Soundfield Representation, Reconstruction and Perception

Glenn Dickins

B.Sc. (ANU) B.Eng.(Hons) (ANU)

March 2003

A thesis submitted for the degree of Master of Engineering
of The AUstralian National University

Department of Telecommunications Engineering
Research School of Information Sciences and Engineering

This thesis has been prepared and submitted as an original research work. I declare that all sources used in the development and discussion of the research topics have been appropriately acknowledged. The work has not been submitted towards any other degree or qualification.

Glenn Dickins

March 19, 2003.

Abstract

This thesis covers the area of soundfield representation, reconstruction and perception. The complexity and information content of a soundfield presents many mathematical and engineering challenges for accurate reconstruction. After an in-depth review of the field of mathematical soundfield representation, an analysis of the numerical and practical constraints for soundfield reconstruction is presented. A review of work in experimental psycho-acoustics higlights the variability of spatial sound perception. It is shown that the error and uncertainty in perception is of a comparable magnitude to the accuracy achievable by present soundfield systems. Therefore, the effects of hearing adaption, sensory bias, sensory conflict, and contextual memory cannot be ignored. If the listening environment is inappropriate or in conflict with the desired perceptual experience, little is gained from more complex soundfield representation or reconstruction. The implications of this result to the delivery of spatial audio is discussed and some open problems for further exploration and experimentation are detailed.

Acknowledgements

I would like to thank those who have supported and encouraged me to submit this thesis. In particular I would like to thank my wife, Karen, whose understanding and support has been the greatest gift for this and all other endeavours we have embarked on together. I would also like to acknowledge the interest and faith of my friends and family in my chosen area of research.

I would like to give paritucular acknowledgement to Lake Technology. While working at Lake I was able to visit many of the world's leading research and development facilities in the area of spatial sound and virtual reality. To work towards the development and commercialization of Dolby Headphone was the fulfillment of a vocational dream and a great experience. I would like to thank and acknowledge all those who shared in this experience and also contributed to the discussions and ideas of this thesis. In particular I would like to give credit to David McGrath whose ideas and passion for audio was a source of inspiration and motivation.

I would also like to thank my supervisor, Professor Rodney Kennedy with whom I have enjoyed many discussions and conversations relating to this thesis and other areas. I give further thanks to Rod for providing an environment in which I could pursue audio signal processing and spatial sound at the Australian National University. Through this opportunity I have been able to work on problems and ideas that have been rewarding both professionally and personally.

Contents

List of Figures

List of Tables

List of Symbols

Symbol	**Definition**
$\mathbf{A} \quad A_n$	Ambisonic excitation signal vector and individual term.
$\mathbf{B} \quad B_n^m$	Modal excitation signal vector and individual term.
c	Speed of sound (approximately 320 ms^{-1})
$\mathbf{E}\left[\cdot\right]$	Expectation value – statistical mean.
f	Frequency (Hz)
f_n	Ambisonic directional polynomial $\sum_{l,m,n} \gamma_{l,m,n} x^l y^m z^n$
$\mathbb{G}$	Speaker gain matrix.
g_i	Gain parameter for speaker or point source.
$h_n^{(1)}\left(.\right)$	Hankel function of the first kind, order n.
i	Imaginary number – $\sqrt{-1}$
$j_n\left(.\right)$	Spherical Bessel Function $j_n\left(x\right) = J_{n+\frac{1}{2}}\left(x\right)/\sqrt{x}$
$J_n\left(.\right)$	Ordinary Bessel Function of the first kind.
k	The wave number. Defined as $k = 2\pi f/c = \omega/c = 2\pi/\lambda$
l, m, n, p, q, s	Enumeration variables for summations and expansions.
$\mathbf{N}\left(\mu, \sigma\right)$	Normal distribution with mean μ and standard deviation σ.
N	Number of terms or order of expansion. Number of speakers.
$P \; P\left(x, y, z, t\right)$	Absolute pressure at a point
$p \; p\left(x, y, z, t\right)$	Relative pressure at a point around average pressure P_{avg}
$p_{l,m,n}$	Taylor series coefficient of soundfield $\frac{\partial^l}{\partial x^l}\frac{\partial^m}{\partial y^m}\frac{\partial^n}{\partial z^n}p$
$P_n^m\left(.\right)$	Associated Legendre Polynomials
R	Radius of sphere for region of reconstruction or integration.
r, θ, ϕ	Components of a vector in $\mathbb{R}^3$ in spherical polar coordinates.
	$x = r\sin\theta\cos\phi \qquad y = r\sin\theta\sin\phi \qquad z = r\cos\theta$
S	Surface of an enclosed region
$\mathbb{S}$	Speaker contribution matrix.

Symbol	**Definition**
t	Time
$\mathbf{U}$ $\quad U$	Speaker excitation vector and individual speaker excitation.
V	Volume of region.
$\mathbf{x}, \mathbf{y}, \mathbf{d}, \mathbf{r}$	Position vectors in $\mathbb{R}^3$
$\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \widehat{\mathbf{d}}, \widehat{\mathbf{r}}$	Unit vectors in $\mathbb{R}^3$. $\widehat{\mathbf{x}} = \mathbf{x} / \left\Vert \mathbf{x} \right\Vert$ etc.
x, y, z	Components of a vector in $\mathbb{R}^3$ in Cartesian coordinates.
α_n^m	Modal expansion coefficient.
$\gamma_{l,m,n}$	Ambisonic polynomial coefficient for $x^l y^m z^n$
λ	Wavelength (m)
Ω	Range of solid angle
ω	Angular frequency ($\omega = 2\pi f$)
$\boldsymbol{\delta}\left(a, b\right)$	Kronecker-Delta function. $\boldsymbol{\delta}\left(a, b\right) = 1$ iff $a = b$
$\left\vert \cdot \right\vert$	Absolute value or complex magnitude
$\left\Vert \cdot \right\Vert$	Euclidian distance or L_2 norm
$\left\lceil \cdot \right\rceil$	Integer ceiling. $\left\lceil x \right\rceil$ =smallest integer $\geq x$
$\frac{\partial^l}{\partial x^l}$	Partial derivative of order l with respect to variable x.
$\bigtriangledown^2$	Laplacian operator

List of Papers

The following papers and documents have been authored and submitted in conjunction with the work on this thesis.

- G. Dickins and R. Kennedy, "Towards optimal sound field reconstruction," *Presented at the106th Convention of the Audio Engineering Society*, Munich, May 1999. Audio Engineering Society Preprint 4925.
- G. Dickins, M. Flax, A. McKeag, and D. McGrath, "Optimal 3d speaker panning," *Proceedings of the AES 16th International Conference on Spatial Sound*, Finland, April 1999. Audio Engineering Society, pp. 421–426.
- G. Dickins, P. Flanagan, and L. Layton, "Real-time virtual acoustics for 5.1," *Proceedings of the AES 16th International Conference on Spatial Sound, Ravoniemi*, Finland, April 1999. Audio Engineering Society, pp. 136–140.
- G. Dickins, "What is surround sound?" *Acoustics Australia*, vol. 26, no. 3, December 1998.
- G. Dickins, P. Flanagan, and L. Layton, "The SP1 – acoustic simulation for post production," *Presented at the 106th Convention of the Audio Engineering Society*, Munich, May 1999.
- G. Dickins, "Automated time alignment and equalization of a speaker array for soundfield reproduction," *Presented at the 6th Australian Regional Convention of the Audio Engineering Society*, September 1996. Audio Engineering Society Preprint 4317.
- G. Dickins, D. McGrath, A. McKeag, A. Reilly, L. Layton, B. Conolly, R. Cartright, R. Buttler, and S. Bartlett et al "Binaural simulation studies and experiments," *Technical Reports*, Lake Technology, Sydney, 1996-2000. Internal Discussion Papers – Unpublished.

Chapter 1

Introduction

1.1 Overview

Sound plays an important part in our awareness of the environment. For those of able hearing, sound provides a constant perception of things around us. Hearing, perhaps more than any other sense, provides us with a sense of connection and presence [1,2]. Since Edison first recorded on a foil in 1877, many engineers have worked to improve the fidelity and realism of audio recordings. The transporting of sound through wires dates back to 1877 with Bell and the telephone and via radio waves with Marconi in 1895. Now devices that record, play and transport sound are so common place that the ability to record and transport an audio signal is now taken for granted.

The fidelity of modern recording equipment is extremely high – and arguably approaches the limits of perception of the human auditory system. However, recordings are rarely mistaken for real auditory events. Most efforts in audio recording and reproduction have concentrated on the temporal fidelity of the wave field – frequency range, dynamic range and distortion [3]. To truly create an immersive sound field it is necessary to match the complexity of the spatial excitation and nuances present in an real sound field. The ability to synthesize, capture and recreate an immersive sound field is a significant challenge.

Within the areas of telepresence[1] and entertainment, the ability to create a compelling immersive sound field has many applications. In an age of information, hearing is currently an under-utilized sense. Application areas include simulation [8], collaborative environments [9], virtual reality [2], medicine [10], education, data visualization, information dis-

[1] Telepresence : the perception of presence within a physically remote or simulated site [4, 5]. For a discussion of this area see the Telepresence issue of the BT Technology Journal [6] in particular [7].

plays[2], augmented reality [11] and many entertainment applications [12]. There is also an immediate market for improved spatial sound with multi-channel audio being the fastest growing segment of the consumer audio market [13].

The thesis begins by looking at the mathematical foundations of spatial audio with a goal of determining the theoretical and practical limits of accurate sound field representation and recreation. It is established that the information content in an accurate sound-field is high [14], thus the cost of aiming for exact representation and reconstruction is prohibitive. By investigating the complexity of the mathematical task it is possible to determine fundamental limitations. By comparing different techniques and representations for sound fields, it can be shown that there is a strong equivalence between representations. Thus fundamental limits apply to all systems for capturing and reconstructing spatial sound fields.

Headphones can be considered as an alternative to free field reconstruction for the delivery of spatial sound and are depicted by some as the optimal approach [15]. The physical link to the listener reduces the need to reconstruct the sound over a volume of space. Headphones can be considered an effective way of delivering a sound field at two points. Theoretically, headphones should be able to accurately control the excitation of the eardrum, with the correct excitation the correct sound field may be perceived.

