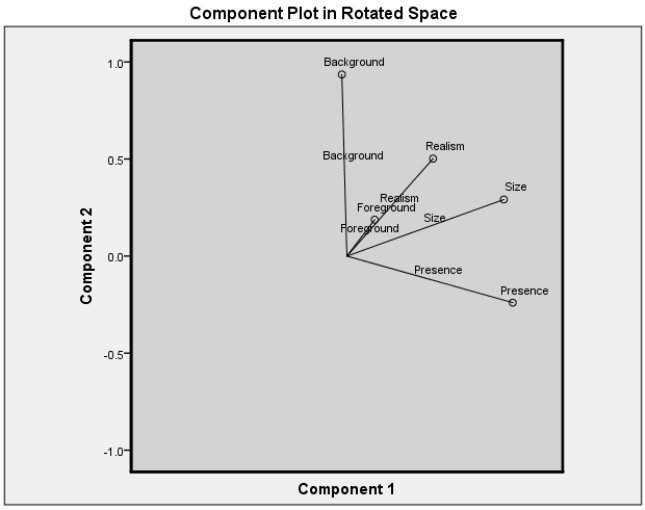


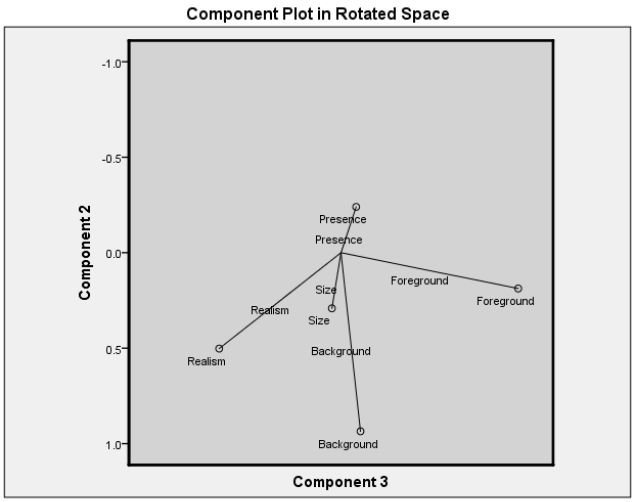
Figure 9.21 Scree Plot Showing Eigenvalue versus Component Number

The scree plot seems to show a definite dip at the third component, it has been detailed by (Field, 2009) that the component at the point of inflexion should be discarded and everything to the left of the inflexion point kept, however it seems important that the third component be kept since it has an Eigenvalue >.7.

The PCA analysis was rerun specifying the extraction of three components which accounted for 83% of the total variance. The first component accounted for 35% of the variance, the second component accounted for 24% of the variance and finally the third component accounting for 23% of the variance.

A varimax rotation showed that *presence*, *size* and *realism* loaded onto the first component, with the attribute *background* loading onto the second component and lastly the *foreground* loading onto the third component. *See Figure 9.22*.





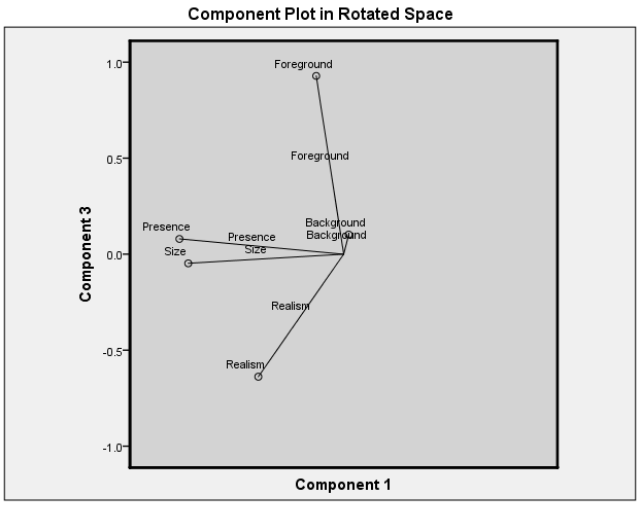


Figure 9.22 A, B, C Principal Component Plot for VBAP with Height System

9.7.2 VBAP 2D Surround System

The PCA analysis was run for the VBAP 2D surround systems, the resultant scree plot can be seen below in *Figure 9.23*.

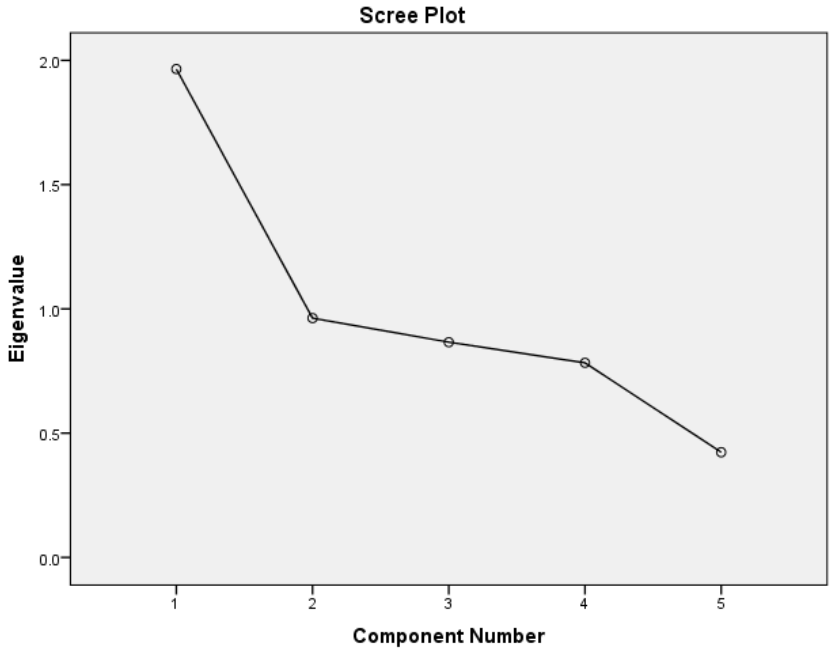


Figure 9.23 Scree Plot Showing Eigenvalue versus Component Number (VBAP 2D Surround System)

It can be seen from the scree plot that there is a definite break at the fourth component, this would suggest in retaining three factors.

The PCA was rerun in order to extract three factors. The first factor explained 39% of the variance, the second factor explained 19% of the variance and the final factor explained 17% of the variance. Using the varimax rotation this showed that the first factor included the attributes *foreground*, *background* and *realism*, the second component included the attribute *presence* and finally the third component included the attribute *size*. The component plot can be seen below in *Figure 9.24.*

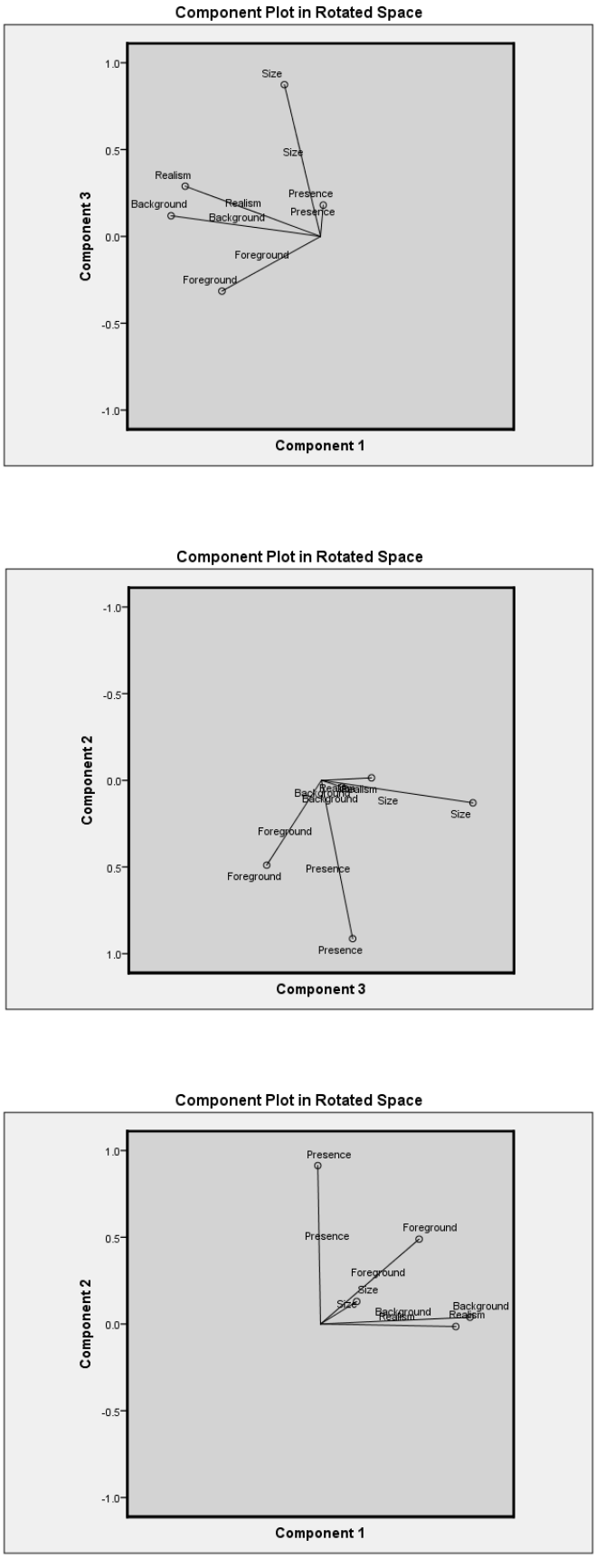


Figure 9.24 Principal Component Plot for VBAP 2D Surround System

9.7.3 Subjective Results Summary

Overall across all samples and rendering methods it is apparent that the listeners can discriminate between the mono and stereo systems and between the surround systems to a statistically significant degree using the derived scales. Five attribute scales were used *foreground, background, realism, presence* and *size*. The results have shown that there is no significant difference between the horizontal surround systems and the systems with height for four of the attributes for either the ambisonics or VBAP rendered soundscenes. It was discussed in *Chapter 8 Section 8.4.1* that less frequent terms like *roomy* or 'sense of room' were used to describe the surround systems with height and other more frequent terms were *open* and *spacious*. It was speculated that surround sound with the inclusion of height could convey to the listener a better sense of the space in which the material was recorded compared to surround systems in the horizontal plane. This seems to have been confirmed in the subjective test with the attribute *size* showing significant results for the surround systems with height, for not only the ambisonic with height systems, but also the VBAP with height systems which used a different set of samples and rendering methods.

Furthermore the majority of participants who took part in the subjective test described in this chapter were not part of the focus group, but could still understand the attributes derived from the focus group and use them in the subjective test. In addition, they were able to rate the surround systems with height as providing the best perception of the attribute *Size*.

Since there are a number of ways of rendering spatial audio and a number of speaker layouts which can be used, it is necessary to derive scales that can have a wide application. The analysis of attribute scales was started by combining all of the data for the ambisonic and VBAP 2D and 3D surround systems. This showed that there were weak correlations between attribute scales.

Correlations were then carried out for the surround systems seperated by rendering method and it was found that the correlations for the ambisonic surround systems showed that *background, presence* and *realism* were moderately correlated, whilst for the VBAP systems the attributes *size* and *presence* were moderately correlated and similar to the ambisonic systems *background* and *realism* were also moderately correlated.

In order to compare between 2D and 3D systems and to investigate if there was any redundancy between scales and to observe the dimensions of the space, a separate PCA was run for the ambisonic 2D and 3D surround systems and also for the VBAP 2D and 3D systems.

For the ambisonic surround systems 2D and 3D two principal components were extracted for both the horizontal and with height systems. The first principal component explaining 57% of the variance and being related to the attributes *background*, *realism*, *presence* and *size*. This suggests that the scales had been used in a similar manner.

A slightly different result was found for the VBAP rendered systems since there was less correlation between the attributes. In addition using the PCA analysis three factors were extracted for the with height system and the system in the horizontal plane which showed an equal distribution of the variance across all three factors, suggesting that the attribute scales were more orthogonal compared to the ambisonic 2D and 3D systems.

In terms of the ambisonic surround systems these results suggest that there is redundancy between scales, however using the VBAP rendered soundscenes the PCA does not suggest this. In light of these results it is apparent that decisions whether scales can be reduced cannot be made and that further testing is required to establish the performance of the scales using soundscenes of similar composition to that of the soundscenes supplied by the BBC.

9.8 Objective Measure IACC

The IACC is an objective measure which has been shown to be related to the perception of spatial attributes like spatial impression, envelopment and spaciousness, it has been used primarily in concert hall acoustics (Beranek et al., 1996) and in reproduced sound (Mason and Rumsey, 2002), (Hirst, 2006), (Choisel and Wickelmaier, 2006b). In the previous *Chapters 6* and *7* the IACC has been used as an objective measure and has been correlated to mean subjective scores of envelopment, the measure has also shown that by reproducing the soundfield from additional directions, for example using height channels contributes to a lower IACC measure. It is stated by Mason and Rumsey (2002) that of the acoustical measurements stated in the international standard (ISO, 2000) the IACC is the objective measure most likely to be successful in reproduced sound.