The performance of headphone sound presentation is seen to be variable [16][3]. In some cases performance similar to free field localization is observed [17] whilst in others constant and significant errors have been observed [18]. Although there are differing opinions, the problem with headphone audio presentation is generally accepted to arise from numerical deficiencies in the measurement and simulation of responses. However, in some situations, a poor numerical approximation can create a completely convincing effect. It is perhaps the design of the environment, control of the subjects expectations and placement of visual cues that is more influential than the numerical accuracy. The perception of spatial audio is as much in the mind as in the math.

The latter part of this work seeks to quantify this premise through research in to the psychology of hearing. Following on from a rigorous approach to spatial audio, the uncertainty of perception, as observed in many psychological experiments, can be characterized and compared to the fundamental limits of sound field reconstruction. The literature on the psychology of hearing is significant and cannot be ignored when the true end goal is to create the perception of an auditory event, not just recreate a sound field. The areas of the perception of spatial sound and the interaction of the various senses are discussed in detail.

[2] ICAD – International Community for Auditory Display – is a research body very active in the development and application of the audio sense for information display. `http:\\www.icad.org`

[3] Research and development for spatial audio at Lake Technology, Australia 1995-2000. Contribution to the design, development and commercialization of the Dolby Headphone algorithm. Unpublished work.

As noted in some other papers in this area, it is a multidisciplinary pursuit with a combination of systems engineering and human perception [19]. A commentary is provided here on the problem of creating convincing and effective spatial audio. The problem encompasses the mathematical, practical and psychological aspects of virtual audio. It is apparent that the best approach will consider all of these elements – consideration of any one aspect in isolation is not likely to give significant advances in performance. The objective is to provide guidance for the design of practical systems for the presentation of spatial audio.

1.2 Literature Review

This thesis provides a combined review of the fields of mathematical soundfield representation and psycho-acoustics. Previous works concentrate on either the mathematics of sound field reconstruction, or the psychological aspects of spatial hearing. Since these are fairly diverse fields, few works provide an in-depth analysis of both issues simultaneously. Table 1.1 table sets out some of the literature mapped into several categories.

It is interesting to note the evolution of the field among these areas over time. Figure 1.1 provides a graphical representation of this by enumerating the number of papers to appear each year. Although the literature review is quite thorough, it cannot be considered complete, and is some-what biased towards recent publications. However, examining this figure shows several interesting trends that relate to research work in Spatial Sound and the subject of consideration for this thesis.

The availability of DSP processors suitable for processing audio signals in real time has increased the level of interest in the mathematics and reproduction of spatial sound. From 1993 we see first the mathematical work increase, followed by systems that actually targeted a human listener. Interest in spatial sound systems using simple psycho-acoustic models peaked in 1998/9. The pioneering work of Gerzon in the early 70s and 80s, [14, 78], was well ahead of the technology capabilities. As this type of system has been only some-what capable of delivering the perception of good spatial sound, interest in the mathematical modelling and representation of the sound-field continues to increase to match the technological possibilities for implementation.

Interest in the explanation of spatial hearing through experimentation with simple stimulus, is a continuing field. The complexity and nuances of spatial hearing offer boundless potential for experimentation in this area. Most research in the area of perception attempts to break down the process of sound localization into independent components. There is an argument

Soundfield analysis and reconstruction. In depth analysis of the mathematics and numerical performance of a soundfiled or transucer array system without consideration of the perceptual issues. Essentially the problem of pure soundfield representation and reconstruction.	[3, 13, 20–46]
Spatial Sound Reproduction. Works on the reconstruction of a soundfield with the goal of a human subject as the listener. Typically they involve some simple models of the perceived location or spatial characteristics based on Interaural Time-Delays (ITD) and Interaural Intensity Differences (IID) or simple mathematical models based on phase or intensity geometry at the listening location.	[14, 47–78]
Localization phenomena with simple synthetic stimulus. Includes the lateralization experiments with phase and intensity effects of binaural hearing. These tend to be very contrived experiments to try and reveal the underlying processes of spatial hearing. The mathematical complexity of sound fields are not considered.	[79–103]
Localization phenomena with real sound sources. Experiments using real sound sources in combination with other senses to demonstrate psychological effects of spatial hearing and spatial awareness. Real sound sources are used with attempts to remove, disguise or mislead the visual cues associated with the sound source. Little consideration is given to mathematical issues.	[15, 80, 104–126]
Localization experiements with the simulation of complex sound fields. Typically the use of headphones to present spatial audio stimulus and measure the pshycological and performance effects of comninations of sensory stimulus. Some consideration of both the math and practical constraints along with the subjective perception of the sound.	[18, 127–149]

Table 1.1: Broad Categorization of Spatial Sound Literature

that this takes spatial hearing out of context and is not appropriate for understanding true spatial sound perception, [83, 89].

As computational systems become more capable at creating synthetic spatial sound, experimentation into the perception and response of subjects in virtual auditory environments had become an increasing area of interest. However, the study of spatial sound perception of real sound sources is largely neglected. Interestingly, prior to 1980, without suitable computational resources, this was the predominant mode of research into spatial hearing.

It is the premise of this thesis, that with the advent of computers and audio processing technologies, work on the perception of real sound sources is rarely carried out. It is the general assumption that provided the numerical accuracy of a system is sufficient, a synthetic presentation of spatial sound is suitable for experimentation. This is an assumption that goes largely unchallenged, to the point that when a subject's perception differs from the intended spatial sound presented, the cause is ultimately assumed to be numerical error or imperfection.

Few works combine both the latest in synthetic sound field modelling and reconstruction, with reference to the observations and properties of spatial hearing of real sound sources.

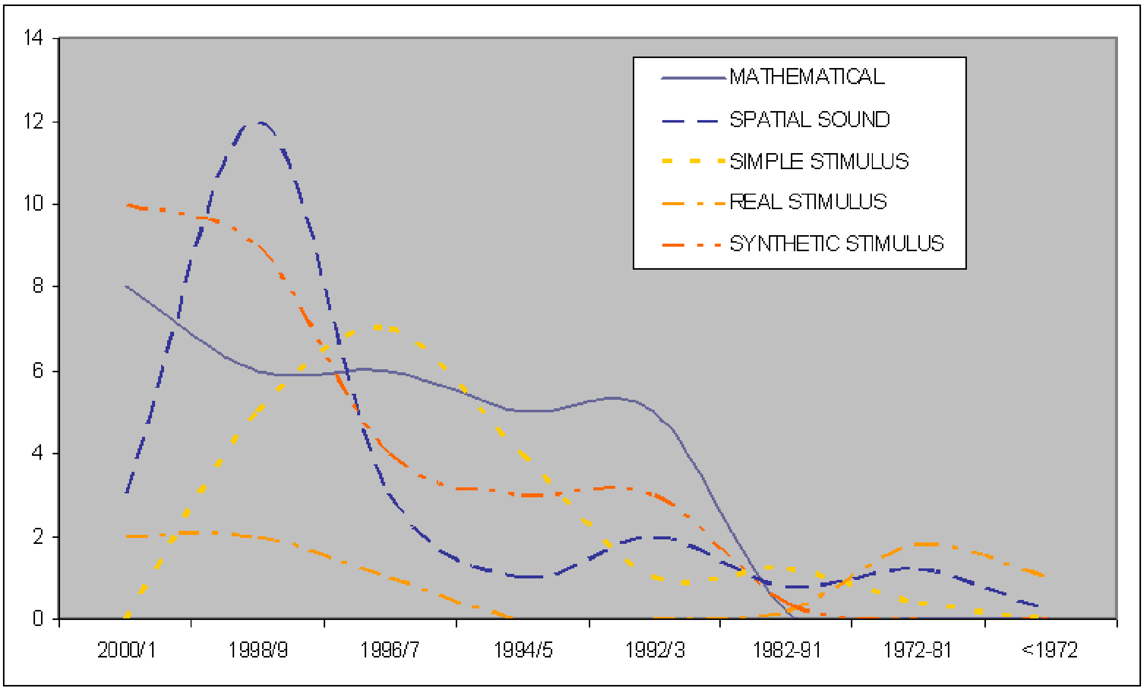

Figure 1.1: Timeline for Literature on Spatial Sound

This thesis sets out two complementary arguments regarding the issues in reconstruction of spatial sound :-

- The numerical benefits of increased channels and complexity in the representation and reconstruction of spatial sound is providing diminishing returns. Practical issues due to position uncertainty and environmental interaction, along with fundamental numerical limitation, indicate marginal improvements for substantially increased complexity.
- Hearing is an extremely adaptive, biased and associative sense that requires a great degree of consistency and continuity, both with itself and with other senses for deterministic results. A review of research in the field of psychology reveals just how much variation can be expected in spatial sound perception with real sound sources.

This leads to the premise that advances in the virtual presentation of spatial sound to achieve a determined perception by a subject are best obtained by considering not only the numerical presentation, but the factors of the subject's psychology and state of mind. These factors may include the environment, other sensory cues, method of introduction and the level of "cooperation" of the subject.

There are some other works that begin to consider both the reconstruction and perception of the sound field. This is the fifth category represented in Table 1.1 and has shown increasing interest in recent times. Another category of works, most of them recent, address specifically the perceptual implications of virtual audio. These works start to look at how the perceptual environment of a virtual auditory display differs from reality and therefore what features and characteristics of the sound field are most likely to benefit the perceptual results, [4, 5, 8, 11, 127, 132, 143, 144, 147, 148, 150–157]. The concepts and issues originated in these previous works are developed further throughout this thesis.

In summary, this thesis provides a basis to compare the significance of the perceptual and psychological effects against the limitations of practical sound field systems.

1.3 Current Spatial Sound System

Audio and Surround Sound systems have a strong commercial market. Music and audio is a popular form of entertainment both in isolation and in combination with visual information through movies and videos. Movie theatre and public auditoriums have a concern for sound as well as vision. The stereo and TV are common house-hold appliances and are now integrated into the Home Theatres. To provide a context for the area of research for this thesis, this section provides a brief review of some of the commercially available surround sound systems. Additional reviews of surround sound systems and the evolution of multi-channel audio can be found in [12, 57, 62, 158, 159].

What is Spatial Audio? It is the subtle characteristics of a sound field that allows us to place sound objects in space and time and also obtain a feel for the physical environment in which those sound sources are present. Spatial Audio reveals information about both active and passive acoustic objects surrounding us. We make constant use of spatial audio in our present state. It provides a sense of presence and awareness of things around us. The goal of a "Surround Sound" system is essentially to present spatial audio to listeners.