It was of interest to discover if the IACC was correlated with any of the attribute scales used in the previous subjective evaluation, since these scales relate to spatial aspects of the reproduced sound scene such as the spatial distribution, which refers to how the sound scene is distributed around the listener, similar to apparent source width or envelopment.

To measure the IACC the HATS (head and torso simulator) was placed at the listening position in the semi anechoic chamber. Each of the sound scenes was reproduced over each of the respective systems and captured using RME interface which was connected to a laptop. *See Figure 9.25.*

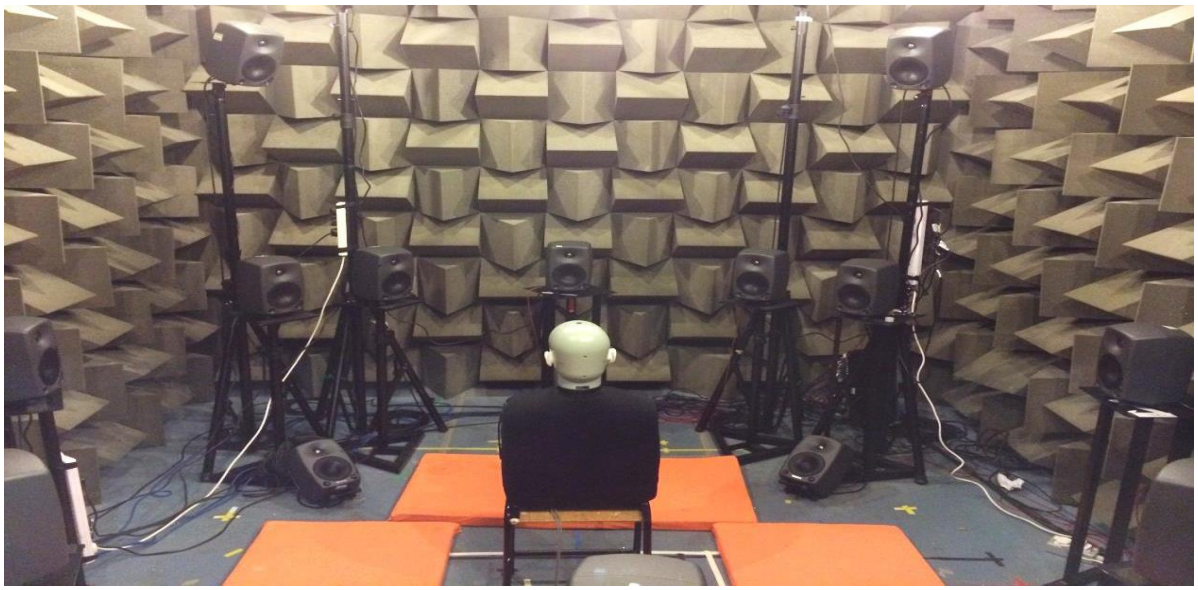


Figure 9.25 Binaural Recording of Reproduction Systems Using the HATS

The resultant binaural recordings were then analysed to extract measures of the IACC using the method outlined in *Chapter 3, Section 3.10.1*. *Table 9.32* below shows the IACC for each of the ambisonic systems and stimulus.

System	IACC	IACC
	Applause	Choir
Mono	0.96	0.98
Stereo	0.28	0.68
5.1	0.32	0.65
Octagon_1st Order	0.22	0.34
Octagon_3rd Order	0.24	0.5
Octagon_Cube_1st Order	0.21	0.42
Octagon_Cube_3rd Order	0.21	0.44

Table 9.32 Linear IACC Value for each System and Stimuli (Ambisonic Systems)

For the ambisonic systems it can also be seen that the value of the IACC for the two 2D octagon layouts have a difference of 0.02 whilst the 1st order and 3rd order 3D Octagon_Cube have identical values using the applause sample.

Whilst for the choir sample the difference in IACC between the two horizontal octagon systems is much greater with a difference of 0.16 and only a difference of 0.02 between the two 3D surround systems. The same method was also used for the systems which were rendered using VBAP. It is also interesting to note that the IACC values for the VBAP systems using the music sample did follow an expected trend with the values of IACC decreasing with increasing number and positioning of speakers, *See Table 9.33.*

System	IACC	IACC
	Drama	Music
Mono	0.94	0.92
Stereo	0.87	0.8
Five	0.9	0.72
Nine	0.88	0.62

Table 9.33 Linear IACC Value for each System and Stimuli (VBAP Systems)

The values of IACC for the VBAP systems overall are very high indicating a more correlated soundfield at the listeners ears, this could have been because of the strong centre channel element in both of the samples, especially the drama sample which in addition to the narrator which was placed in the centre channel, towards the end of sample there is a voice over by the presenter. This would have increased the similarity in the signals reaching the HATS system which would have led to the higher IACC values, this phenomenon is explained by

(Hidaka et al., 1995) where a sound wave arriving from straight ahead will excite the two ears in a similar manner. Further evidence to support this theory can be seen when observing the stereo IACC value which is slightly lower than the 2D surround system (Five) and the 3D surround system (Nine), due to the fact that the stereo system would use a phantom centre rather than the dedicated centre channel used for the other systems.

Overall it can be observed that the lowest values of IACC are for the systems with height. This is true for both the ambisonic systems and the VBAP systems, with the exception of the 1st order octagon using the choir sample which had the lowest IACC, and for the drama sample where the lowest IACC was for the stereo system.

9.8.1 Correlation between Objective and Subjective Measures

Five different attributes were used derived from focus group discussions, for each of the attributes a correlation coefficient has been calculated between the average subjective score and the linear IACC value, to identify if there is an attribute to which IACC has the strongest correlation.

It can be seen that for the samples rendered using ambisonics, the *foreground* attribute subjective scores have the strongest correlation with the linear IACC. *See Table 9.33.* This is also true for the VBAP rendered material using the drama sample, however overall the strongest correlations with the subjective scores were for the VBAP rendered music stimuli and in particular the attributes *size* and *realism*. *See Table 9.34*

Stimuli	Attribute	Correlation Coefficient	Significance p<0.05
Applause	Foreground	-0.89	0.00
	Background	-0.87	0.01
	Presence	-0.83	0.01
	Size	-0.88	0.00
	Realism	-0.84	0.01
Choir	Foreground	-0.81	0.01
	Background	-0.76	0.03
	Presence	-0.75	0.03
	Size	-0.71	0.04
	Realism	-0.75	0.03

Table 9.34 Correlation between Mean Subjective Score and Linear IACC for Each Attribute using Pearsons Correlation Coefficient (Ambisonic Systems)

Stimuli	Attribute	Correlation Coefficient	Significance p<0.05
Drama	Foreground	-0.93	0.03
	Background	-0.81	0.09
	Presence	-0.72	0.14
	Size	-0.82	0.09
	Realism	-0.79	0.11
Music	Foreground	-0.95	0.03
	Background	-0.95	0.03
	Presence	-0.95	0.02
	Size	-0.96	0.02
	Realism	-0.96	0.02

Table 9.35 Correlation between Mean Subjective Score and Linear IACC for Each Attribute using Pearsons Correlation Coefficient (VBAP Systems)

It can be observed that there are high correlations between the objective measure and the mean subjective score for each attribute scale. In order to investigate the correlations further the mono system was removed from the analysis and the correlations re examined. *See Tables 9.36 and 9.37.*

Stimuli	Attribute	Correlation Coefficient	Significance p<0.05
Applause	Foreground	-0.72	0.10
	Background	-0.54	0.27
	Presence	-0.64	0.17
	Size	-0.33	0.53
	Realism	-0.51	0.31
Choir	Foreground	-0.27	0.60
	Background	-0.93	0.86
	Presence	-0.13	0.81
	Size	0.00	1.00
	Realism	-0.15	0.78

Table 9.36 Correlation between Mean Subjective Score and Linear IACC for Each Attribute using Pearsons Correlation Coefficient (Ambisonic Systems) mono system removed

Stimuli	Attribute	Correlation Coefficient	Significance p<0.05
Drama	Foreground	-0.56	0.62
	Background	0.65	0.55
	Presence	0.59	0.60
	Size	0.24	0.84
	Realism	0.47	0.69
Music	Foreground	-0.55	0.35
	Background	-0.85	0.35
	Presence	-0.86	0.34
	Size	-0.88	0.31
	Realism	-0.88	0.31

Table 9.37 Correlation between Mean Subjective Score and Linear IACC for Each Attribute using Pearsons Correlation Coefficient (VBAP Systems) mono system removed

It is interesting to observe overall that the removal of the mono system from the correlations dramatically reduces the correlations. For the ambisonic systems using the applause sample the strongest correlation are for the *foreground* attribute, whilst for the choir sample this is for the *background* attribute. Looking at the correlations between the IACC and the mean subjective score for the VBAP systems show that for the drama sample the highest correlation is for the *background* attribute, whilst for the music sample the highest correlations are for the attributes *size* and *realism*. What can be noticed looking at the music stimulus in Table 9.37 is that across all attributes with the exception of *foreground* the IACC has maintained a strong negative correlation with the mean subjective score.

9.8.2 Objective Measures Summary

The initial correlations for the IACC measures have shown a strong negative correlation with the mean subjective score, the strongest statistical correlation was with the VBAP rendered music. However the VBAP rendered sound scenes showed the highest IACC as previously mentioned, possibly due to the strong centre channel component in both the drama and music sound scene.

Further analysis of the correlations was carried out by removing the mono system from the analysis. This showed that the strength of the correlations was reduced with the highest correlation being found for the ambisonics rendering in terms of the *background* attribute

using the choir sample. Overall the music sample rendered using VBAP maintained high correlations between all of the attributes with the exception of the *foreground* attribute.
The IACC measurement method used was a simplified version which does not take into account frequency weightings or seperation of sources based on background and foreground content. Therefore caution is needed when interpreting the results. It may be that other variations of IACC measure might be more suitable like the method proposed by Mason (2002) the perceptually grouped interaural cross correlation (PGIACC) which includes a more detailed model of the hearing system and measures seperately the IACC based on the source related content and the environment related content of the sound scene.

A number of metrics were also combined to measure spatial quality in the QESTRAL project (Rumsey et al., 2008) generally these were based on variants of the IACC, like the IACC0 which was the mean value of IACC calculated across 22 frequency bands with the dummy head facing forward. In addition the IACC90 was also introduced which was the same calculation as the IACC0 but with the dummy head at 90° to the source, the product of the two measures was also combined IACC0*IACC90 as this gave an indication of how diffuse the soundfield was in 360°. In addition to the variants of IACC in the QESTRAL study was the EntropyL and the TotEnergy, the EntropyL metric was measured from the left ear signal of a dummy head recording and was designed to measure the change in soundfield density, since it was found by Conetta (2011) that this had an effect on the perceived direct envelopment, the TotEnergy metric measured the overall playback level and was the root mean square of the pressure value measured by a pressure microphone placed at the listening position and was inspired by the work of Soulodre et al. (2002) who found that the overall playback level had an effect on the perceived envelopment.

Further investigation using one or a combination of the previously described metrics may provide better results, but more importantly provide the ability to discriminate between the attributes utilised.

9.9 Criticisms of the Method

There are a number of areas within this work which could have had an effect on the results. Firstly only four of the original focus group panel were able to take part in the subjective test, this would not have had a direct effect on the results but unfortunately precludes observing whether the variance for each of the scales was reduced as a result of the focus group. In

addition only four different types of sound source were used. This has limited the generalisability of the results. It would have been better to have used a more varied selection of audio samples. No training was given to the participants regarding the five attributes, it possibly would have been better to provide training to each listener to reinforce their understanding. However this would have increased the time of the test.

9.10 Chapter Discussion

The reported work in this chapter has concluded the descriptive analysis process. A number of systems which have included surround sound systems in the horizontal plane and with height systems have been assessed using five descriptors derived from a descriptive analysis.

Using the five attribute scales it has been observed that the use of the *size* scale, derived to scale the perceived size of the reproduced sound scene, has shown that the inclusion of height channels provides a significantly better sense of the size of the reproduced space than surround in the horizontal plane only. This outcome has been established using two different rendering methods and stimuli. In the case of the ambisonic system this was only true for the higher order system using the applause sample, whilst for the VBAP rendered system this finding was true for the drama sample.

It has also been revealed that for one of the attributes there seems to be an advantage in using higher order ambisonics, since it has been found that for the attribute *presence* increasing the ambisonic order has the same effect on the perception of presence as the addition of height channels.

Subjective results summary:

- *Foreground* attribute: no significant improvement over stereo with the addition of height
- *Background* attribute: no significant improvement over 2D surround with the addition of height
- *Presence* attribute: significantly enhanced with 3D (HOA) higher order ambisonics using applause sample compared 2D surround systems
- *Realism* attribute: 3D HOA system significantly adds to the perceived realism using applause sample compared to 5.1 system

- *Size* attribute: 3D HOA using applause sample and 3D VBAP using drama sample provides a better sense of the reproduced space compared to 2D surround systems

9.10.1 Independence of Scales

It is necessary to have attribute scales which can be used to provide subjective measures over a wide range of systems allowing comparisons of a number of systems. Therefore an initial investigation into the correlation between scales was carried out, firstly combining the attribute scores across all surround systems for the ambisonic and VBAP systems. This showed that there were weak correlations between attributes.

For further comparison, correlations between attributes separated by rendering method was then carried out between ambisonic and VBAP rendered surround systems. This showed that there were slightly different correlations between attributes, although it was found that for the different rendering methods the attributes *background* and *realism* were moderately correlated which was common a common factor.

It is also of interest to utilise the PCA to inspect the perceptual space of the 2D and 3D surround systems seperately.It appears for the ambisonic rendered 3D surround systems and also the 2D surround systems in the horizontal plane, there have been similar moderate to strong correlations between the attributes *presence*, *size*, *realism* and *background*. Although for the 3D surround system using the VBAP rendering the correlations between the attributes *presence*, *size* and *realism* were moderate and the correlations between attributes for the VBAP rendered 2D surround system the only correlations were between *realism* and *background.*

The PCA for the ambisonic 3D surround systems found two components the first was related to the attributes *background*, *size*, *realism* and *presence* and the second to *foreground*. There was a similar finding for the ambisonic 2D surround system with the exception that *realism* was shifted more towards the *foreground* attribute. The PCA for the VBAP 3D surround system extracted three components, the first component contained *presence*, *size* and *realism*, the second component contained *background* and finally the third component contained the attribute *size*. For the VBAP 2D system the first factor was the attributes *realism*, *background* and *foreground*, the second factor was *presence* and the third component was *size*.

One notable aspect for the PCA involving the ambisonic 2D surround systems and also the 3D surround systems was the fact that the percentage of variance explained by the first factor was greatest and was the same 57%. However for the VBAP 2D surround systems and 3D surround systems the percentage of variance explained was almost equally distributed for the three factors extracted, this was expected since the correlations between attributes for the VBAP systems was much weaker. Furthermore, the weaker correlations between attributes for the VBAP rendered systems suggest that the scale usage for these systems was more orthogonal. One explaination for this is that the soundscenes for the ambisonic systems were relatively simple and included clearly distinguishable *foreground* and *background* content. On the other hand the VBAP rendered soundscenes were more complex as these had been mixed and produced by the BBC soundmixers. This may have meant that listeners were using other decisions when scaling attributes because they had been produced and mixed differently to that of the ambisonic sound scenes. It was reported by Berg and Rumsey (2006) that complex stimuli used in their listening test may have affected their results by causing listeners to focus on different parts of the auditory scene. This would explain the weaker correlations and the distribution of the variance in the PCA for the VBAP systems.

9.10.2 Relationship between Attributes

It has been described by Blauert (1996) that when a soundfield is presented to a listener, the listener spontaneously arrives at a conceptual image regarding the type, size and properties of the real or reproduced space. In addition to this Zacharov and Koivuniemi (2001) state that the size and type of acoustic space also plays a role in the perceived spatial sound quality. Accounting for the strong correlations between attributes for the ambisonic systems in this work, Rumsey (2002) defines presence as the sense of being in an enclosed space and states that this implies that the listener is able to sense the boundaries of the space around them and feel present in the space. In addition Berg and Rumsey (2006) have discussed the fact that in their work subjects made the distinction between being in the room with the sound source and the fact that the sound source was being reproduced through speakers using the term *naturalness* this could be interpreted as the *realism* attribute in this work, furthemore it is also explained by Berg and Rumsey that it is also expressed by the term *presence*, he also states "The impression of the room seems to be essential for the feeling of presence". This follows the trend found in this work where there has been high correlations between the attributes *presence* and *size*. Further Gabrielsson and Lindström (1985) also make reference to room related attributes describing the attribute 'Spaciousness' their definition meaning: "that the

reproduction is spacious, that it sounds open, has breadth and depth, fills up the room, gives a feeling of presence, the opposite is a reproduction that sounds closed, shut up, narrow, without feeling of presence". It is interesting to note that the second most used term for the 3D surround systems during the elicitation was the term 'spacious', and other terms were 'open' and 'natural'. In addition Letowski (1989) defines the term 'spaciousness' which he states is "the attribute of an auditory image in terms of which the listener judges the distribution of sound sources and the size of the acoustical space". It is also interesting to note that a subjective test carried out by Avni et al. (2013) using binaural reproductions found that as the ambisonic order increased the reproduced room was perceived as more spacious.

9.10.3 Attribute Scales

Similar attribute scales have been realised by other researchers using descriptive analysis, particularly relevant is the work of Zacharov and Koivuniemi (2001) who used a number of recording techniques, including ambisonics recording and reprodcution systems, which included a 3D surround system. From their structured elicitation attribute scales were realised which were similar to this study, like naturalness, broadness and sense of space, the descriptions of these attributes can be seen in *Table 3.4.* These attributes were comparable with the attribute scales derived in this study like *realism*, *size* and *spatial distribution.*

The attribute scale *spatial distribution* was realised from the descriptors which encompassed *surrounding*, *surrounded*, *wide*, *distributed*, *behind*, *spread* and also *envelopment*. These demonstrate that a number of terms are used to describe the perception of the soundfield relative to the listener and could be interpreted as envelopment or in some cases apparent source width, where the term wide or spread has been used.

In the previous *Chapters 6* and *7* the attribute envelopment was used explicitly to asses a number of reproduction systems which included 2D and 3D surround systems using a more general term to cover the perception of envelopment in multichannel reproduced sound . It is suggested from the elicitation and subsequent focus group sessions, that the spatial distribution scale could be considered a more complete rating scale which captures the sensation of envelopment in its various guises and from the grouping of the descriptors of the focus group listening panel would capture more information about the sound scene compared to the envelopment scale used in this work. Further to this the classification of the spatial distribution into *foreground* and *background* is particularly important, since multichannel

reproductions are capable of reproducing two types of envelopment according to Conetta (2011) which are direct envelopment and indirect envelopment, indirect envelopment is described as originating from the reverberant part of the soundscene, whilst direct envelopment originates from placing anechoic or dey sources around a listener.

It was discussed during the focus group sessions that listeners could use two modes of listening when auditioning the soundscenes, one where they could decide to listen to the direct sounds or the sounds which are most prominent, or they could decide to focus on the reverberant or indistinct parts like room tone and that these were based on foreground and background content. Originally this is described as perceptual grouping or auditory streaming and are based on the theories of Bregman (1990) which is an analysis of the way human hearing organises sound scenes into events or objects. It has also been discussed by Darwin (1991) where he refers to these as primitive grouping mechanisms, where the listener does not need to be familiar with what they are going to hear and that these mechanisms were used in evolving humans in order to be able to seperate speech from background sounds. These theories have also been utilised in the work carried out by Zielinski et al. (2003) where they have classified soundscenes based on their composition in terms of foreground/background content. Griesinger (1997) also bases his theories on the different types of spatial impression on the theories of Bregman, in that background events, which is the reverberation of the acoustical environment, leads to what he terms as spaciousness and envelopment and that they naturally occur in a large space. This would tie in with the correlation of the *background* attribute with the previously described *size*, *presence* and *realism* attributes. Furthermore, it can also be observed that for the attributes *foreground* and *background* there has remained weak correlations across all test conditions, further evidence that listeners have used these two methods of analysis.

In terms of the *size* scale it could be argued that the ratings using this scale to a large extent would be dependent on the soundscene used, for example a soundscene recorded in a small non reflective environment even though it was reproduced over a 3D system would still be rated as small regardless whether it is reproduced over a 3D or 2D system.

What does seem to be clear is that there is a connection between *presence* and *size*. As previously highlighted in *Section 9.10.2* by Rumsey (2002) that presence is defined as the sense of being in an enclosed space and being able to sense the boundaries. Therefore these

scales seem logical and in terms of the *size* attribute have found differences between the 2D and 3D systems.

In terms of the *realism* scale it may not always be appropriate, since soundscenes may be created for use over 3D systems but the main aim might not be to reproduce a sound scene which is true to life. However, for this study which used recorded soundscenes in real spaces this was relevant. However what became apparent is that the drama sample supplied by the BBC R&D, which was not part of the initial elicitation, was created from scratch and was ficticious therefore the *realism* scale might not be appropriate, this is also stated by Rumsey (2002) that to aim for an accurate reconstruction of a natural soundfield is missing the point.

9.10.4 PCA Grouping of Attributes

It would seem from previous research that there is a connection between the definiton of spaciousness and the attributes size, presence and naturalness or realism. The PCA output suggests combining the attribute scales for size, realism and presence in terms of the ambisonic height systems, since these attributes come out as the first principal component and explain the largest percentage of the overall variance for these systems. However as previously discussed the justification for this in terms of the VBAP systems are not so clear since the attribute scale usage by the listeners according to the correlations and the PCA do not justify this.

In the case of the ambisonic systems if these scales were combined the question remains as to the overall description of the scale. Also as previously explained in terms of the VBAP systems the scale usage was more orthogonal therefore combining the scales in respect to these systems on the basis of the PCA is not justified. However it may be that the attributes *size*, *presence* and *realism* could be combined into a scale of spaciousness based on the previously described definitons. Overall it appears that the scales should be used in further testing since only four different types of sound scene were used and it appears from the results that there are differences which could be due to the differences in composition of the soundscenes, this has been found by a number of other researchers not only in the case of surround with height but also using surround in the horizontal plane only for example (Gabrielsson and Sjögren, 1979), (Guastavino and Katz, 2004), (Choisel and Wickelmaier, 2007), (Guastavino et al., 2007), (Paquier et al., 2011), (Silzle et al., 2011). Therefore further testing would be needed to clarify this point.

9.10.5 Objective Measure

Observing the measured IACC values for the height systems in general these were lower. This suggests greater dissimilarity between the two signals at the listeners ears which is connected to listener preference (Schroeder, 1979). Furthermore IACC has been shown to be highly correlated to aspects of spatial impression which include apparent source width and envelopment (Mason et al., 2005), (George et al., 2006). Looking at the IACC measures for the drama sample, this had high IACC values. This was the only sound scene of the four used which was produced from samples and not recorded in a reverberant environment, this would mean that it did not contain naturally occuring reflections from the performance space. Furthermore, the drama sample was found to provide the best sense of the attribute size but had high values of IACC which is counter intuitive, therefore this suggests that the inclusion of height and the overhead bird sounds was more suggestive rather than based on reverberant cues.

The IACC values of the applause sample, choir sample and of the music were all recorded in reverberant environments therefore the recordings contained reflections naturally occuring in the space leading to lower IACC values. Yet for the Elbow music sample and the choir sample reproduced using the 3D surround systems, these were not judged as being significantly different to that of the other surround systems using the five attribute scales.

9.10.6 Similarity between Samples

Investigating the similarity between the two samples which showed significance for the attribute *size* was that they clearly had audible height components. For example the applause sample which was recorded in the Royal Albert hall London during a Proms concert had clearly audible reflections from all around and also from above due to the energy of the applause to excite the space where the sample was recorded. Whilst the drama sample which had been mixed by BBC producers had clearly the sound of birds overhead, this would have been audible to the listeners. This is a slightly different case to the applause sample, since this was synthetic but would have suggested to the listener an outdoor space with the sky above, which would have accounted for the higher mean score for the *size* attribute. The opposite is true for the choir sample which mainly contained the direct sound of the singers in the front horizontal channels and the room tone all around including above. But because the singers were recorded in a large hall (orchestra hall at the University of Detmold) they were not enough to excite the hall as much as for example as the applause, therefore the rooms reflections were not strong enough to be noticeable in the height channels. In the case of the

music sample this was of the group Elbow and was recorded in the Manchester cathedral, however this piece was further produced by the BBC soundmixers using creative mixing by placing the keyboard sounds and other instruments around the listener and faintly in the height channels. When considering the drama sample as previously mentioned the sound of seagulls in the height channels is suggestive of the sky above a person creating a natural height cue, but the Elbow music sample had the sounds of keyboards and instruments faintly in the height channels which would not be suggestive of height cues. It is also hypothesized that the content in the horizontal channels may have masked the subtle content in the height channels, since they were quieter versions of the horizontal channels, therefore would not have been as noticeable to the listeners compared with the drama sample.

Another difference between the ambisonics and VBAP material is the recording and production of the material. Firstly the ambisonic samples were captured with an Eigenmike which as a result captures the Z component or height elements naturally occuring in the recording space.

It has been reported previously by Guastavino and Katz (2004) that their comparison of surround horizontal and with height system that it was difficult to recommend a system that would suit all types of sound scene. However, they state that their 3D system was more suited to indoor material and 2D to outdoor material and their 1D or stereo system to frontal musical sound scenes. This to a certain extent has been reflected in this work, although there was no real outdoor recorded material, however the drama sample which was an artificially created outdoor scene was significantly different to the surround system in the horizontal for the attribute size, therefore this result does not suggest favouring the horizontal surround system in this case for outdoor sound source material. It was also observed in general that the height systems used to reproduce the choir material received lower mean scores than that of the surround systems in the horizontal. One reason for this could be that as explained by Rumsey (2002) and Guastavino and Katz (2004) that providing listeners with a spatial reproduction which is true to the scene may provide listeners with too much information and be disturbing to them.