1.3.1 Types of Spatial Sound System

Acoustical events generate sound pressure waves or variations at the point of observation. This can be described as a soundfield – pressure variations about the mean air pressure. A listener at the observation point perceives this sound field predominantly by hearing. To simulate or reproduce an acoustical event we can either attempt to recreate the soundfield or just the effect on the ears [66]. The first method is known as holographic or soundfield techniques. The second method is known as binaural or transaural[4].

Holographic and soundfield systems aim to recreate the acoustical pressure field over a region of space. Such a system is independent of the listener in the sense that a representation of an acoustical event is recreated even if no listener is present. Soundfield systems typically use a multitude of speakers to reproduce sounds originating from different directions. The goal may not always be to recreate the exact sound field, but rather something which will be perceived as similar or convey the appropriate information. This type of system is well suited to delivering surround sound to a large audience. However to accurately represent a

[4] Binaural typically refers to the delivery of sound directly to the ears through headphones. Transaural generally refers to delivering a binaural signal through speakers [70, 71] – that is essentially controlling the sound field only at the ear locations. Processing of the binaural signals must be carried out to cancel the speaker cross talk and compensate for head related transfer functions.

soundfield over a large region of space and large bandwidth a very large number of channels would be required [58, 66, 72, 78]. To reproduce a soundfield accurately with a bandwidth of just 1kHz over a 10cm sphere would require more than 100 channels [23].

The principle of binaural and transaural systems is to reproduce an appropriate stimulation of the listener's auditory senses. By reconstructing the sound only in the region of the ears (as with headphones) it is possible to control the excitation of the listener's eardrums. This can result in the perception of complex spatial sound. Binaural systems deal with only two channels of information. Complexities arise with binaural systems due to the fact that each individual has a unique head related transfer function (HRTF) [38], however, with only two points of control, they have the advantage of a smaller acoustical reconstruction region. Less information is required to produce a surround sound experience and individual HRTFs and transducer transfer functions to the ear-drum can be fairly accurately measured and compensated for. This approach is considered to be an "optimal" approach to spatial sound delivery [15] and performs better than low order soundfield methods.

Once spatial audio is reduced to a binaural form, information is lost and the complete soundfield cannot be recreated. Without compensating for head movement, the listener cannot experience the effect of moving through the soundfield – small rotations of the head to help resolve sound source locations, do not have the desired effect. Full bandwidth transaural systems become very sensitive to the listener's location [49, 63] and speaker layout [59] – locations must be within a few centimeters. More speakers are needed to deliver to multiple listeners [48] and system performance is significantly degraded in practical reverberant rooms [22]. Even a theoretically optimal system would not perform well near and above 15kHz [25].

Figure 1.2 depicts the three classifications of spatial sound systems as set out above and Table 1.2 sets out a comparison of the advantages and disadvantages of these systems.

All three systems, in essence create some form of sound field. True soundfield systems create a soundfield that exists over a region of space independent of the listener. Binaural systems reconstruct the soundfield only in the listeners ears, by direct coupling through headphones or attached transducers of some form. Transaural systems aim to reconstruct a soundfield in the proximity of the listeners ears using remote and unattached transducers through cross-talk cancelling or small region sound field control. Binaural and Transaural systems may deliver a more accurate pressure signal to the ear-drum, but the suffer from being static (headphones) or incorrectly sensitive to listener movement (transaural).

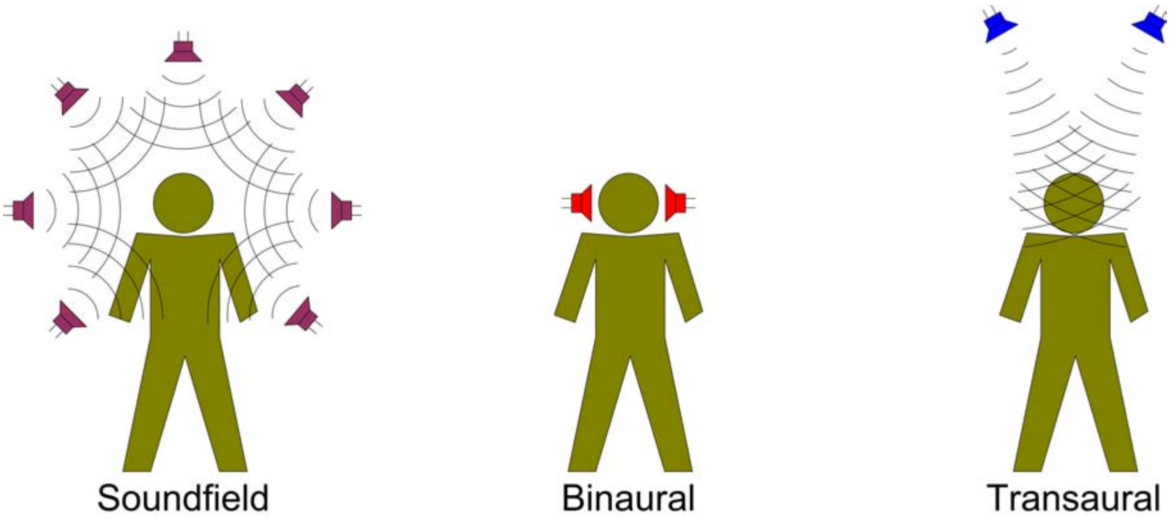

Figure 1.2: Diagramatic Representation of Three Classifications of Spatial Sound System

Type of System	Advantages	Disadvantages
Soundfield	Independent of listener and HRTFs. Extends to large listening regions for multiple listeners. Inherently allows freedom of movement for listener.	Only accurate over small listening areas. Large volumes require huge number of channels. Not suitable for private listening.
Binaural	Accurate delivery of spatial sound. Private listening. Listener always in ideal location.	Sensitive to individual HRTFs. Does not inherently cater for head movement. Physical presence of headphones effects perception.
Transaural	Accurate spatial sound delivery when in correct position. No physical presence of headphones.	Sensitive to individual HRTFs. Very sensitive to listener position. Does not inherently cater for listener movement.

Table 1.2: Comparison of Spatial Sound Systems

1.3.2 Current Practical Systems

To understand the scope for practical and commercial application of spatial audio delivery, it is necessary to look at the history and extent of use of several common sound formats.

- **Stereo** – although strictly not confined to two channels [70, 159], since the LP phonograph it has become synonymous with two channel recording and playback. In its basic form a stereo system is a soundfield type system which can create the illusion of virtual sound sources between the two speakers where the listener is positioned centrally at the "sweet spot". As the listener moves away from this region, the quality of the sound field imaging will degrade – sound images initially between the speakers will become unstable and collapse to the nearest speaker. Good quality stereophonic equipment can provide a very enjoyable listening experience, though not truly spatial.

- **Quadraphonic** – several techniques were developed for delivering more than two channels of audio in the early 70s. Typically the additional two channels were encoded on top of the existing two channels. There were several standards, each requiring its own decoder. Quad systems represent a soundfield type surround system covering 360 degrees in a planar array. Program material was not well mixed, and the technology cost and confusion at the time prevented any significant commercial penetration.

- **Matrix Stereo** – in the mid to late 70s, matrix techniques similar to those used for quadraphonic were used for film sound [159]. Four channels – left, centre, right and surround – could be encoded on only two audio channels. The popular Dolby MP Matrix encoded the extra channels using matched and opposing phase additions to the left and right channels (Figure 1.3 below from [160]). The passive matrix decoding process is degenerative and can only achieve 3dB channel separation. "Dolby Stereo" became the standard distribution format for film sound using optical encoding of the Dolby MP Matrix two-channel sound-track on the film [159].

  Rather than creating an accurate sound field, these systems are concerned with reproducing a desired cinematic experience – centre channel for dialog, surround channel for envelopment and ambience. With video reorders and CDs in the 80s, matrix stereo surround found its way into the home.

- **Ambisonics** – a hierachical system for the optimal representation of sound fields developed in the early 70s. It is based on spherical harmonic expansions of the wave equation [14, 74, 75, 78]. The techniques involved have been developed from a combination of acoustical control and psycho-acoustic principles to correctly match important auditory localization cues. Ambisonics is not tied to any particular speaker layout and represents a very flexible and generic way of recording, simulating and re-creating three dimensional sound. It has found applications in large audio displays for virtual reality, artistic presentations and theme park entertainment.

Niche consumer decoders offer Ambisonic decoding and a limited amount of program material is available. Although the system is well founded in theory and can deliver excellent results, in practice it has not found a market beyond surround enthusiasts [161]. The process of correctly setting up an Ambisonic decoder and speaker array is quite complex. Hopefully support for this format will continue – it represents a good option for true sound field recordings on emerging multi-channel audio media [60, 65]. Other works and extension on Ambisonics deal with hierachical formats surround sound transmission and multi-speaker array reconstruction [69, 162, 163].

- **Active Matrix Decoding** – to improve the decoding process active steering was introduced. Through monitoring the relative levels of the decoded channels, the intended sound direction can be estimated. An additional gain stage is included in the decoder to "steer" the sound in that direction, emphasizing the effective channel separation. Dolby Pro Logic incorporates active steering and became available in consumer products in 1987. This gave an effective channel separation for a steered sound source of 37dB [160].

  The process of active steering can introduce artifacts. Various decoding algorithms improved on the process with improved steering logic, banded frequency steering and increased number of output channel – Circle Surround, Pro Logic II and Logic 7. Active decoding of matrix stereo is by far the largest installed base of surround sound systems with over 50 million units sold[5]. All matrix formats have the fundamental limitation that the derived channels are not totally independent.

- **Multi-channel Digital Formats** – digital media offers the ability to store multiple audio channels. With the main motivation being for theatre sound Dolby Digital 5.1 (AC-3) and DTS[6] were introduced around 1992. Both use compression algorithms to reduce the multi-channel sound track to a manageable amount of digital data [164]. Dolby Digital typically uses a higher compression ratio than DTS. Qualitative comparison of the two compression schemes is a contentious issue, both offer 6 independent channels of audio for a cinema configuration.

  The 5.1 environment has become a defacto standard and refers to three front channels, two rear or surround channels and a low frequency effect (LFE) channel (Figure 1.4).

[5] Data from Dolby Laboratory Statistics, July 2002.

[6] DTS used as the trademark for the digital audio format created by Digital Theatre Systems Inc.

The rear channels are usually line arrays of speakers in a theatre to reproduce a diffuse sound field. Dolby Digital has secured a greater market share due to its use as a standard in Laserdisc, DVD and digital TV. Movie material in the home is brought to life by multi-channel digital audio. As directors and sound engineers are learning to take full advantage of the format, new release movies are delivering powerful and interesting surround sound. The goal is to create a compelling and enveloping soundfield rather than an accurate soundfield as speaker geometries and installations vary considerably.