9.11 Summary

Five attribute scales derived from a descriptive analysis were used in a subjective test to rate a number of reproduction systems. This included surround systems in the horizontal plane

and also surround systems with height. Using the five scales it has been observed that the attribute scale *size* used to scale the perceived size of the reproduced sound scene has been found to be significant for the with height systems. This has not only been found for the systems using ambsionics as the rendering method, but also for the systems using VBAP. One caveat to this is that for both rendering methods the results have been dependent on the soundscene used. Summarising the attribute scales it was found that:

- *Foreground* attribute: no significant improvement over stereo with the addition of height
- *Background* attribute: no significant improvement over 2D surround with the addition of height
- *Presence* attribute: significantly enhanced with 3D (HOA) higher order ambisonics using applause sample compared 2D surround systems
- *Realism* attribute: 3D HOA system significantly adds to the perceived realism using applause sample compared to 5.1 system
- *Size* attribute: 3D HOA using applause sample and 3D VBAP using drama sample provides a better sense of the reproduced space compared to 2D surround systems

For the 2D and 3D ambisonic surround systems there were strong correlations between the attributes size, presence, realism and background. A subsequent PCA found that the greatest amount of variance was explained by these factors, which would suggest reducing these attributes to one scale.

The correlation between attributes for the VBAP rendered systems were weak and the subsequent PCA found that the variance was spread equally between the three factors. The correlation between attributes and the subsequent PCA suggest that for the VBAP rendered system there are no grounds to reduce the scales into a single scale.

No clear reccomendation can be given as to the type of sound scene suited to surround systems with height, but it does appear from this research that the height content should be distinctive enough so that it can be heard by the listeners.

The outcome of the PCA highlights two different scenarios. It is hypothesized that this is an effect of the the soundscene used, since the soundscenes for the ambisonic recorded material

were relativley simple, on the other hand the VBAP rendered soundscenes were produced by the BBC so were more varied and complex. This would have caused listeners to focus on different parts of the sample and use the attribute scales in a different manner. It is suggested that in order to get a clearer understanding of the scale usage, further testing should be carried out using the five attribute scales to assertain the validity of the scales for use across a large range of spatial reproduction methods and also to assertain if there are any redundant scales.

10 Summary and Conclusions

This chapter provides a summary and conclusions of the research and experimentation that has resulted in this thesis. The research carried out which is detailed in this thesis was to establish the benefits of the inclusion of height in surround sound listening experience. The approach taken involved assessing localisation in with height systems, determining the impact of the height channels on envelopment and the use of more explorative subjective methods to uncover the differences between 2D and 3D surround systems.

10.1 Chapter 2

In *Chapter 2* a number of areas concerned with previous research in spatial audio have been documented. The review started with an account of the development of spatial audio systems from stereo to the surround systems in use today and also the development of 3D surround systems. It was detailed that 3D surround systems or systems with height are not a new concept since they had been proposed before, but had not been fully considered due to the failings of quadraphonics.

The review then moved onto methods of rendering 3D sound scenes over speakers and it was detailed that there are two main methods, VBAP and ambisonics. The strengths of each method was detailed and it was decided to utilise ambisonics for rendering 3D sound scenes since VBAP is solely a panning method and ambisonics offers the ability to record in 3D as well as a panning method and also the flexibility reproduce sound scenes in different resolutions. However VBAP rendered sound scenes are included in the final subjective test since it was necessary to investigate if the derived scales had wider applicability.

Other areas of 3D reproduction were also considered this included the ability of humans to localise elevated sound sources. From previous psychoacoustic research it was detailed that there were a number of peculiarities in human hearing with elevated sound sources. It was highlighted that:

- Virtual sound sources in the median plane would have no interaural differences and would rely solely on the spectral content to resolve its position
- Sources with high frequency content were easier to locate in the vertical plane than sources with only low frequency content

- More cues are available to a listener when sources are placed off the median plane including ITD and ILD and also spectral cues

It was identified that previous research into the ability of listeners to locate sound sources in the median plane using two different resolutions of ambisonics had found that there was no significant difference between the two resolutions. From this the question was asked:

- Does increasing the ambisonic order improve the localisation of elevated sources in the median plane?
- Are there differences in localisation accuracy when virtual sources are placed in the median plane?

In addition there were also further questions in terms of localisation in with height surround reproduction which included:

- Does the use of higher order ambisonics improve localisation accuracy for sources in the frontal plane?
- Are there differences in localisation accuracy depending on the sound source used?

In the literature review it was also identified that localisation is not the only important attribute in spatial audio. It was highlighted that previous research had identified that envelopment is one of the attributes that separates multichannel systems from stereo systems and that it is one of the attributes driving multichannel development. It had been found that recent research had considered the effect of the number and positioning of height channels on the perception of envelopment and that an increase in elevated channels continued to provide improvement on the perceived envelopment compared to increasing the channels in the horizontal. The question was therefore asked:

- How does the number and positioning of elevated channels impact on the perception of envelopment?

Finally it was discussed that the attributes realised in previous spatial audio research had been arrived at using horizontal only surround systems. Therefore using these attributes to assess 3D surround systems may not reveal in what way 3D systems are adding to the spatial quality, as a result it was thought that a more exploratory method was needed to uncover the perceptual differences between 2D and 3D systems, an exploratory method exists which is called descriptive analysis, in short, a descriptive analysis allows a group of

listeners to provide descriptors elicited using a number of different audio stimulus, these descriptors can then be used to derive rating scales through group discussions. The main advantage of descriptive analysis is its less restrictive in the response from the listener. A small number of researchers had previously suggested attributes that may be concerned with 3D surround reproduction, however these were not realised through any descriptive analysis. Although it was highlighted that a small number of researchers have utilised descriptive analysis using 1st order ambisonics with height systems and have found contradictory results. Finally, utilising the attributes from previous research might not uncover the perceptual differences between with height systems and 2D surround systems and previous research using descriptive analysis has utilised only 1st order ambisonics, therefore the question was asked:

- Are there any unique attributes concerned with 3D surround systems?
- Does higher order ambisonics improve the listening experience?

10.1.1 Chapter 3

Chapter 3 starts with an account of the development of reproduction systems from mono to stereo and then onto 2D multichannel systems and the perceptual benefits gained from the inclusion of additional reproduction channels. The development of 3D systems are then detailed and the claimed perceptual benefits of the development of surround systems using height are also detailed, it is also illustrated that current 3D surround systems feature differing numbers of channels. The chapter then moves onto the rendering methods used to create 3D soundscenes and the benefits of each method. Ambisonics is then introduced as the method chosen to render 3D soundscenes due to its flexibility in terms of encoding and decoding in different resolutions and its ability to record in 3D. The mechanisms behind human hearing are also explored and the perceptual consequences of the inclusion of 3D surround are also detailed. It is reported that the addition of the surround channels allow the reproduction of rear sound effects and lateral energy which is important for the perception of envelopment. Envelopment has its origins in concert hall acoustics but is also valid in spatial audio reproduction. It is identified that envelopment is the consequence of lateral reflections and that in concert hall acoustics reflections from elevated positions do not enhance envelopment. With the development of surround systems to include height channels the impact on the localisation and envelopment is considered and is set out as the topic of further investigation in the thesis.

10.1.2 Chapter 4

In *Chapter 4* a localisation experiment is detailed which investigates whether there is a difference in localising an elevated virtual sound source presented in the median plane using a 3D loudspeaker array, in addition three different ambisonic orders are used 1, 2 and 3rd order. Based on the previous research it was concluded that a sound source which contained a large amount of high frequencies like pink noise would be appropriate for this experiment. From this localisation experiment it was concluded:

- There is a large amount of variability in the elevated virtual source positions
- There is no significant difference between ambisonic orders

10.1.3 Chapter 5

A further localisation test is carried out utilising virtual sources in the frontal plane, this was to investigate if there was any significant differences in localisation for virtual sources presented in the frontal plane, since there are more cues available to the listener. In addition two different sound sources were used pink noise and female speech and these were panned using three different ambisonic orders 1st, 2nd and 3rd order. The results of this experiment found that:

- 3rd order ambisonics had a significantly lower localisation error than that of 1st and 2nd
- 3rd order ambisonics did not match the localisation accuracy of real sound source

10.1.4 Chapter 6

It had been identified that envelopment is one of the attributes driving multichannel development and that in previous research it had been found that envelopment is a highly rated attribute. For these reasons an investigation was carried out into the impact of the height channels on the perception of envelopment. The investigation was initialised by utilising a computer simulation of a number of horizontal and with height speaker layouts. An objective measure was also used which is closely related to human perception called IACC. This allowed the measurement of binaural signals in order to quantify the degree of correlation between the ear signals which has been linked to spatial impression and envelopment.

- A large range of IACC values were found between 3D surround systems
- 3D surround systems provided a lower IACC than the current 5.1 standard

From the simulation it was discovered that there were a large range of IACC values across each of the systems, therefore a subjective test was executed using the systems previously simulated. The subjective test was designed to evaluate envelopment. From this experiment it was found:

- No significant difference between 2D and 3D surround systems for the perception of envelopment
- Most consistent scores for envelopment were for the 3D system with the least number of height channels
- Strong negative correlation between objective measure and simulated soundfield
- Strong negative correlation between objective measure and real soundfield

10.1.5 Chapter 7

The previous experiment utilised a decorrelated noise source investigating envelopment, this reduced the external validity, and therefore it was proposed that a second test should be carried out using real recorded sound scenes. Three different sound scenes that were captured using 3D ambisonic microphones were utilised. To reproduce these sound scenes two different ambisonic orders are used. Several different 2D systems were used and two different 3D systems, one of these using four height channels, since the previous experiment had shown this to have the most consistent rating. Again IACC was used to measure the soundfields of each system. The results of this experiment were:

- The addition of height reduces the value of IACC
- For two of the sound scenes there was a significant difference between the 3D systems and the 2D systems
- There was a strong negative correlation between the IACC and the mean subjective score

10.1.6 Chapter 8

In *Chapter 8* it was considered that in the previous chapters the subjective attribute envelopment had been used to assess 2D and 3D systems, it was considered that this restricts the amount of information captured and that there was a need for a more exploratory method. Therefore a descriptive analysis is initiated, this started with an elicitation process and then moves to the focus group discussions. A total of 3 separate focus group sessions were carried out in order to realise the attribute scales to be used in the subjective test. The outcome of this first stage was:

- Collection of 380 attributes
- Reduction of 380 descriptors to unique scales
- Five scales realised, *Spatial Distribution* (*Foreground*), *Spatial Distribution* (*Background*), *Realism*, *Size* and *Presence*

10.1.7 Chapter 9

In *Chapter 9* an experiment is detailed which utilises the five attribute scales realised from the elicitation. Various 2D and 3D surround systems were used and in addition to ambisonic rendered sound scenes in two different ambisonic orders 1st and 3rd, two VBAP rendered sound scenes were also introduced as further points of comparison. From the subjective evaluation it was found:

- *Foreground* attribute: no significant improvement over stereo with the addition of height
- *Background* attribute: no significant improvement over 2D surround with the addition of height
- *Presence* attribute: significantly enhanced with 3D (HOA) higher order ambisonics using applause sample compared 2D surround systems
- *Realism* attribute: 3D HOA system significantly adds to the perceived realism using applause sample compared to 5.1 system
- *Size* attribute: 3D HOA using applause sample and 3D VBAP using drama sample provides a better sense of the reproduced space compared to 2D surround systems

Objective measurements were also made of the various sound scenes using a KEMAR placed at the listening position. From this it was found:

- Introduction of height channels reduces the value of IACC
- Strong correlations between IACC and mean subjective score
- IACC was strongly correlated with all attribute scales, however subsequent removal of mono system reduced the statistical correlation between attribute scales

Independence between scales was also investigated, a PCA analysis was also carried out in order to investigate the underlying dimensions of the data.

- Combining the data for the surround 2D and 3D systems for the ambisonic and VBAP systems it was found that there were weak correlations between attribute scales

- Different correlations between scales was found for ambisonic surround systems combined versus VBAP surround systems combined.
- Separation of 2D and 3D ambisonic systems for comparison found similar correlations, different correlations between 2D and 3D VBAP surround systems
- PCA found 2 dimensional space and similar grouping of attributes for ambisonic 2D and 3D surround systems
- PCA found a 3 dimensional space and different grouping of attributes between 2D and 3D VBAP surround systems

From this it appears that the attribute scales were successful in finding perceptual differences between 2D and 3D surround systems. However it appears that from the PCA analysis there has been a slightly different attribute scale usage which could be due to the nature of the different sound scenes used in terms of the ambisonic and VBAP systems. This suggests that further subjective testing using the attribute scales is required to clarify this.

10.2 Summary

The question was asked in the introduction does the addition of height significantly enhance the listening experience?

From this research it has been found that the addition of height does significantly enhance the listening experience, although this is dependent on the sound scene used. It was found that for simulated diffuse soundfields and soundfields with random characteristics like applause there was no significant difference with the addition of height to improve the perception of envelopment. In addition it has been found that the use of higher order ambisonics (HOA) does not improve the perception of envelopment, but using (HOA) can reduce the localisation error for sources panned with an elevation in the frontal plane but not the median plane. It has also been found that the use of (HOA) to reproduce 3D sound scenes significantly enhances the perception of presence compared to 2D systems and that 3D HOA provides a significantly better sense of realism. Finally it was also found that the addition of height using HOA and VBAP rendered sound scenes provided the best sense of the size of the reproduced soundfield. This work provides only a small insight into the addition of height to the surround listening experience and therefore it has also been identified that there are a number of other areas that require further investigation.

10.3 Further Work

A number of areas are highlighted in this section of which could be extended in order to provide further information.

10.3.1 Localisation in the Median Plane

Further work may be carried out by developing the experimental method in order to provide further information by extending the stimuli to other types like speech or sound sources with significant low frequencies. In addition the ambisonic rendering method could also be optimised by using a dual band decoding method instead of only a one band method. The ambisonic order could also be increased so as to increase the frequency limit of correct reproduction.

10.3.2 Localisation in the Frontal Plane

Again this experimental work could be further developed by improving the test design and including additional stimuli which includes different high frequency content. The ambisonic decoding method could also be optimised by utilising dual band decoding and further optimisation taking into account the irregular speaker array. The ambisonic order could also be increased firstly to ensure the correct reproduction of high frequencies and secondly to investigate the order required to match the localisation accuracy of a real sound source.

10.3.3 Evaluation of Envelopment in With Height and Horizontal Surround Part 1

In the assessment of the perception of envelopment a simple computer simulation was used and an approximation of a diffuse soundfield was implemented. The computer simulation could be extended to use more sophisticated methods in order to investigate the perceptual effects of reflections from above a listener on subjective attributes. A further area of investigation could be different elevations of height speakers in order to investigate if different elevations have an effect on the perceived envelopment. Further to this listeners could be trained using the with height systems in order to make them more sensitive to the differences between 2D and 3D systems.

10.3.4 Evaluation of Envelopment in With Height and Horizontal Surround Part 2

The subjective testing using 2D and 3D surround systems utilised only three different types of sound scene. Further work could explore the envelopment in with height systems further, in particular using more diffuse or random sounfields. Since it was apparent from the two types of signal, applause sample and the noise sample used in the previous subjective test that there was no significant difference between the 2D and 3D systems and that this could possibly be due to the limitations of human hearing.

10.3.5 Perception of 3D Surround Systems

It was established that the results from the subjective test using the five attribute scales produced two different outcomes and that it was concluded that this could have been caused by the composition of the sound scenes. Further work could utilise the scales to test firstly a larger number of systems and also a broader range of stimuli to better establish the validity of the scales

11 Appendix 1 Matlab Code for the Creation of Decoding Matrices

```
%clear all
%clc
index_a =0;
for a= 5:5:360;
%a = input ('enter azimuth >')
%e = input ('enter elevation >')
a = deg2rad(a)
index_a= index_a+1

index_e =0;
for e= 5:5:360;

index_e= index_e+1
e = deg2rad(e)
%Bformat Encoder=================================================================%
b = [(1),cos(a)*cos(e),sin(a)*cos(e),sin(e),(3*(sin(e)^2)-
1)/2,sqrt((3)/2)*cos(a)*(sin(2*e))...

,sqrt((3)/2)*sin(a)*(sin(2*e)),sqrt((3)/2)*cos(2*a)*(cos(e)^2),sqrt((3)/2)*
sin(2*a)*(cos(e)^2),...
    sin(e)*(5*(sin(e)^2)-3)/2,sqrt((3)/8)*cos(a)*cos(e)*(5*(sin(e)^2)-
1),sqrt((3)/8)*sin(a)*cos(e)*(5*(sin(e)^2)-1),...

sqrt((15)/2)*cos(2*a)*sin(e)*cos(e)^2,sqrt((15)/2)*sin(2*a)*sin(e)*cos(e)^2
,sqrt(5/8)*cos(3*a)*(cos(e)^3),...
    sqrt((5)/8)*(sin(3*a))*(cos(e)^3)]

%Prealocate create Matrix of
zeros=============================================%
c = zeros (16,16)
%SpeakerAzimuths=================================================================
%
o1=[45 135 225 315 45 90 135 180 225 270 315 360 45 135 225 315];

%SpeakerElevations===============================================================
%o2=[-45 -45 -45 -45 0 0 0 0 0 0 0 0 45 45 45 45];

o1=deg2rad(o1);
o2=deg2rad(o2);

%CreateSpeakerMatrix=============================================================%
c(1,1:16)=(1);
c(2,:)=cos(o1).*cos(o2);
c(3,:)=sin(o1).*cos(o2);
c(4,:)=sin(o2);
c(5,:)= (3.*(sin(o2).^2)-1)/2;
c(6,:)= (sqrt((3)/2).*cos(o1)).*(sin(o2.*2));
c(7,:)= (sqrt((3)/2).* sin(o1)).*(sin(o2.*2));
c(8,:)= (sqrt((3)/2).* cos(o1*2)).*(cos(o2).^2);
c(9,:)= (sqrt((3)/2).*sin(o1*2)).*(cos(o2).^2);
c(10,:)= sin(o2).* (5.*(sin(o2).^2)-3)/2;
c(11,:)= sqrt((3)/8).* cos(o1).*cos(o2).*(5.*(sin(o2).^2)-1) ;
c(12,:)= sqrt((3)/8).* sin(o1).*cos(o2).*(5.*(sin(o2).^2)-1);
```

```
c(13,:)= sqrt((15)/2).* cos(2.*(o1)).*(sin(o2).*cos(o2).^2);
c(14,:)= sqrt((15)/2).* sin(2.*(o1)).*(sin(o2).*cos(o2).^2);
c(15,:)= sqrt((5)/8).* cos(3.*(o1)).*(cos(o2).^3);
c(16,:)= sqrt((5)/8).* sin(3.*(o1)).*(cos(o2).^3);

%createdecoderflavour=====================================================%
x = [1 0.861 0.861 0.861 0.612 0.612 0.612 0.612 0.612 0.305 0.305 0.305
0.305 0.305 0.305 0.305];

%creatediagonalmatrix======================================================%
r = diag(x);

%psuedoinversecmatrix======================================================%
d = pinv(c)

%multiplymatrixwithflavour=================================================%
j = d*r;

%factorinencodedsourceangle================================================%
s = j*b'

s_norm(:,index_a,index_e)=s;

[xx yy zz] = sph2cart(a,e,abs(s(5)));

xxx(index_a,index_e)=xx;
yyy(index_a,index_e)=yy;
zzz(index_a,index_e)=zz;

if s(5)>=0
    colour(index_a,index_e)=1;
elseif s(5)<0
    colour(index_a,index_e)=0;
end
end
end

surf(xxx,yyy,zzz,colour)


disp ('Decode cube_8_3R')


%from Malham "3D acoustic space and its simulation"
```

12 Appendix 2 Evaluation of Localisation in the Median Plane

Thank you for agreeing to take part

- Three different ambisonic orders are being investigated in terms of their virtual imaging accuracy,

- Please indicate your judgement of the sound source position by shouting out the position using the protocol explained:

Example, right 25 degrees would be: **RIGHT 25** left 25 degrees would be **LEFT 25** and the additional elevation in the positive 25 degrees would be **UP 25** or 25 degrees in the negative elevation would be **DOWN 25**

- The pole in front will allow you to judge the elevation

- You will be given a short practice session to allow you to become acquainted with the reporting method

- After the practise session the evaluation will begin

- The noise burst will be repeated 3 times

- Please face forward and do not move your head whilst sound is playing

- The test will last approximately 20 minutes

Thank you for taking part!

13 Appendix 3 Evaluation of Localisation In The Frontal Plane

Thank you for taking part

- You are asked to indicate the position of a sound source
- A microphone has been placed in the chamber for you to respond
- Before commencing the test proper you will be provided with a practice session where you will be played a noise burst from three different positions to help you become acquainted with the process
- The noise burst will be repeated 3 times
- In order to respond to the position of the noise source please respond to the horizontal position of the sound source by shouting left 30 for example for a source 30 degrees to the left or right 30 for a source 30 degrees to the right, for the elevation of a source please shout up 30 for a source with an elevation of 30 degrees or down 30 for a source 30 degrees below you.
- The markers placed in front of you can be used to guide your decision of elevation and the chart can help guide your decision as to the azimuth of the source
- Please feel free to ask questions about anything you do not understand

Thanks again

14 Appendix 4 Example Code for Multichannel Simulation

```
%currently set for the Auro 3D/
9.1=========================================
Fs= 44100;
%ReadInSignal=============================================================%
signal1=wavread ('white.wav');
signal2=wavread ('whitei.wav');
signal3=wavread ('whiteii.wav');
signal4=wavread ('whiteiii.wav');
signal5=wavread ('whiteiiii.wav');
signal6=wavread ('whiteiiiii.wav');
signal7=wavread ('whiteiiiiii.wav');
signal8=wavread ('whiteiiiiiii.wav');
signal9=wavread ('whiteiiiiiiii.wav');
%heightChannelsRead
InHRTF=====================================================%
lspL4030= wavread('R40e328a.wav');lspR4030= wavread('R40e032a.wav');
lspL40330= wavread ('R40e032a.wav');lspR40330= wavread('R40e328a.wav');
lspL40120= wavread ('R40e238a.wav');lspR40120= wavread ('R40e122a.wav');
lspL40240= wavread  ('R40e122a.wav');lspR40240= wavread ('R40e238a.wav');
%HorizontalChannelsReadInHRTF=============================================
=========%
lspL360= wavread ('R0e000a.wav'); lspR360= wavread ('R0e000a.wav');
lspL30=  wavread ('R0e330a.wav'); lspR30=  wavread ('R0e030a.wav');
lspL330= wavread ('R0e030a.wav'); lspR330= wavread ('R0e330a.wav');
lspL120= wavread ('R0e240a.wav'); lspR120= wavread ('R0e120a.wav');
lspL240= wavread ('R0e120a.wav'); lspR240= wavread ('R0e240a.wav');
%ConvolveChannelsHeightPlus===============================================
lspLeft4030= conv(signal1,lspL4030);lspRight4030= conv(signal1,lspR4030);
lspLeft40330= conv(signal2,lspL40330);lspRight40330=
conv(signal2,lspR40330);
lspLeft40120= conv (signal3,lspL40120);lspRight40120=
conv(signal3,lspR40120);
lspLeft40240= conv (signal4,lspL40240);lspRight40240=
conv(signal4,lspR40240);
%ConvolveChannelsHorizontal===============================================
lspLeft360= conv (signal5,lspL360);lspRight360= conv (signal5,lspR360);
lspLeft30= conv (signal6,lspL30);lspRight30= conv  (signal6,lspR30);
lspLeft330= conv (signal7,lspL330);lspRight330= conv (signal7,lspR330);
lspLeft120= conv(signal8,lspL120);lspRight120= conv (signal8,lspR120);
lspLeft240= conv (signal9,lspL240);lspRight240= conv (signal9,lspR240);
%Auro3DAdd Up All Speaker Contributions For
L/R=======================================================================
L=0.09*(lspLeft30+lspLeft120+lspLeft240+lspLeft330+lspLeft360+lspLeft4030+l
spLeft40120+lspLeft40240+lspLeft40330);

R=0.09*(lspRight30+lspRight120+lspRight240+lspRight330+lspRight360+lspRight
4030+lspRight40120+lspRight40240+lspRight40330);

wavwrite(L,Fs,'3DL')
wavwrite(R,Fs,'3DR')

corr=xcorr(L,R,50,'coeff')
tau=[-1.13:1.13/50:1.13];
[C,I]=max(corr);
iacc=max(corr)
itd=tau(I)
```

15 Appendix 5 Evaluation of Envelopment in With Height and Horizontal Surround Systems Part 1

Firstly thank you for agreeing to take part

- You are asked to judge the perceived envelopment of a number of different audio samples

- Before commencing the test you will be given a short training session where you will be played an example of an enveloping source and then a non-enveloping source.

- You will then be provided with an I Pad which will allow you to switch between seven different reproductions.

- There are two pages with seven different sliders and playback buttons you should indicate your score using the sliders and once you have scored all the reproductions you can then move to the next page

- A good strategy, it is suggested that you listen through to each of the examples before you start scoring the different versions

- Please feel free to ask questions about anything you do not understand

- The test should last approximately 20-30 minutes

16 Appendix 6 Evaluation of Envelopment in With Height and Horizontal Surround Systems Part 2

First of all thanks for taking part.

- Before commencing the test you will be given two aural examples to demonstrate ‘no envelopment’ and ‘envelopment’

- Then you are asked to rate the magnitude of envelopment of a number of audio samples which are:

1. Applause
2. Music
3. Atmos at a football match

- These will be looped samples of 30 seconds duration.

- For the music and atmos there will be six samples of each on two pages of the test interface.

- For the applause sample there will be eight of each on two pages of the test interface.

- Using the test interface you can switch between samples and compare whilst playback continues

- A good strategy is to listen through to each of the samples first to get a first impression of each one before grading.

- When you have finished grading each of the samples on page two of the test interface a ‘Save as’ dialogue box will appear. Please shout finished and I will load up the next bank of samples.

  Thanks again

17 Appendix 7 Exploring Spatial Perception In 3D Surround Systems Test Instructions

Firstly, thank you for taking part in this subjective evaluation

You are being asked to generate descriptive words for each of the presented sound scenes to help characterise their qualities. PLEASE FOCUS ON SPATIAL ATTRIBUTES AND THEN IF NECESSARY TIMBRAL ONES.