- **Virtual Surround Sound** – Multi-channel surround formats raise an issue of convenience for consumers with full 5.1 setups being costly and impractical in domestic situations. New technologies using binaural and transaural techniques are emerging. The term "virtual surround" is being used to describe products that create the impression of a surround sound system using headphones or two speakers [165].

  Headphone systems use convolutional or infinite impulse response filters to simulate the directional and room response for the virtual sound channels at left, centre, right and surround. The resulting auditioning experience can be vastly superior to a simple 2 channel down mix. Virtual systems using speakers use filters on the surround channels to create the impression of a diffuse sound field to the side or behind the listener. As with all transaural systems, there is a trade off between the quality of the surround effect and the range of listening locations it will cover ("sweet spot" size). They are well suited to an individual listener in a known location [165].

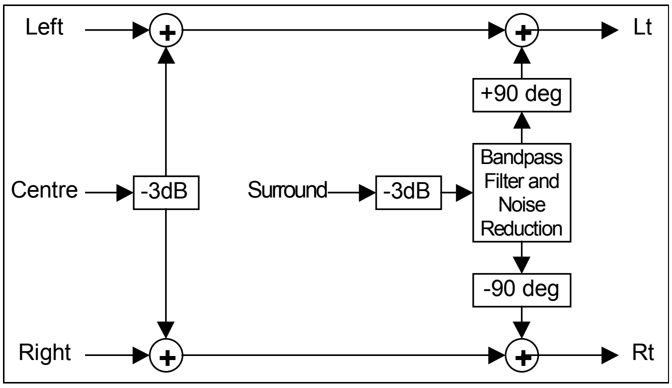

Figure 1.3: Schematic of Dolby MP Matrix Encoder

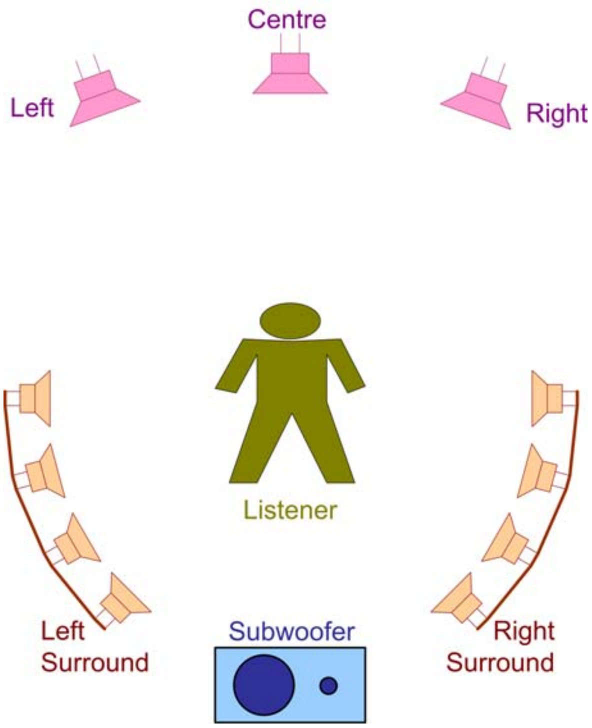

Figure 1.4: Configuration for 5.1 Speaker Array for Consumer Surround Sound

1.4 Problem Definition

Where the goal is to generate spatial sound for human perception, the ultimate goal is the correctness or appropriateness of the perception. Too often it is assumed that this is best achieved by creating a numerically correct soundfield over a region of space or at the listener's ears. As pointed out by Dermont Furlong [72], too much of audio engineering is a "black art" since the goal is rarely well defined. In particular reference to the use of a domestic audio system for recreating concert music he states

> "the function of the audio recording and reproduction system should be to optimally reconstruct the concert hall **experience** for the domestic living room listener"

The goal is really "plausibility" rather than "authenticity" [166], so it may not be necessary to reproduce accurately all physical aspects. What is important is that the features of the original soundfield that are perceptually important are preserved [72].

The issue of experience is not one that is well dealt with amongst the engineering disciplines. Since the experience of a listener is subjective and largely unobservable, very little work attempts to optimize the experience of a spatial sound reconstruction system. Instead, the engineering approach leans toward optimizing the reconstruction of the soundfield or spatial sound itself. As will be shown in this thesis, a soundfield is a complex entity and the pursuit of perfection in reconstruction is a boundless problem.

The problem addressed by this thesis is in a sense, where should our best efforts be placed to optimize the **experience** delivered by a spatial sound system. This is addressed as three component questions:-

- How numerically good can spatial sound systems get in practice? How accurate can a soundfield or binaural signal be recorded, created and delivered?
- How reliable and predictable is perception. Assuming that spatial sound delivery can be perfected, what uncertainty exists in the perception and experience of a spatial sound?
- Can we compare the two measures of uncertainty. Which is the bigger uncertainty and what can we do to mitigate the influential factors?

Essentially the overall question is,

> **"What is more important – accurate soundfield / binaural reconstruction or the management and control of perceptual influences?"**

Through the analysis of spatial sound reconstruction it is shown that binaural delivery of spatial sound is less complex and more feasible for a single controlled listener. As already stated in Section 1.3.1, headphones can be considered optimal [15]. From the work and experience of the author [16] it was already known that for binaural delivery headphones, increasing the level of accuracy and individually matched Head Related Transfer functions and headphone transducer and coupling responses, was not the most effective means of creating a compelling headphone presentation.

The design and delivery of a compelling 5 channel headphone virtualizer [167] provided an opportunity to qualitatively assess this observation over many subjects. Despite the use of low quality headphones, non-individualized, poorly equalized transfer functions and non-ideal demonstration environments, the illusion of static, compelling and external spatial sound was extremely consistent. Rather than attempting to measure and reconstruct individual and well matched binaural responses, the demonstration was optimized with for perceptual considerations. Lightweight headphones were used, with a real 5 channel speaker array present the subjects were initially introduced to real sound sources similar to that of the virtual audio and provided with visual cues, the acoustic properties of the virtual space were similar to the real space and the transition to the headphone delivery was made gradually. The demonstration audio was typically a test voice panning around the speaker array. At the end of the demonstration, subjects removed the headphones and were shocked to find the room silent. This demonstration was instrumental in the interest and commercial adoption of Lake's technology as Dolby Headphone[7].

So under the right conditions, a poor virtual simulation was convincing. In other experiments, laborious measurement and correction of an individuals response to achieve the most numerically accurate results proved less than satisfactory. Particularly when that subject auditioned the virtual audio in an environment that was in conflict to the environment of the response measurement – that is with conflicting visual cues or substantially different room acoustics.

In any situation where a binaural presentation fails to deliver compelling spatial sound, it is always the numerical accuracy and complexity of the system that is assumed to be the

[7] Commercial information on Dolby Headphone is available at www.dolbyheadphone.com along with a some-what simplified technical explanation of the process.

cause. It is very difficult to show experimentally that this is not the case. We cannot substitute the virtual spatial audio for real spatial audio in the same listening configuration and assess the perceptual performance independent of any “synthetic” or reconstructed acoustic components.

A premise behind this thesis is that this it is the perceptual more than numerical or system limitations that causes binaural simulation of spatial audio to fail or become ambiguous. This leads to a significant problem and question addressed by the thesis,

> **“Is headphone presentation of binaural spatial sound good enough?”**

The problem is to resolve this question with an answer along the lines of “Yes they are, as given the same situation, even real sound would create a similar misleading or ambiguous perception”. There is no expectation that this argument will divert the blame from the numerical binaural simulation – the belief that we hear accurately and unambiguously is far too strong. Rather the problem and argument is presented to those working in the field of binaural spatial sound reconstruction to answer the question above, and also provide a measure of sanity before enormous effort is placed into what may be an unrewarding pursuit.

1.5 Chapter Overview

This thesis examines the conjecture that the numerical precision considering the pursuit of numerical precision in isolation will not lead us to compelling spatial audio. Truly convincing audio may not be possible without a truly convincing environment. This thesis is divided up into a set of chapters moving from the mathematical into the psychological aspects of spatial sound perception and onto a comparison and conclusions.

Chapter 2 deals with the possible mathematical representations of sound fields. In order to record, recreate and analyze the performance of soundfield systems, it is important to have a suitable mathematical framework. This chapter covers several different frameworks and shows a result of equivalence between them. The most comprehensive representation is shown to be that resulting from the natural harmonic solutions of the wave equation.

Given a framework in which we can represent and analyze a soundfield, Chapter 3 deals with the problem of reconstructing such a soundfield around a listener. As the desired bandwidth and size of the reconstruction region increases, the order of the soundfield and number of channels required will also increase. Chapter 3 draws some fundamental mathematical and practical limitations to the ability to accurately recreate a soundfield. Headphones present an alternate way of delivering spatial sound. With only two points of reconstruction, binaural systems are considerably less complex than soundfield systems.

The quality and compellingness of spatial sound as perceived by a listener will have a degree of uncertainty. Chapter 4 looks at the process of sound perception. There is a wealth of information in the last century of experimental psychology relating to spatial hearing. A substantial amount of research and experimental work demonstrates the adaptive, ambiguous and biased nature of spatial sound perception. Chapter 4 draws from this background specific examples and data to provide an indication of the uncertainty and error introduced by the actual perception of either a soundfield or a binaural presentation. We cannot simply be compelled to perceive a specific spatial sound event through accurate acoustical reconstruction alone. If the audio sensory input is in conflict with other sense modalities or does not represent a coherent and plausible reality the perception may not be as desired.

Chapter 5 provides a comparison of the issues in reconstruction of a spatial sound with the psychological issues discussed in the previous chapter. The results suggest that the error in perception is comparable with the error in the soundfield or binaural reconstruction. This leads to the argument that current conventional systems can provide an adequate auditory stimulus – increased numerical accuracy will only give marginal perceptual improvement. Numerical accuracy does not imply a compelling spatial sound system. Similarly, the subjective measure of whether a system provided a compelling audio simulation cannot be used as an indicator of numerical accuracy.