Please take some time to listen to each of the excerpts, then

1. Firstly consider each sound scene in isolation
2. After you are satisfied with your responses then consider terms which describe differences between samples

Please avoid attitudinal statements such as ‘better’ ‘worse’ ‘good’ ‘bad’ and use descriptive words for example ‘bright’ or ‘spacious’ or ‘confined’ etc... and please also avoid just describing the content of the sample e.g. the instrumentation…

Audio Samples

There are 3 different types of audio samples these vary in the number of different presentations:

- Sample of a small jazz ensemble has 5 different versions
- Choir sample has 7 different versions
- Applause has 7 different versions

Collecting Your Response

You will be provided with an IPad with which you can control and switch between presentations.

A text box is provided under each of the playback buttons for you to type in your responses (see picture).

When you have exhausted all the descriptors for the given sample please call out finished and the next set of samples will be loaded for you.

18 Appendix 8 Exploring the Perception of 3D Surround Systems

Thank you for agreeing to take part in this listening test, this test is designed to evaluate a number of different reproduction configurations using five different attributes.

To evaluate the systems four types of sound excerpts are used: Choir, Applause, Radio Drama and a live music concert

- The choir and Applause have seven different samples each
- The radio drama and live music concert have four different samples each

You are asked to rate each of the samples on five different attributes these are:

1. Foreground Distribution
2. Background Distribution
3. Presence
4. Realism
5. Size

The definitions for each of the attributes are as follows:

Foreground Distribution Definition

Foreground sources are described as being: mainly close and clearly perceived audio sources

Ends of scale: Biased-Uniform

Where the term biased meant that the sound field was comming from a particular direction, for example a mono source directly in front of the listener. Whilst uniform meant that the soundfield was equally spread all around the listener.

Background Distribution Definition

Background sources are described as being: room response, reverberant sounds, unclear, “foggy”

Ends of Scale: Biased-Uniform

Where the term biased meant that the sound field was comming from a particular direction, for example a mono source directly in front of the listener. Whilst uniform meant that the soundfield was equally spread all around the listener.

Presence

Question: Do you feel part of the sound-scene?

Ends of Scale: Detached-Immersed

I feel detached from the sound scene, I feel immersed in the sound scene

Realism

Question: How well does the reproduction match your expectations?

Ends of Scale: Unrealistic- Realistic

Size

Question: What's your impression of the physical size of the soundfield?

Ends of Scale: Small-large

User Interface

You will be provided with an IPad with sliders and playback buttons (see picture) and at the top of the interface will be displayed the attribute you are currently asked to grade on, each attribute will be presented on a separate page. You can freely switch and compare the different versions as much as you wish.

Once you have finished grading all the samples for a particular group a save as dialog box will appear please shout finished and the next set of samples will be loaded.

IMPORTANT: EACH OF THE SAMPLES CAN BE GRADED AT INTERMEDIATE POINTS ON THE SCALE NOT JUST AT EITHER END!

THANKS AGAIN!

19 References

AIRO, E., PEKKARINEN, J. & OLKINUORA, P. 1996. Listening to music with earphones: an assessment of noise exposure. *Acta Acustica united with Acustica,* 82**,** 885-894.

ANDO, Y. 1985. *Concert Hall Acoustics* Berlin, Springer-Verlag.

ANDO, Y. 1998. *Architectural acoustics: blending sound sources, sound fields, and listeners*, New York, Springer-Verlag

ASHBY, T., MASON, R. & BROOKES, T. 2011. Prediction of Perceived Elevation Using Multiple Pseudo-Binaural Microphones. *130th Audio Engineering Society Convention.*

ASHBY, T., MASON, R. & BROOKES, T. 2013. Head Movements in Three-Dimensional Localization. *134th Audio Engineering Society Convention.*

AUDESSY. 2014. *Audessy DSX* [Online]. Available: http://www.audyssey.com/technologies/dsx [Accessed 19/05/14 2014].

AURO3D. 2014. *Auro 3D Listening Formats* [Online]. Available: http://www.auro-3d.com/system/listening-formats/ [Accessed 15/6/2014 2014].

AVNI, A., AHRENS, J., GEIER, M., SPORS, S., WIERSTORF, H. & RAFAELY, B. 2013. Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution. *The Journal of the Acoustical Society of America,* 133**,** 2711.

BAMFORD, J. S. & VANDERKOOY, J. 1995. Ambisonic sound for us. *99th Audio Engineering Society Convention*

BANG, O. A. 1992. Music for Archemedes B&O.

BARBOUR, J. 2003a. Elevation Perception: Phantom Images in the Vertical Hemisphere. *Australisian Computer Music Association Conference*

BARBOUR, J. L. 2003b. Elevation perception: Phantom images in the vertical hemi-sphere. *24th Audio Engineering Society Conference.*

BARRON, M. 1971. The subjective effects of first reflections in concert halls—the need for lateral reflections. *Journal of Sound and Vibration,* 15**,** 475-494.

BARRON, M. 2009. *Auditorium acoustics and architectural design,* Oxon, Routledge.

BARTLETT, B. 1991. *Stereo microphone techniques,* Oxford, Focal Press.

BARTLETT, B. 2013. *Practical Recording Techniques: The step-by-step approach to professional audio recording,* Oxford, Taylor & Francis.

BATES, E. 2009. *The Compositon and Performance of Spatial Music* PhD PhD Thesis.

BAUME, C. & CHURNSIDE, A. 2010. SCALING NEW HEIGHTS IN BROADCASTING USING AMBISONICS. *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*

BECH, S. 1990. Listening tests on loudspeakers: a discussion of experimental procedures and evaluation of the response data. *8th Audio Engineering Society Conference.*

BECH, S. 1998. The influence of stereophonic width on the perceived quality of an audiovisual presentation using a multichannel sound system. *Journal of the Audio Engineering Society,* 46**,** 314-322.

BECH, S. 1999. Methods for subjective evaluation of spatial characteristics of sound. *16th Audio Engineering Society Conference.*

BECH, S., GEORGE, S., RUMSEY, F. & ZIELINSKI, S. 2008. Evaluating the Sensation of Envelopment Arising from 5-Channel Surround Sound Recordings. *124th Audio Engineering Society Convention*

BECH, S. & ZACHAROV, N. 2006. *Perceptual Audio Evaluation,* London, John Wiley&Sons Ltd.

BEGAULT, D. R. 1991. challenges to the successful implementation of 3D surround. *Journal of the Audio Engineering Society,* 39**,** 864-870.

BENJAMIN, E., LEE, R. & HELLER, A. 2006. Localization in horizontal-only ambisonic systems. *121st Audio Engineering Society Convention*

BERANEK, L. 2004. *Concert halls and opera houses: music, acoustics, and architecture,* New York, Springer.

BERANEK, L. L., AMERICA, A. S. O. & PHYSICS, A. I. O. 1996. *Concert and opera halls: how they sound,* Published for the Acoustical Society of America through the American Institute of Physics.

BERG, J. 2006. How do we determine the attribute scales and questions that we should ask of subjects when evaluating spatial audio quality. *Proc. Int. Workshop on Spatial Audio and Sensory Evaluation Techniques.*

BERG, J. 2009. The contrasting and conflicting definitions of envelopment. *126th Audio Engineering Society Convention*

BERG, J. & RUMSEY, F. 1999a. Spatial attribute identification and scaling by repertory grid technique and other methods. *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction.*

BERG, J. & RUMSEY, F. 1999b. Spatial attribute identification and scaling by repertory grid technique and other methods. *16th Audio Engineering Society Conference.*

BERG, J. & RUMSEY, F. 2000. Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction. *109th Audio Engineering Society Convention*

BERG, J. & RUMSEY, F. 2001. Verification and correlation of attributes used for describing the spatial quality of reproduced sound. *19th Audio Engineering Society Convention*

BERG, J. & RUMSEY, F. 2006. Identification of quality attributes of spatial audio by repertory grid technique. *Journal of the Audio Engineering Society,* 54**,** 365-379.

BERG, J. & RUMSEY, F. J. 2002. Validity of selected spatial attributes in the evaluation of 5-channel microphone techniques. *112th Audio Engineering Society Convention*

BERGE, S. & BARRETT, N. 2010. High Angular Resolution Planewave Expansion. *Proceedings of the 2010 Ambisonics Symposium.*

BERKHOUT, A. 1992. Wave-front synthesis: A new direction in electroacoustics. *The Journal of the Acoustical Society of America,* 92**,** 2396-2396.

BERTET, S., DANIEL, J. & MOREAU, S. 2006. 3d sound field recording with higher order ambisonics-objective measurements and validation of spherical microphone. *120th Audio Engineering Society Convention*

BERTET, S., DANIEL, J., PARIZET, E. & WARUSFEL, O. 2013. Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. *Acta Acustica united with Acustica,* 99**,** 642-657.

BLAUERT, J. 1996. *Spatial Hearing* London, MIT Press.

BLAUERT, J. & LINDEMANN, W. 1986. Auditory spaciousness: Some further psychoacoustic analyses. *The Journal of the Acoustical Society of America,* 80**,** 533.

BLOOM, P. J. 1977. Creating source elevation illusions by spectral manipulation. *J. Audio Eng. Soc,* 25**,** 560–565.

BLUMLEIN, A. 1931. *Improvements in and relating to sound transmission sound recording and sound reproducing systems*

BOEHM, J. 2011. Decoding for 3-D. *130th Audio Engineering Society Convention*

BRAUN, S. 2011. Localization of 3D Ambisonic Recordings and Ambisonic Virtual Sources. *In:* FRANK, M. (ed.). UniversitÄat fÄur Musik und darstellende Kunst Graz, Austria.

BREGMAN, A. S. 1990. *Auditory Scene Analysis* Cambridge, MA, MIT Press.

BS.2159-4, I.-R. 2012. Multichannel sound technology in home and broadcasting applications.

BUTLER, R. A. & HUMANSKI, R. A. 1992. Localization of sound in the vertical plane with and without high-frequency spectral cues. *Attention, Perception, & Psychophysics,* 51**,** 182-186.

CAMPO, E., BALLESTER, J., LANGLOIS, J., DACREMONT, C. & VALENTIN, D. 2010. Comparison of conventional descriptive analysis and a citation frequency-based descriptive method for odor profiling: An application to Burgundy Pinot noir wines. *Food quality and preference,* 21**,** 44-55.

CAPRA, A., FONTANA, S., ADRIAENSEN, F., FARINA, A. & GRENIER, Y. 2007. Listening tests of the localization performance of Stereodipole and Ambisonic systems. *123rd Audio Engineering Society Convention.*

CARLILE, S. 1996. *Virtual auditory space: Generation and applications,* Austin RG Landes.

CATTELL, R. B. 1966. The scree test for the number of factors. *Multivariate behavioral research,* 1**,** 245-276.

CENGARLE, G. 2012. *3D audio technologies: applications to sound capture, post-production and listener perception.* PhD PhD, Universitat Pompeu Fabra.

CHAFFEY, R. & SHIRLEY, B. G. 2007. Up-Mixing and Localisation-Localisation Performance of Up-Mixed Consumer Multichannel Formats. *122nd Audio Engineering Society Convention*

CHAPMAN, M. 2009. "Symmetries of Spherical Harmonics: Applications to Ambisonics. *Ambisonics Symposium, .* Graz.

CHAPMAN, M., RITSCH, W., MUSIL, T., ZMÖLNIG, I., POMBERGER, H., ZOTTER, F. & SONTACCHI, A. 2009. A STANDARD FOR INTERCHANGE OF AMBISONIC SIGNAL SETS Including a file standard with metadata. *1st Ambisonics Symposium*

CHERRY, E. C. 1953. Some further experiments upon the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America,* 26**,** 554-559.

CHOISEL, S. & WICKELMAIER, F. 2006a. Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound. *Journal of the Audio Engineering Society,* 54**,** 815-826.

CHOISEL, S. & WICKELMAIER, F. 2006b. Relating auditory attributes of multichannel sound to preference and to physical parameters. *120th Audio Engineering Society Convention.*

CHOISEL, S. & WICKELMAIER, F. 2007. Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *The Journal of the Acoustical Society of America,* 121**,** 388.

CHURNSIDE, A. 2011. Elbow in 3D Sound. Available: http://www.bbc.co.uk/rd/blog/2011/11/elbow-in-3d-sound [Accessed 8/12/14].

CHURNSIDE, A. 2012. Pinocchio. Available: http://www.bbc.co.uk/blogs/legacy/researchanddevelopment/2012/12/pinocchio.shtml [Accessed 8/12/14].

CHURNSIDE, A., PIKE, C. & LEONARD, M. Musical Movements- Gesture Based Audio Interfaces. 131st Audio Engineering Society Convention 2011.

CLAYPOOL, B., VAN BAELEN, W. & VAN DAELE, B. 2012. Auro 11.1 versus object-based sound in 3D. Barco.

CONETTA, R. 2011. *Towards the automatic assessment of spatial quality in the reproduced sound environment.* PhD, University of Surrey.

COX, T., DAVIES, W. & LAM, Y. 1993. The sensitivity of listeners to early sound field changes in auditoria. *Acta Acustica united with Acustica,* 79**,** 27-41.

CROCKETT, B. G., SEEFELDT, A. & SMITHERS, M. 2004. A New Objective Measure of Perceived Loudness. *117th Audio Engineering Society Convention*

DANCEY, C. & REIDY, J. 2004. *Stats Without Maths For Psychology Using SPSS For Windows,* London, Prentice Hall.

DANIEL, J. 2000. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia.* PhD Doctorial Thesis University of Paris

DANIEL, J. 2003. Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format. *23rd Audio Engineering Society Conference*

DANIEL, J. 2009. Evolving Veiws on Higher Order Ambisonics: from technological Veiws to Pragmatic Concerns *AMBISONICS SYMPOSIUM.* Graz.

DANIEL, J., MOREAU, S. & NICOL, R. 2003. Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging. *114th Audio Engineering Society Convention*

DANIEL, J., RAULT, J. B. & POLACK, J. D. 1998. Ambisonics encoding of other audio formats for multiple listening conditions. *105th Audio Engineering Society Convention*

DARWIN, C. 1991. The relationship between speech perception and the perception of other sounds. *Modularity and the motor theory of speech perception.*

DAVIS, M. F. 2003. History of spatial coding. *Journal of the Audio Engineering Society,* 51**,** 554-569.

DAVIS, R. & STEPHENS, S. 1974. The effect of intensity on the localization of different acoustical stimuli in the vertical plane. *Journal of Sound and Vibration,* 35**,** 223-229.

DOLBY. 2011. *Dolby History* [Online]. Available: http://www.dolby.com/us/en/about/history.html [Accessed 11/10/2013 2013].

DOLBY 2012. Dolby Atmos Next Generation Audio For Cinema. *In:* INC, D. L. (ed.). San Francisco.

DOLBY. 2014. *Authentic Cinema Sound By Dolby Atmos* [Online]. Available: http://www.dolby.com/us/en/consumer/technology/movie/dolby-atmos-details.html [Accessed 19/05/2014 2014].

EISLER, H. 1966. Measurement of Perceived Acoustic Quality of Sound Reproducing Systems by Means of Factor Analysis. *The Journal of the Acoustical Society of America,* 39**,** 484-492.

EVANS, M. J. 1998. Obtaining accurate responses in directional listening tests. *104th AES Convention*

FARRAR, K. 1979. Soundfield microphone. *Wireless World,* 85**,** 48-50.

FIELD, A. P. 2009. *Discovering statistics using SPSS*, Sage Publications Limited.

FURUYA, H., FUJIMOTO, K., YOUNG JI, C. & HIGA, N. 2001. Arrival direction of late sound and listener envelopment. *Applied Acoustics,* 62**,** 125-136.

GABRIELSSON, A. & LINDSTRÖM, B. 1985. Perceived sound quality of high-fidelity loudspeakers. *Journal of the Audio Engineering Society,* 33**,** 33-53.

GABRIELSSON, A. & SJÖGREN, H. 1979. Perceived sound quality of sound-reproducing systems. *The Journal of the Acoustical Society of America,* 65**,** 1019.

GARDENER, B. & MARTIN, K. 1997. *HRTF Measurement of a KEMAR Dummy-Head Microphone.* [Online]. Available: http://sound.media.mit.edu/ resources/ KEMAR. html, .

GEORGE, S., ZIELINSKI, S. & RUMSEY, F. 2006. Feature extraction for the prediction of multichannel spatial audio fidelity. *Audio, Speech, and Language Processing, IEEE Transactions on,* 14**,** 1994-2005.

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**,** 1013-1031.

GERZON, M. 1992. General Metatheory of Auditory Localisation. *92nd Audio Engineering Society.* .

GERZON, M. A. 1973. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society,* 21**,** 2-10.

GERZON, M. A. 1980. Practical periphony: The reproduction of full-sphere sound. *65th Audio Engineering Society Convention*

GERZON, M. A. 1983. Ambisonics in multichannel broadcasting and video. *74th Audio Engineering Society Convention*

GERZON, M. A. & BARTON, G. J. 1992. Ambisonic decoders for hdtv. *92nd Audio Engineering Convnetion*

GORZEL, M., KEARNEY, G. & BOLAND, F. 2014. Investigation of Ambisonic Rendering of Elevated Sound Sources. *55th Audio Engineering Society Conference.*

GRIESINGER, D. 1996. Spaciousness and envelopment in musical acoustics. *101st Audio Engineering Society Convention.*

GRIESINGER, D. 1997. The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces. *Acta Acustica united with Acustica,* 83**,** 721-731.

GUASTAVINO, C. & KATZ, B. F. G. 2004. Perceptual evaluation of multi-dimensional spatial audio reproduction. *The Journal of the Acoustical Society of America,* 116**,** 1105-1115.

GUASTAVINO, C., LARCHER, V., CATUSSEAU, G. & BOUSSARD, P. 2007. Spatial audio quality evaluation: comparing transaural, ambisonics and stereo. *Proceedings of the 13th International Conference on Auditory Display, Montréal, Canada.*

HAAS, H. 1972. The influence of a single echo on the audibility of speech. *Journal of the Audio Engineering Society,* 20**,** 146-159.

HAMASAKI, K., HATANO, W., HIYAMA, K., KOMIYAMA, S. & OKUBO, H. 2004a. 5.1 and 22.2 multichannel sound productions using an integrated surround sound panning system. *117th Audio Engineering Society Convention.*

HAMASAKI, K. & HIYAMA, K. 2003. Reproducing spatial impression with multichannel audio. *24th Audio Engineering Society Conference: International Conference: Multichannel Audio, The New Reality.*

HAMASAKI, K., HIYAMA, K., NISHIGUCHI, T. & OKUMURA, R. 2006. Effectiveness of height information for reproducing the presence and reality in multichannel audio system. *120th Audio Engineering Society Convention*

HAMASAKI, K., NISHIGUCHI, T., HIYAMA, K. & ONO, K. 2004b. Advanced multichannel audio systems with superior impression of presence and reality. *116th Audio Engineering Society Convention.*

HARDIN, R. H. & SLOANE, N. J. A. 2014. *"T Designs"* [Online]. Available: http://neilsloane.com/sphdesigns/ [Accessed 1/25/2014 2014].

HARTMANN, W. M. 1983. Localization of sound in rooms. *The Journal of the Acoustical Society of America,* 74**,** 1380-91.

HEBRANK, J. & WRIGHT, D. 1974a. Are two ears necessary for localization of sound sources on the median plane? *J. Acoust. Soc. Am,* 56**,** 935-938.

HEBRANK, J. & WRIGHT, D. 1974b. Spectral cues used in the localization of sound sources on the median plane. *The Journal of the Acoustical Society of America,* 56**,** 1829-1834.

HELLER, A., BENJAMIN, E. & LEE, R. 2010. Design of ambisonic decoders for irregular arrays of loudspeakers by non-linear optimization. *129th Audio Engineering Society Convention*

HELLER, A., LEE, R. & BENJAMIN, E. 2008. Is My Decoder Ambisonic? *125th Audio Engineering Society Convention*

HELLER, A. J., BENJAMIN, E. M. & LEE, R. 2012. A toolkit for the design of ambisonic decoders. *Linux Audio Conference.*

HIDAKA, T., BERANEK, L. L. & OKANO, T. 1995. Interaural cross-correlation, lateral fraction, and low-and high-frequency sound levels as measures of acoustical quality in concert halls. *The Journal of the Acoustical Society of America,* 98**,** 988.

HIRST, J. 2006. *Spatial Impression in Multichannel Surround Sound Systems.* PhD Doctorial Thesis University of Salford

HIRST, J. M., DAVIES, W. J. & PHILIPSON, P. J. 2006. Adaptation of concert hall measures of spatial impression to reproduced sound. *120th Audio Engineering Society Convention*

HIYAMA, K., KOMIYAMA, S. & HAMASAKI, K. 2002. The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field. *113th Audio Engineering Society Convention*

HOLLERWEGER, F. 2006. *Periphonic sound spatialization in multi-user virtual environments.* Masters Master's Thesis, Austrian Institute of Electronic Music and Acoustics (IEM).

HOLMAN, T. 2008. *Surround sound: up and running*, Taylor & Francis.

HORSBURGH, A. J., DAVIS, R. E., MOFFAT, M. & CLARK, D. F. 2012. Subjective Assessments of Higher Order Ambisonic Sound Systems in Varying Acoustical Conditions. *133rd Audio Engineering Society Convention*

HOWARD, D. & ANGUS, J. 2006. *Acoustics and Psychoacoustics,* Oxford, Focal Press.

IHLE, M. 2003. Differential microphone arrays for spectral subtraction. *Eighth International Workshop on Acoustic Echo and Noise Control.* Citeseer.

ISO 2000. Acoustics:Measurement of the reverberation time of rooms with reference to other acoustical parameters. *ISO 3382:2000.*

ITU 2001. BS. 1534-1. Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA). *International Telecommunications Union, Geneva*.

ITU 2011. Multichannel sound technology in home and broadcasting applications.

ITU 2012. Document 6C/85-E, Sensation of listener's envelopment and localization of phantom sound images on loudspeaker configurations with height channels.

JOLLIFFE, I. 2005. *Principal component analysis*, New York, Springer

KAISER, F. 2011. A hybrid approach for three-dimensional sound spatialization. IEM.

KAMEKAWA, T. & MARUI, A. 2010. Developing Common Attributes to Evaluate Spatial Impression of Surround Sound Recording. *40th Audio Engineering Society Conference*

KEILER, F. & BATKE, J. M. 2010. Evaluation of Virtual Source Localization Using 3-D Loudspeaker Setups. *128th Audio Engineering Society Convention*

KENDALL, G. S. 2010. Spatial perception and cognition in multichannel audio for electroacoustic music. *Organised Sound,* 15**,** 228-238.

KIM, S., KO, D., NAGENDRA, A. & WOSZCZYK, W. 2013. Subjective Evaluation of Multichannel Sound with Surround-Height Channels. *135th Audio Engineering Society Convention*

KIM, S., LEE, Y. W. & PULKKI, V. 2010. New 10.2-Channel Vertical Surround System (10.2-VSS); Comparison Study of Perceived Audio Quality in Various Multichannel Sound Systems with Height Loudspeakers. *129th Audio Engineering Society Convention.*

KIM, S. & MARTENS, W. L. 2007. Verbal elicitation and scale construction for evaluating perceptual differences between four multichannel microphone techniques. *122nd Audio Engineering Society Convention*

KO, S., OH, E., PARK, S. H. & SHIM, H. 2010. Perceptual Evaluation of Spatial Audio Quality. *129th Audio Engineering Society Convention.*

LETOWSKI, T. 1989. Sound quality assessment: cardinal concepts. *87th Audio Engineering Society Convention.*

LIEBETRAU, J., SPORER, T., KORN, T., KUNZE, K., MANK, C., MARQUARD, D., MATHEJA, T., MAUER, S., MAYENFELS, T. & MÖLLER, R. 2007. Localization in spatial audio-from wave field synthesis to 22.2. *123rd Audio Engineering Society Convention*

LOKKI, T. & PÄTYNEN, J. 2011. Lateral reflections are favorable in concert halls due to binaural loudness. *The Journal of the Acoustical Society of America,* 130**,** EL345-EL351.

LOKKI, T., PÄTYNEN, J., KUUSINEN, A., VERTANEN, H. & TERVO, S. 2011. Concert hall acoustics assessment with individually elicited attributes. *The Journal of the Acoustical Society of America,* 130**,** 835.

LONG, M. 2005. *Architectural acoustics,* London, Elsevier.

MALHAM, D. 2003. Higher order Ambisonic systems. *abstracted from'Space in Music-Music in Space', an Mphil thesis by Dave Malham, submitted to the University of York in April*.

MALHAM, D. G. 1992. Experience with a Large Area 3D Ambisonic Sound Systems. *PROCEEDINGS-INSTITUTE OF ACOUSTICS,* 14**,** 209-209.

MANTEL, J. 1975. Optimum Multiphony. *50th Audio Engineering Society Convention*

MARSHALL, A. H. 1967. A note on the importance of room cross-section in concert halls. *Journal of Sound and Vibration,* 5**,** 100-112.

MARTIN, G. 2002. Interchannel interference at the listening position in a five-channel loudspeaker configuration. *113th Audio Engineering Society Convention*

MARTIN, G. 2005. A New Microphone Technique for Five-Channel Recording. *118th Audio Engineering Society Convention.*

MARTIN, G., WOSZCZYK, W., COREY, J. & QUESNEL, R. 1999a. Controlling phantom image focus in a multichannel reproduction system. *PREPRINTS-AUDIO ENGINEERING SOCIETY.*

MARTIN, G., WOSZCZYK, W., COREY, J. & QUESNEL, R. 1999b. Sound source localization in a five-channel surround sound reproduction system. *107th Audio Engineering Society Convention*

MASON, R. 2002. *Elicitation and measurement of auditory spatial attributes in reproduced sound.* PhD Doctoral, University of Surrey.

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**,** 1337-1350.

MASON, R. & RUMSEY, F. J. 2002. A comparison of objective measurements for predicting selected subjective spatial attributes. *112th Audio Engineering Society Convention*

MEYER, J. & ELKO, G. W. 2008. Handling spatial aliasing in spherical array applications. *Hands-Free Speech Communication and Microphone Arrays.* IEEE.

MOORE, D. J. 2009. *The development of a design tool for 5-speaker surround sound decoders.* PhD PhD, University of Huddersfield.