An important issue raised in Chapter 5 is the asymmetry in spatial sound perception experiments. It is possible to present real spatial sound with alternate or distracting visual cues – the example and possibilities of the ventriloquist demonstrates this. However it is extremely difficult to provide a visual indication of the absence of an audio source – an invisible speaker. The point to note here is that **the visual cue of an absence is very different to the absence of a visual cue**. Most experiments will conceal the sound source or suppress the visual sense – this only creates uncertainty rather than providing a strong contradiction. This specific asymmetry is discussed in Section 5.5 and a specific experiment is discussed to help resolve this. This experiment has been conducted previously and from the results of this experiment, it is argued that even a real sound source will be perceived incorrectly, if there is a compelling visual cue that suggests the absence or impossibility of there being an object generating that sound at the perceived spatial location.

Chapter 6 brings together the research, results and discussions of the thesis to answer the specific questions raised in Section 1.4. Suggestions are made for where research effort should be placed, as well as outlining some problems of further interest. From this thesis it is apparent that the perception of a spatial sound system will depend on much more than just numerical accuracy. Spatial sound systems are already good enough to provide plausible compelling and consistent spatial audio to match other sensory input – how well it deceives the listener depends on the listeners willingness to believe. Although this initially appears to be a weak argument, it is the research and ideas presented in this thesis that make it compelling and ultimately shows that it will be a limitation of any spatial sound system, no matter how accurate.

Chapter 2

Soundfield Representation

This chapter deals with the mathematical representation of a soundfield. As set out in Section 1.3.1, soundfield systems deal with the representation or reconstruction of the acoustic activity over a region of space. Some soundfield systems do include reference to basic models of human auditory spatial perception, however these are typically just simple expressions of geometry [68, 77] such as the phase vector or intensity vector[1]. Generally, soundfield systems are concerned with the numerical accuracy of the representation and reconstruction of a soundfield over a region of space rather than psychoacoustic and perception issues. This chapter is from the author's work carried out in 1998/9 towards a paper which was presented at the 106th Convention of the Audio Engineering Society in Munich 1999 [30].

In Section 2.1 we introduce a soundfield as a varying pressure field over a region of air. We can use a classic Taylor series representation to represent such a soundfield. The concept of spherical harmonics and the theory behind Ambisonics is introduced as an alternate representation in Section 2.2. This is shown to be a more efficient representation than the Taylor series, however they represent similar information. Beyond this, the more complete form of modal analysis is introduced in Section 2.3 to show a natural set of spatial basis functions for representing soundfields. These also relate to the spherical harmonics, but include radial dependence for a complete spatial description of a soundfield based on eigenfunctions of the wave equation. Full modal analysis provides a sound theoretical foundation to examine the spatial and frequency bandwidth constraints of sound field representations. This is used in Chapter 3 to investigate and address some of the issues of sound field recording and playback. Another technique known as Wavefield Synthesis is discussed in Section 2.4.

Ambisonics is shown to be a subset of the full modal analysis and the chapter concludes

[1] The phase vector and the intensity vector are the vector sums of the amplitude and energy of the incident sources (speakers).

in Section 2.5 with a discussion on the utility of using modal analysis as a framework for analysing soundfield representation and reconstruction.

2.1 Taylor Series Approximation

We can think of a soundfield as the variations of pressure over a region of air around the average pressure

$$p\left(x,y,z,t\right)=P\left(x,y,z,t\right)-P_{avg} \tag{2.1}$$

A Taylor series expansion of the instantaneous pressure field about a point gives an intuitive Cartesian representation of the sound field. Since the soundfield is constrained by the physics of acoustical propagation, it can be assumed to be analytic and thus is suited to a Taylor series representation. The Taylor field expansion is defined by the partial derivatives of the sound field around a point,

$$p_{l,m,n}\left(t\right)=\left.\frac{\partial^l}{\partial x^l}\frac{\partial^m}{\partial x^m}\frac{\partial^n}{\partial x^n}p\left(x,y,z,t\right)\right|_{x=x_0,y=y_0,z=z_0} \tag{2.2}$$

Given the signals defined by (2.2), the sound field synthesis equation can be written

$$p\left(x,y,z,t\right)=\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}p_{l,m,n}\left(t\right).\left(x-x_0\right)^l\left(y-y_0\right)^m\left(z-z_0\right)^n \tag{2.3}$$

Without any loss of generality, we can consider the expansion about the coordinate origin, reducing this to,

$$p\left(x,y,z,t\right)=\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}p_{l,m,n}\left(t\right)x^ly^mz^n \tag{2.4}$$

The sound field is reduced to a set of time varying coefficients. For complete description this is an infinite set. For an approximate description we can truncate the order of the expansion. For a Taylor series expansion, we can define the order of the expansion as the highest power term that is used. At each successive order (N), an additional number of terms are introduced related to the possible combinations of l,m,n such that $l+m+n=N$. A higher order representation will more accurately represent the sound field. The terms for a Taylor series expansion up to third order are shown in Table 2.1.

Order	New Terms	Total Terms	Taylor Expansion Terms
0	1	1	1
1	3	4	x, y, z
2	6	10	x^2, xy, xz, y^2, yz, z^2
3	10	20	x^3, x^2y, x^2z, xy^2, xyz, xz^2 y^3, y^2z, yz^2, z^3
N	$\frac{1}{2}\left(N^2+3N+2\right)$	$\frac{1}{6}\left(N^3+6N^2+11N+6\right)$	

Table 2.1: Taylor Series Terms

2.2 Ambisonic Representation

Ambisonics is built on an encoding of a soundfield using the direction of arrival of sounds. It is known that plane waves form a suitable basis set for describing arbitrary sound fields [168]. The theory of Ambisonics is a hierachical model for representing and reproducing soundfields which also includes optimal decoder theorems for reproducing sound fields based on psychoacoustic criteria. These psycho-acoustic models are simple first and second order models based on optimizing the reconstructed phase and energy vectors of the soundfield [68]. In this paper we only consider the sound-field representation at the centre of Ambisonic theory. In this framework, the sound field is represented by a set of signals that would be obtained by microphones with specific directionality patterns [14, 76–78]. Ambisonic techniques have received a lot of attention and formed the basis of practical sound field recording apparatus [74, 75].

An arbitrary sound field can be represented by an arbitrary superposition of plane waves. This form of representation is also known as the Herglotz wave function [168],

$$p\left(\mathbf{x},t\right)=\int_{\Omega}e^{ik\mathbf{x}.\mathbf{d}}g\left(\widehat{\mathbf{d}}\right)ds\left(\widehat{\mathbf{d}}\right) \tag{2.5}$$

where $\mathbf{x}\in\mathbb{R}^3$ and $\widehat{\mathbf{d}}\in\Omega$ is the unit vector integrand. This equation shows that any soundfield is can be represented as a function over solid angle of the "directionality" of the sound. The spherical harmonics of associated Legendre polynomials form a natural basis to represent such a class of functions. Ambisonic signals are defined by a set of channels having the directionality properties of the associated Legendre polynomials. Where the direction of the incident sound is represented as a unit-vector in Cartesian coordinates the channel gains or sensitivity are given in the Table 2.2. It can be seen that the Ambisonic expansion has fewer terms than the Taylor series expansion for the same order. This is discussed further in Section 2.5.

The coefficients of these functions normalize the "pickup" energy of the channel (the sec-

Order	**New Terms**	**Total Terms**	**Ambisonic Directional Sensitivity**
0	1	1	1
1	3	4	$\sqrt{3}x,\ \sqrt{3}y, \sqrt{3}\ z$
2	5	9	$\frac{\sqrt{15}}{2}\left(x^2-y^2\right),\ \frac{\sqrt{15}}{2}xy,\ \frac{\sqrt{15}}{2}xz$ $\frac{\sqrt{15}}{2}yz,\ \frac{\sqrt{5}}{2}\left(3z^2-1\right)$
3	7	16	$\sqrt{\frac{35}{8}}x\left(x^2-3y^2\right), \sqrt{\frac{35}{8}}x\left(x^2-3x^2\right)$ $\sqrt{\frac{21}{8}}x\left(1-5z^2\right), \sqrt{105}xyz$ $\sqrt{105}\left(x^2-y^2\right)z, \frac{\sqrt{7}}{2}\left(5z^3-3z\right)$
N	$2N+1$	$(N+1)^2$	

Table 2.2: Ambisonic Directional Channel Gains

ond order norm of the function across the unit sphere). This set of functions represents an orthonormal basis over the constraint $x^2+y^2+z^2=1$ resulting from the unit vector representation of direction. These functions are spherical harmonics, related to the associated Legendre polynomials (discussed further in later section).

Note that the terms in the Taylor series and Ambisonic series seem similar in nature. However, the Taylor terms represent the basis functions for representing the pressure field itself while the Ambisonic series represent a basis function for encoding a sound-field based on directional sensitivity of a "theoretical" microphone to infinite plane waves.

The Ambisonic reconstruction equations assume the contribution of a speaker array at the origin to be a plane wave from the incident speaker direction. The contribution of each speaker to the origin, $\mathbf{S}_n$, is the Ambisonic coefficients for that direction scaled by the reciprocal of the distance. The set of speaker contribution vectors forms the speaker contribution matrix $\mathbb{S}$. An arbitrary Ambisonic soundfield is then reconstructed by applying the pseudo inverse of the speaker contribution matrix to the encoded Ambisonic signal,

$$
\begin{aligned}
\mathbf{S}_n &= \frac{1}{\|\mathbf{x}_n\|}\begin{bmatrix} 1 & \sqrt{3}\widehat{x}_n & \sqrt{3}\widehat{y}_n & \sqrt{3}\widehat{z}_n & \frac{\sqrt{15}}{2}\left(\widehat{x}_n^2-\widehat{y}_n^2\right) & \dots \end{bmatrix}^T \\
\mathbb{S} &= \begin{bmatrix} \mathbf{S}_1 & \mathbf{S}_2 & \mathbf{S}_3 & \dots & \mathbf{S}_N \end{bmatrix} \\
\mathbf{u}(t) &= \begin{bmatrix} u_1(t) \\ u_2(t) \\ \dots \\ u_N(t) \end{bmatrix} = \mathbb{S}^T\left(\mathbb{S}.\mathbb{S}^T\right)^{-1}\begin{bmatrix} A_1(t) \\ A_2(t) \\ \dots \\ A_M(t) \end{bmatrix}
\end{aligned}
\tag{2.6}
$$

where $\mathbf{x}_n=(x_n,y_n,z_n)$ is a speaker position, $\widehat{\mathbf{x}}_n=(\widehat{x}_n,\widehat{y}_n,\widehat{z}_n)=\mathbf{x}_n/\|\mathbf{x}_n\|$ is a speaker direction unit vector, $\|\cdot\|$ is the Euclidian norm, $\mathbf{u}(t)$ represents the N speaker signals and

$A_m(t)$ are the M Ambisonic signals. Ambisonics effectively separates the encoding and decoding processes making it independent of the number of transducers used for recording or playback. It has been considered as a "generic" format [60] and compared to other popular encoding formats [65]. Because of its hierachical nature and relative simplicity of the encoding and decoding equations it has been used in professional audio recording and reproduction (See Section 1.3.2).