MORIMOTO & M. MAEKAWA, Z. 1989. Auditory spaciousness and envelopment. *Proc 13th ICA* Belgrade.

MORIMOTO, M. 1997. The role of rear loudspeakers in spatial impression. *103rd Audio Engineering Society Convention*

MORIMOTO, M. 2002. The relation between spatial impression and the precedence effect. *Proc. International Conf. on Auditory Display.*

MORRELL, M. J. & REISS, J. D. 2009. A Comparative Approach to Sound Localisation within a 3D Sound Field. *126th Audio Engineering Society Convention*

MURRAY, J., DELAHUNTY, C. & BAXTER, I. 2001. Descriptive sensory analysis: past, present and future. *Food research international,* 34**,** 461-471.

NACHBAR, C., ZOTTER, F., DELEFLIE, E. & SONTACCHI, A. 2011. ambiX-A Suggested Ambisonics Format. *3rd Ambisonics Symposium.*

NAKAYAMA, T., MIURA, T., KOSAKA, O., OKAMOTO, M. & SHIGA, T. 1971. Subjective assessment of multichannel reproduction. *Journal of the Audio Engineering Society,* 19**,** 744-751.

NEHER, T. 2004. *Towards a spatial ear trainer.* PhD, University of Surrey.

NEUKOM, M. 2007. Ambisonic panning. *123rd Audio Engineering Society Convnetion*

NISBETT, A. 1970. Happy Birthday Ludwig. *Studio Sound* IPC Media Limited

OODE, S., SAWAYA, I., ANDO, A., HAMASAKI, K. & OZAWA, K. 2011. Vertical Loudspeaker Arrangement for Reproducing Spatially Uniform Sound. *131st Audio Engineering Society Convention.*

PAQUIER, M., KOEHL, V., NICOL, R. & DANIEL, J. 2011. Subjective assessment of microphone arrays for spatial audio recording. *Proceedings of Forum Acusticum 2011.*

PERNAUX, J. M., EMERIT, M. & NICOL, R. 2003. Perceptual Evaluation of Binaural Sound Synthesis: the Problem of Reporting Localization Judgments. *114th AES Convention.*

PHILIPSON, P., HIRST, J. & WOOLLARD, S. 2002. Investigation into a Method for Predicting the Perceived Azimuth Position of a Virtual Image. *112th Audio Engineering Society Convention*

PIELEANU, I. 2004. *LOCALIZATION PERFORMANCE WITH LOW-ORDER AMBISONICS AURALIZATION.* Masters Master of Science Rensselaer Polytechnic Institute.

POHLMANN, K. C. 2005. *Principles of digital audio,* New York, McGraw-Hill New York.

POPPER, A. N. & FAY, R. R. 2005. *Sound source localization,* USA, Springer.

PULKKI, V. 2001. Localization of amplitude-panned virtual sources II: Two-and three-dimensional panning. *Journal of the Audio Engineering Society,* 49**,** 753-767.

RATLIFF, P. 1974. *Properties of hearing related to quadraphonic reproduction,* Research Department, Engineering Division, BBC.

RAYLEIGH, L. 1875. On our perception of the direction of a source of sound. *Proceedings of the Musical Association***,** 75-84.

ROBINSON, C. Q., MEHTA, S. & TSINGOS, N. 2012. Scalable Format and Tools to Extend the Possibilities of Cinema Audio. *SMPTE Motion Imaging Journal,* 121**,** 63-69.

ROFFLER, S. K. & BUTLER, R. A. 1968. Factors that influence the localization of sound in the vertical plane. *The Journal of the Acoustical Society of America,* 43**,** 1255.

ROHR, L., CORTEEL, E., NGUYEN, K.-V. & LISSEK, H. 2013. Vertical Localization Performance in a Practical 3-D WFS Formulation. *Journal of the Audio Engineering Society,* 61**,** 1001-1014.

RUMSEY, F. 1998a. Subjective assessment of the spatial attributes of reproduced sound. *15th Audio Engineering Society Conference.*

RUMSEY, F. 1998b. Subjective assessment of the spatial attributes of reproduced sound. *Proceedings of the Audio Engineering Society 15th International Conference: Audio, Acoustics and Small Space.*

RUMSEY, F. 2002. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society,* 50**,** 651-666.

RUMSEY, F. 2005. *Spatial Audio,* Oxford, Focal Press.

RUMSEY, F. 2006. Spatial audio and sensory evaluation techniques-context, history and aims. *IOSR Workshop* University of Surrey Institute of Sound Recording

RUMSEY, F., ZIELINSKI, S., JACKSON, P., DEWHIRST, M., CONETTA, R., GEORGE, S., BECH, S. & MEARES, D. 2008. QESTRAL (Part 1): Quality evaluation of spatial transmission and reproduction using an artificial listener. *125th Audio Engineering Society Convention.*

SANTALA, O. & PULKKI, V. 2009. Resolution of Spatial Distribution Perception with Distributed Sound Source in Anechoic Conditions. *126th Audio Engineering Society Convention.*

SAWAYA, I., IRIE, K., SUGIMOTO, T., ANDO, A. & HAMASAKI, K. 2013. Discrimination of Changing Loudspeaker Positions of 22.2 Multichannel Sound System Based on Spatial Impressions. *134th Audio Engineering Society Convention*

SAWAYA, I., OODE, S., ANDO, A. & HAMASAKI, K. 2011. Size and Shape of Listening Area Reproduced by Three-dimensional Multichannel Sound System with Various Numbers of Loudspeakers. *131st Audio Engineering Society Convention.*

SAZDOV, R., PAINE, G. & STEVENS, K. 2007. Perceptual Investigation into envelopment, spatial clarity, and engulfment in reproduced multi-channel audio. *31st Audio Engineering Society Conference*

SCHOEFFLER, M., WESTPHAL, S., ALEXANDER, A., BAYERLEIN, H. & HERRE, J. 2014. Comparison of a 2D-and 3D-Based Graphical User Interface for Localization Listening Tests. *Proc. of the EAA Joint Symposium on Auralization and Ambisonics.*

SCHROEDER, M. R. 1979. Binaural dissimilarity and optimum ceilings for concert halls: More lateral sound diffusion. *The Journal of the Acoustical Society of America,* 65**,** 958.

SHIRLEY, B., OLDFIELD, R., MELCHIOR, F. & BATKE, J. M. 2013. Platform Independent Audio. *Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media***,** 130-165.

SILZLE, A., GEORGE, S., HABETS, E. & BACHMANN, T. 2011. Investigation on the Quality of 3D Sound Reproduction. *Proceedings of ICSA.*

SILZLE, A., STRAUSS, H. & NOVO, P. 2004. IKA-SIM: A system to generate auditory virtual environments. *116th Audio Engineering Society Convention*

SOULODRE, G. A., LAVOIE, M. C. & NORCROSS, S. G. 2002. Investigation of listener envelopment in multichannel surround systems. *113th Audio Engineering Society Convention.*

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**,** 826-840.

SOUTHERN, A. & MURPHY, D. 2007. 2nd Order Spherical Harmonic Spatial Encoding of Digital Waveguide Mesh Room Acoustic Models. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*

STANDARDS, B. 2000. Acoustics:Measurement of the reverberation time of rooms with reference to other acoustical parameters. *BS EN ISO 3382:2000.*

STEINBERG, J. & SNOW, W. B. 1934. Auditory perspective-physical factors. *Transactions of the American Institute of Electrical Engineers,* 53**,** 12-17.

STEINKE, G. 1996. Surround Sound the New Phase, An Overview *100th Audio Engineering Society Convention*

STREICHER, R. & EVEREST, A. 1998. *The New Stereo Soundbook,* USA.

STRYBEL, T. Z. & FUJIMOTO, K. 2000. Minimum audible angles in the horizontal and vertical planes: Effects of stimulus onset asynchrony and burst duration. *The Journal of the Acoustical Society of America,* 108**,** 3092-3095.

STUART, J. R. 1996. The psychoacoustics of multichannel audio. *11th Audio Engineering Society Conference: Conference: Audio for New Media (ANM).*

TALBOT-SMITH, M. 2001. *Audio engineer's reference book,* Oxford, CRC Press.

THEILE, G. 1993. The new sound format: 3/2-stereo. *94th Audio Engineering Society Convention*

THEILE, G. & PLENGE, G. 1977. Localization of lateral phantom sources. *Journal of the Audio Engineering Society,* 25**,** 196-200.

THOMAS, G., SCHREER, O., SHIRLEY, B. & SPILLE, J. Combining panoramic image and 3D audio capture with conventional coverage for immersive and interactive content production. International Broadcasting Convention, Amsterdam, 2011. 8-13.

THURLOW, W. R. & RUNGE, P. S. 1967. Effect of induced head movements on localization of direction of sounds. *Journal of the Acoustical Society of America,* 42**,** 480-488.

TOHYAMA, M. & SUZUKI, A. 1989. Interaural cross-correlation coefficients in stereo-reproduced sound fields. *The Journal of the Acoustical Society of America,* 85**,** 780.

TOMLINSON, H. 2000. *5.1 Surround Sound Up and Running,* USA, Butterworth Heinemann.

TORICK, E. 1998. Highlights in the history of multichannel sound. *Journal of the Audio Engineering Society,* 46**,** 27-31.

TRAVIS, C. 2009. A new mixed-order scheme for ambisonic signals. *AMBISONICS SYMPOSIUM.*

TREVIÑO, J., OKAMOTO, T., IWAYA, Y. & SUZUKI, Y. 2011. Evaluation of a new Ambisonic decoder for irregular loudspeaker arrays using interaural cues. *Ambisonics Symposium.*

TSINGOS, N., CHABANNE, C., ROBINSON, C. & MCCALLUS, M. 2011. Surround Sound with Height in Games Using Dolby Pro Logic IIz. *41st Audio Engineering Society International Conference.*

WAKUDA, A., FURYA, H., FUJIMOTO, K., ISOGAI, K. & ANAI, K. 2003. Effects of arrival direction of late sound on listener envelopment. *Acoustic Science and Tech,* 24.

WALLACH, H. 1939. On sound localization. *The Journal of the Acoustical Society of America,* 10**,** 270-274.

WALLACH, H., NEWMAN, E. B. & ROSENZWEIG, M. R. 1949. A Precedence Effect in Sound Localization. *The Journal of the Acoustical Society of America,* 21**,** 468-468.

WANKLING, M., FAZENDA, B. & DAVIES, W. J. 2012. The assessment of low-frequency room acoustic parameters using descriptive analysis. *Journal of the Audio Engineering Society,* 60**,** 325-337.

WARREN, R. M. 2008. *Auditory Perception an Analysis and Synthesis,* Cambridge, Cambridge University Press.

WATKINSON, J. 1998. *The art of sound reproduction,* Oxford, Taylor & Francis.

WELLS, J. 2010. Modification of Spatial Information in Coincident-Pair Recordings. *128th Audio Engineering Society Convention*

WENZEL, E. M., ARRUDA, M., KISTLER, D. J. & WIGHTMAN, F. L. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America,* 94**,** 111-123.

WIGGINS, B. 2004. *An investigation into the real-time manipulation and control of three-dimensional sound fields.* PhD Doctoral Thesis, University of Derby.

WIGHTMAN, F. L. & KISTLER, D. J. 1989. Headphone simulation of free-field listening: II. Psychophysical validation. *Journal of the Acoustical Society of America,* 85**,** 868-878.

WIGHTMAN, F. L. & KISTLER, D. J. 1992. The dominant role of low-frequency interaural time differences in sound localization. *Journal of the Acoustical Society of America; Journal of the Acoustical Society of America,* 91**,** 1648-1661.

WILLIS, B. 1999. Holman Conducts First Public Demo of '10.2' Surround Sound. *Stereophile* [Online]. Available: http://www.stereophile.com/news/10489/ [Accessed Nov 2012].

YOST, W. A. 1994. *Fundamentals of hearing: An introduction,* London Academic Press.

ZACHAROV, N. 2000. *Perceptual studies on spatial sound reproduction systems.* PhD PhD, Helsinki University of Technology.

ZACHAROV, N. & KOIVUNIEMI, K. 2001. Unravelling the perception of spatial sound reproduction: Analysis & external preference mapping. *111th Audio Engineering Society International Convention.*

ZIELINSKI, S., BROOKS, P. & RUMSEY, F. On the use of graphic scales in modern listening tests. 123rd Audio Engineering Society Convention 2007.

ZIELINSKI, S. K., RUMSEY, F. & BECH, S. 2003. Effects of down-mix algorithms on quality of surround sound. *Journal of the Audio Engineering Society,* 51**,** 780-798.

ZIELINSKI, S. K., RUMSEY, F., KASSIER, R. & BECH, S. 2005. Comparison of basic audio quality and timbral and spatial fidelity changes caused by limitation of bandwidth and by down-mix algorithms in 5.1 surround audio systems. *Journal of the Audio Engineering Society,* 53**,** 174-192.

ZOTTER, F. & FRANK, M. 2012. All-Round Ambisonic Panning and Decoding. *Journal of the Audio Engineering Society,* 60**,** 807-820.

ZWICKER, E. & FASTL, H. 1999. Psychoacoustics, Facts and Models. Springer, Berlin, 2nd updated edition.