Note that with Ambisonics, the speaker gain matrix used ($\mathbb{S}^T \left(\mathbb{S}.\mathbb{S}^T\right)^{-1}$) is frequency invariant and real valued. This comes from the assumption of the reconstruction speakers being at an infinite distance and creating ideal plane waves at the origin. Ambisonics, by virtue of its formulation, creates a sweet spot region of reconstruction around the origin whose size is frequency dependent.

2.3 Modal Analysis

Sound propagates as a variation of the instantaneous air pressure around the mean pressure. For typical listening conditions, the change in pressure is of the order of .2 Pascals (Pa) with the threshold of pain being at around $20\,\mathrm{Pa}$. Given the standard atmospheric pressure of $100\,\mathrm{kPa}$, this represents a shift about the mean pressure of 0.02%. Thus for typical audible acoustic events, air can be considered a elastic and lossless medium. In such conditions acoustic propagation is well modeled by the first order wave equation [46],

$$\nabla^2 p\,(x, y, z, t) = \frac{1}{c^2}\frac{\partial^2}{\partial t^2} p\,(x, y, z, t) \tag{2.7}$$

Separating the time varying solution ($e^{i\omega t}$) from this equation gives the Helmholtz spatial wave equation,

$$\nabla^2 p\,(\mathbf{x}) + k^2 p\,(\mathbf{x}) = 0 \tag{2.8}$$

where k is the wave number ($k = 2\pi f/c = \omega/c = 2\pi/\lambda$).

In a region of space with no sources, the wave equation can be utilized to reduce the amount of data necessary to represent a sound field in that space. This is because any valid sound field is constrained to satisfy the wave equation. Furthermore, the basis function decomposition of the solution to the wave equation (spherical harmonic modes) presents a useful framework for representing sound fields. For example, such a modal analysis has been used as an effective tool in recent work on broad band beam-forming microphone arrays [169–172].

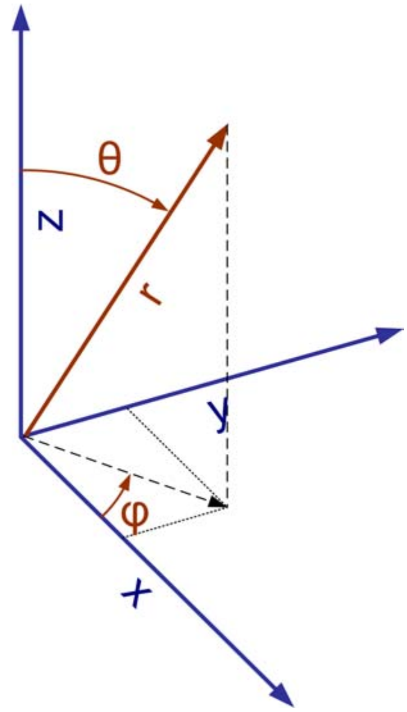

Figure 2.1: Diagram of Spherical Coordinate System

The Taylor series expansion is tied to derivatives of the sound field at a point and the Ambisonics description is a basis function set based on the concept of plane wave directionalities and microphone (or detection) directionality sensitivities. The modal description of a sound field is based on the eigenfunctions of the wave equation.

The solutions are more naturally presented in spherical coordinates where the sound field is represented by the pressure field $p\,(r, \theta, \phi, t)$. The spherical coordinate system used is shown in Figure 2.1. The wave written in spherical coordinates with the appropriate ∇^2 operator [173]

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}p\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}p\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}p=\frac{1}{c^2}\frac{\partial^2}{\partial t^2}p\,(r,\theta,\phi,t) \quad (2.9)$$

In solutions to this equation, the time variation can be separated from the space variables. The time variation solution for the wave equation can be represented by the rotating complex component $e^{i\omega t}$. Spatial solutions to this equation can be represented using the following entire series expansion [168],

$$p(r,\theta,\phi)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}\left(4\pi i^n\alpha_n^m\right)j_n\left(kr\right)Y_n^m\left(\theta,\phi\right) \quad (2.10)$$

where α_n^m are the complex series coefficients, $j_n\,(kr)$ is the spherical Bessel function of integer order n and $Y_n^m(\theta, \phi)$ are the spherical harmonics (orthonormal basis function on the unit sphere) given by

$$Y_n^m(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos\theta) e^{im\phi} \tag{2.11}$$

where $P_n^{|m|}(\cdot)$ are the associated Legendre functions.

Given the orthogonality properties of the spherical harmonics, the complex signal coefficients α_n^m can be calculated by considering the projection of the sound field onto the basis functions,

$$\alpha_n^m = \iiint p(r, \theta, \phi) \left(4\pi i^n j_n(kr) Y_n^m(\theta, \phi)\right) r \sin\theta dr d\theta d\phi \tag{2.12}$$

where the region of interest is a closed volume excluding any sound sources, this can be reduced to an integral on the surface by virtue of the Kirchoff-Helmholtz integral [168, 171, 172],

$$\alpha_n^m = 4\pi i^n \oint_S p(r, \theta, \phi) j_n(kr) Y_n^m(\theta, \phi) dS \tag{2.13}$$

and considering the region of interest as a sphere centred at the origin with radius R

$$\alpha_n^m = 4\pi i^n j_n(kR) \iint p(R, \theta, \phi) Y_n^m(\theta, \phi) \sin\theta d\theta d\phi \tag{2.14}$$

Thus the modal coefficients are frequency dependent $\alpha_n^m(\omega)$ where $k = \omega/c$. Note that initially the rotating complex component of the sound filed was removed, thus the spatial pressure field is implicitly also a function of frequency. To make this explicit

$$\alpha_n^m(\omega) = 4\pi i^n j_n(kR) \iint p(R, \theta, \phi, \omega) Y_n^m(\theta, \phi) \sin\theta d\theta d\phi \tag{2.15}$$

Given a band limited sound field, the signal coefficients $\alpha_n^m(\omega)$will also be of finite extent in the frequency domain. The Fourier transform of these coefficients provides a set of time varying signals representing a modal decomposition of the sound field $B_n^m(t)$. An approximation to a sound field can be obtained by truncating the expansion at a particular effectively reducing the number of terms ($\alpha_n^m(\omega)$ or $B_n^m(t)$). This truncation reduces the spatial resolution of the soundfield representation. Given an order of truncation $n = 0 \ldots N$ the number of complex signal coefficients required $\alpha_n^m(\omega)$ can be seen to increase as $(N+1)^2$ as shown in Table 2.3.

The angular characteristics of some modes are shown in Figures 2.2 through to 2.7. The

Order	**New Channels**	**Total Channels**
0	1	1
1	3	4
2	5	9
3	7	16
N	$2N+1$	$(N+1)^2$

Table 2.3: Channel Numbers for Modal Representation

unique spatial patterns are formed by plotting the modal magnitude at a fixed radius. The angular characteristics of the modes form a complex orthogonal set including structures with planar nodules (Figures 2.2 and 2.3), extended nodes along the Z axis (Figures 2.4 and 2.5) and radially dispersed nodules (Figures 2.6 and 2.7).

The angular structure of the modes is related to the associated Legendre polynomials, which are also related to the Spherical Harmonics of Ambisonics. Noting that $\widehat{z} = \cos\theta$, $\widehat{x} = \cos\phi\sin\theta$, $\widehat{y} = \sin\phi\sin\theta$ and $e^{jm\phi} = \cos m\phi + i\sin m\phi$, using the appropriate trigonometric identities and separating the real and imaginary parts when $m \neq 0$, the polynomials of the Ambisonic an modal spherical harmonics can be shown to be directly related. This is discussed further in Section 2.5.

As for Ambisonics, we can construct a reconstruction matrix for a speaker array using the modal analysis framework. The contribution from each speaker is obtained by assuming the modal excitation at the origin of a point source at the speaker location. From the addition theorem for the expansion of the fundamental point source wave field [168], we can calculate the α_n^m coefficients for a point source at $\mathbf{y}$ [172],

$$\begin{aligned}
\Phi\left(\mathbf{x},\mathbf{y}\right) &= \frac{ik}{4\pi} h_0^{(1)}\left(k\left\|\mathbf{x}-\mathbf{y}\right\|\right) = \frac{e^{ik\|\mathbf{x}-\mathbf{y}\|}}{4\pi\left\|\mathbf{x}-\mathbf{y}\right\|} \\
&= ik\sum_{n=0}^{\infty}\sum_{m=-n}^{n} j_n\left(k\left\|\mathbf{x}\right\|\right) h_n^{(1)}\left(k\left\|\mathbf{y}\right\|\right) Y_n^m\left(\widehat{\mathbf{x}}\right)\overline{Y_n^m\left(\widehat{\mathbf{y}}\right)}
\end{aligned} \tag{2.16}$$

This expansion is only valid for the region about the origin through to the source radius, $\|\mathbf{x}\| < \|\mathbf{y}\|$. Combining this with (2.10), we can solve for α_n^m to obtain the coefficients for a point source as a function of frequency,

$$\alpha_n^m\left(\mathbf{y},\boldsymbol{\omega}\right) = \frac{\left(-i\right)^{n-1}k}{4\pi} h_n^{(1)}\left(k\left\|\mathbf{y}\right\|\right)\overline{Y_n^m\left(\widehat{\mathbf{y}}\right)} \tag{2.17}$$

where $h_n^{(1)}\left(\cdot\right)$ is the (spherical) Hankel function of the first kind, mode n. Given an order of truncation P, and number of speakers N we can construct the speaker contribution matrix $\mathbb{S}$

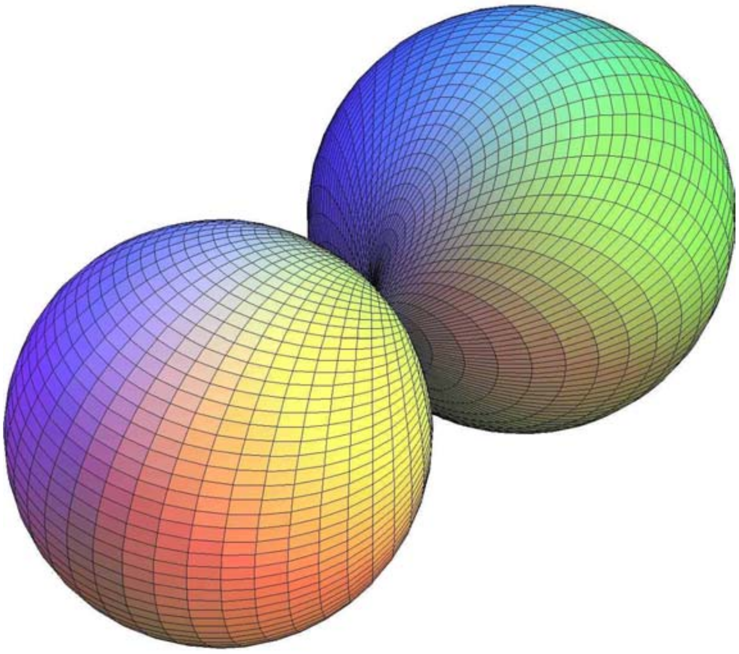

Figure 2.2: First Order Spherical Harmonic $n = 1 \quad m = 1$

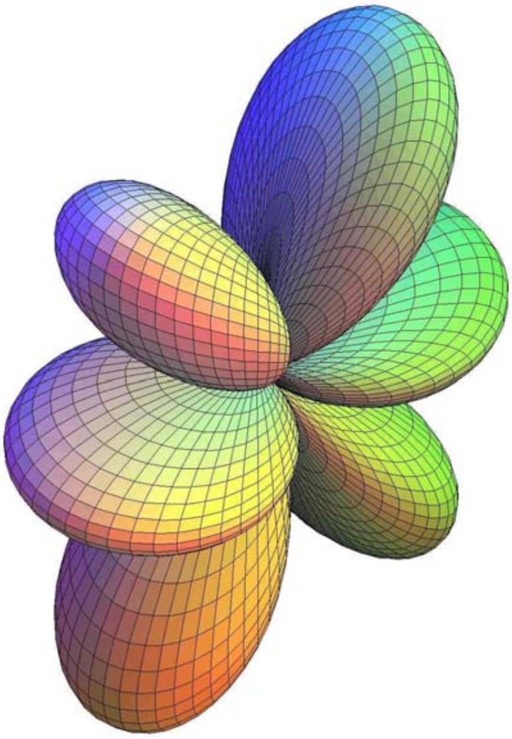

Figure 2.3: Third Order Spherical Harmonic $n = 3 \quad m = 1$

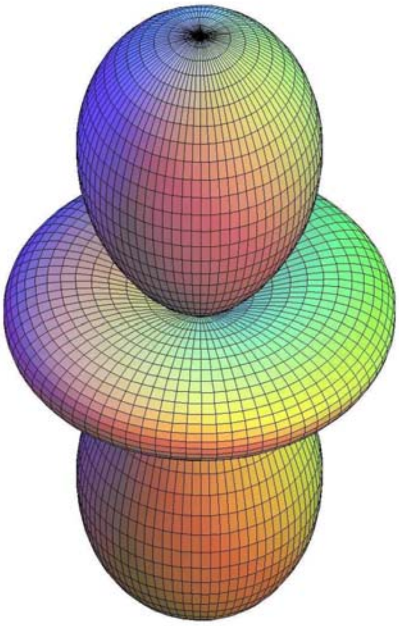

Figure 2.4: Second Order Spherical Harmonic $n = 2 \quad m = 0$

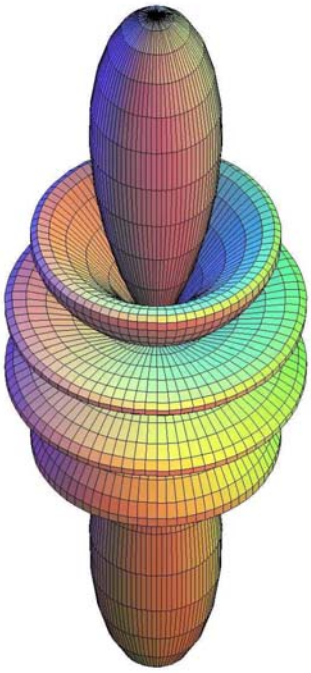

Figure 2.5: Fifth Order Spherical Harmonic $n = 5 \quad m = 0$

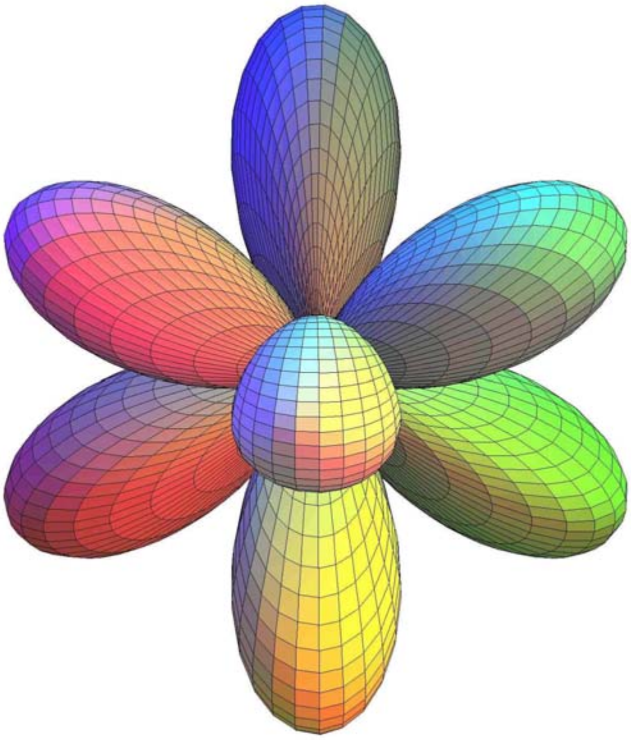

Figure 2.6: Third Order Spherical Harmonic $n = 3 \quad m = 2$

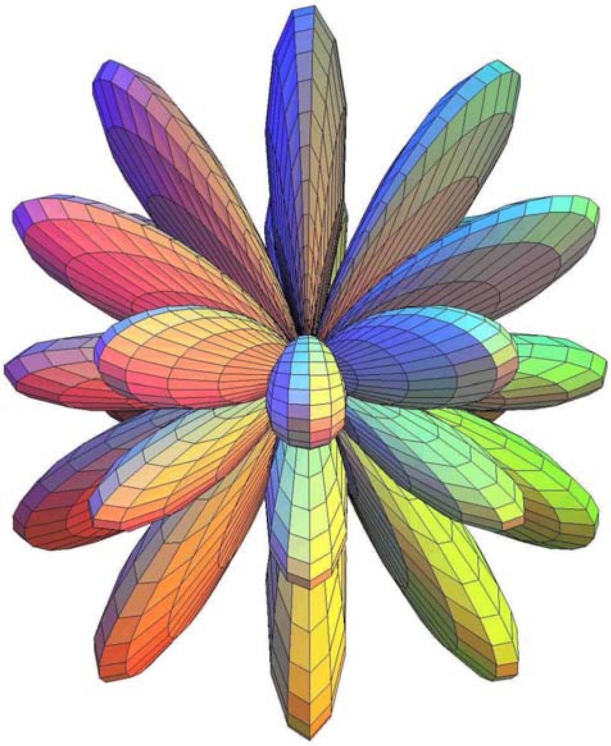

Figure 2.7: Sixth Order Spherical Harmonic $n = 6 \quad m = 4$

(size $(P+1)^2 \times N$) and invert,

$$\begin{aligned}
\mathbf{S}_s(\omega) &= \begin{bmatrix} \alpha_0^0(\mathbf{x}_s, \boldsymbol{\omega}) & \alpha_1^{-1}(\mathbf{x}_s, \boldsymbol{\omega}) & \alpha_1^0(\mathbf{x}_s, \boldsymbol{\omega}) & \ldots & \alpha_P^P(\mathbf{x}_s, \boldsymbol{\omega}) \end{bmatrix}^T \\
\mathbb{S}(\omega) &= \begin{bmatrix} \mathbf{S}_1 & \mathbf{S}_2 & \mathbf{S}_3 & ... & \mathbf{S}_N \end{bmatrix} \\
\mathbf{U}(\omega) &= \begin{bmatrix} U_1(\omega) \\ U_2(\omega) \\ ... \\ U_N(\omega) \end{bmatrix} = \mathbb{S}(\omega)^T \left(\mathbb{S}(\omega) . \mathbb{S}(\omega)^T \right)^{-1} \begin{bmatrix} B_0^0(\omega) \\ B_1^{-1}(\omega) \\ ... \\ B_P^P(\omega) \end{bmatrix}
\end{aligned} \tag{2.18}$$

where $\mathbf{U}(\omega)$ represents the N speaker signals and $B_n^m(\omega)$ are the $(P+1)^2$ modal signals to be reconstructed over the N speakers. It is important to note that the contribution matrix $\mathbb{S}(\omega)$ is a function of frequency. This equation actually creates a set of time domain convolution filters to produce the speaker signals from the modal signals. The time domain formulation can be written,

$$\begin{aligned}
\mathbb{F}(\omega) &= \mathbb{S}(\omega)^T \left(\mathbb{S}(\omega) . \mathbb{S}(\omega)^T \right)^{-1} \\
f_{n,s}^m(t) &= IFFT \left\{ S_{n,s}^m(\omega) \right\} \\
u_s(t) &= \sum_{n=0}^{P} \sum_{m=-n}^{n} f_{n,s}^m(t) \otimes B_n^m(t)
\end{aligned} \tag{2.19}$$

Thus in comparison to the Ambisonics formulation, this approach creates frequency dependent speaker gains (filters) and also takes into account the distance of the speaker from the origin – the Ambisonic decode equation (2.6) assumes an infinite speaker distance. These points are discussed further in section 2.5.

2.4 Wavefield Synthesis

The basic premise of a wavefield system stems from the Kirchoff-Helmholtz integral [174], alternately known as Huygen's principle [168]. An arbitrary sound filed within a closed volume can be generated with a distribution of monopole and dipole sources on the surface of this volume [21, 174] as shown in Figure 2.8. This can be written explicitly

$$p\left(\mathbf{x},\omega\right) = \frac{1}{4\pi}\iint_S \left[\begin{array}{c} p\left(\mathbf{y},\omega\right) \frac{\partial}{\partial \mathbf{n}} \left(\frac{e^{-ik\|\mathbf{x}-\mathbf{y}\|}}{\|\mathbf{x}-\mathbf{y}\|} \right) - \\ \frac{\partial p(\mathbf{y},\omega)}{\partial \mathbf{n}} \left(\frac{e^{-ik\|\mathbf{y}-\mathbf{y}\|}}{\|\mathbf{x}-\mathbf{y}\|} \right) \end{array} \right] ds_{\mathbf{y}} \tag{2.20}$$

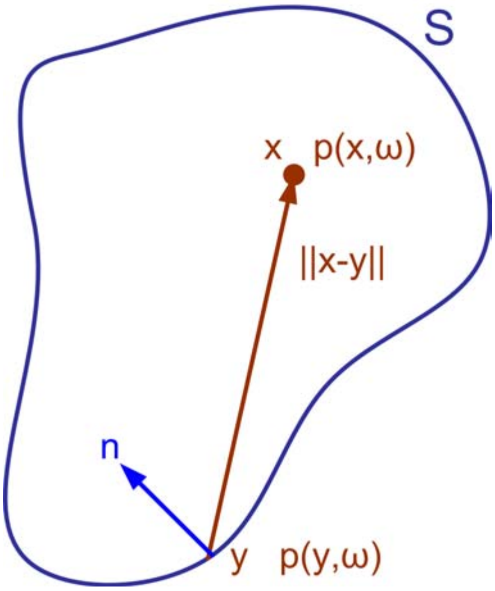

Figure 2.8: Geometry for the Kirchoff-Helmholtz Integral Equation (2.20)

It is apparent that the measurement of the actual sound pressure, $p\left(\mathbf{y},\omega\right)$, and the sound pressure gradient, $\partial p\left(\mathbf{y},\omega\right)/\partial\mathbf{n}$ on the surface of the region will capture the soundfield. This approach is also known as Holophony or acoustical Holography [58].

When a plane surface is considered as the bounding surface for a half space, it can be shown that only measurement of the normal pressure gradient (particle velocity) and monopoles are required for recording and reconstruction [174]. This is known as the first Rayleigh integral,

$$p\left(\mathbf{x},\omega\right) = \frac{1}{4\pi}\iint_S \left[-\frac{\partial p\left(\mathbf{y},\omega\right)}{\partial \mathbf{n}} \left(\frac{e^{-ik\|\mathbf{x}-\mathbf{y}\|}}{\|\mathbf{x}-\mathbf{y}\|} \right) \right] ds_{\mathbf{y}} \tag{2.21}$$

where S is an infinite plane. Finite and discretized approximations of this work well to record and reproduce wave fronts, and a line array can record and reproduce wavefronts in the horizontal plane [21]. This is the basis of Wavefield Synthesis, also known as Holophonics [41]. Wavefield synthesis is suited to large linear arrays of transducers for capturing and reproducing wave fronts.

The nature of this approach is fundamentally restricted to the geometry of the receiver and transducer arrays, and provides for fairly trivial transformations between recorded and reproduction audio signals when the recording and reconstruction arrays have similar geometry [3, 40, 41]. Some recent work has demonstrated the ability to convert the array signals into signals in a "wave space" where they can be transformed for reconstruction by a more generic transducer array [175]. The approach generally uses a sufficiently closely packed array to avoid spatial sampling, however, the recording of all of the signals from the transducer array is not necessarily a compact means of representing the sound field. Wave field synthesis generates fairly accurate large scale reconstruction of low frequency wave fronts, but requires a large number of transducers. Recent practical systems employ around 160 transducers for reconstruction in a horizontal plane [21].

2.5 Soundfield Discussion

The previous sections presented four frameworks for representing soundfields. This section compares these representations in terms of their practical ability to represent a soundfield. Further discussion on practical attempts at soundfield reconstruction is covered in Chapter 3. Here we look at the fundamental information issues relating to the different frameworks for soundfield representation.

2.5.1 Representation Equivalence

Since the representations reflect the same physical entity – a soundfield – there is a degree of equivalence between the representations. To begin with, consider the similarities between Ambisonics and the Taylor series representation. Both frameworks represent a soundfield based on information about the actual sound field at a single point[2]. Both have a hierachical nature, with the Taylor series having $\frac{1}{6}\left(N^3+6N^2+11N+6\right)$ terms for representation to order N and Ambisonics having $\left(N+1\right)^2$ terms for representation to order N. The Taylor

[2] Note that it is not infact possible to measure the required derivatives and directions at a 0 dimensional point, however, the theories are based on infinitesimal differentiation of a continuous pressure field. In practice, a volume is required to measure an approximation of the point derivatives. This issue is raised again further on in this section.

series approximation makes no assumptions about the soundfield being represented other than that it is continuous and sufficiently differentiable. Ambisonics, through the use of directionality, assumes the set of plane waves as a suitable basis function for representing a sound field. A plane wave is an entire solution to the Helmholtz equation (2.8), thus the Ambisonics representation includes the constraint of the wave equation. Hence it is a more compact representation of possible sound fields. It can be shown that the approximation orders are equivalent for Taylor series expansion and Ambisonics and a transformation that allows a mapping is obtained in Appendix A.

The Ambisonic description techniques deal with optimal directionality patterns for the detection and representation of sound fields [78]. The directionality pattern of the Ambisonic polynomials and the Spherical harmonic angular dependence in the modal expansion have a similar structure. In Appendix B it is shown that they are in fact equivalent with an expression derived to generate the Ambisonic polynomials from the Spherical Harmonics. This then leads to a relationship between the Ambisonic representation of a sound field and the modal expansion. Where the Ambisonics definition does not include any reference to the measurement volume, it is shown through the relationship to the modal expansion that this is in fact inherent. Hence, Ambisonics and modal analysis are equivalent.

Wavefield Synthesis and Holophony have been shown to be closely related to Ambisonics [58]. Higher orders of Ambisonics are asymptotically holographic, providing correct sound field reconstruction over a volume [60]. Wavefield synthesis techniques are appropriate with large numbers of transducers recreating a sound field over a larger volume with the speakers close to the listening area (in the near field).

Figure 2.9 shows the equivalence mappings that have been obtained. All representations contain an equivalent amount of information.

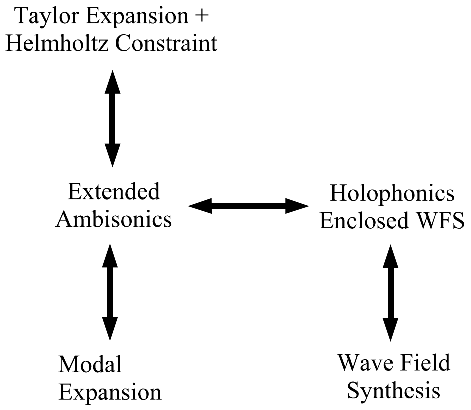

Figure 2.9: Equivalence of Soundfield Representations

2.5.2 Nature of Representations

As Holophonics or Wavefield Synthesis is more suited to a specific geometry, consider the three representations on the left hand side of Figure 2.9 – Taylor Series, Ambisonics and modal expansion. It is evident that all three representations relate to the description of a sound field with respect to some measurement point:

- The Taylor series expansion is constructed from the derivatives of the sound field calculated at a single point.
- Ambisonic representation involves the concept of the direction of arrival of plane sound waves. A direction of arrival can only be defined relative to a single measurement point.
- The modal expansion has a natural coordinate origin around which the basis functions show a strong symmetry.

Neither the direction of arrival of a plane wave, or the derivatives of a soundfield can be measured or calculated from information at a single point. While modal expansion has an explicit dependence on radial or volumetric properties of the soundfield, the Taylor Series expansion and Ambisonics have a similar dependence, however it is implicit. As all three representations contain a similar amount of information, it follows that modal analysis is the most comprehensive and physically sensible framework to adopt. As will be seen in Section 2.5.3 it is a useful framework for analysing the spatial dependence of measurement and reconstruction. It was previously shown in Section 2.3 that the modal analysis reconstruction took into account the distance of the speakers from the listener. modal analysis is also useful for projecting the soundfield behaviour outside of the initial measurement volume [31].

The Taylor series expansion is not naturally efficient in representing the soundfield and, since the derivative of the sound field is taken at a point, is overly sensitive to high frequencies as shown in Appendix A. Ambisonics is simple in theory but the assumption of frequency independent directional microphones, or broad-band beam formers, is not a practical reality [169–171].

In measuring and reproducing sound fields there will always be a trade off. Information theory dictates a trade-off between the number of channels, the region covered, the accuracy and the bandwidth of any system. Modal analysis has been adopted as a useful framework for further work in this area.

2.5.3 Spatial Dependence of Ambisonics

This section looks specifically at the radial dependence of Ambisonics. Despite the belief in some parts of the audio community, it has been shown that Ambisonics does in fact have a "sweet spot" [50]. This follows naturally from the reconstruction equations shown in Section 2.3. The belief that Ambisonics works over much larger volumes than mathematically predicted comes from confusion between the perceived performance of an Ambisonic system and the numerical performance. The issue of perception of spatial sound system performance is taken further in Chapter 4 of this thesis.

Ambisonics also has an implicit measurement volume requirement. For higher order Ambisonics, a finite measurement region, rather than a point microphone is essential. This is shown by considering the Ambisonic signals in relation to the modal structure of the wave equation solutions (B.11). This was derived in Appendix B and it is repeated here for easy reference,

$$A_n^{2m-1}(\omega) = \frac{\sqrt{2\pi}\,(-1)^m\,(i)^n}{R^3} \iiint p(\mathbf{x})\, j_n(k\,\|\mathbf{x}\|)\left(Y_n^m(\widehat{\mathbf{x}}) + Y_n^{-m}(\widehat{\mathbf{x}})\right) d\mathbf{x} \quad (2.22)$$

Since the only radially dependent term in B.11 is the spherical Bessel function $j_n(k\,\|\mathbf{x}\|)$ the volume over which the integral is required to obtain the Ambisonic terms from the sound field is related to properties of the Bessel function. Consider the following approximation for the Bessel function for small arguments,

$$J_n(x) \sim \tfrac{1}{2^n n!} x^n \quad x \rightarrow 0 \quad (2.23)$$

From this we can set out a similar approximation for the spherical Bessel functions considering the identity $j_n(x) = \frac{1}{\sqrt{x}} J_{n+\frac{1}{2}}(x)$,

$$j_n(x) \sim \tfrac{1}{2^{n+\frac{1}{2}}\Gamma(n+\frac{3}{2})!} x^{n-\frac{1}{2}} \quad x \rightarrow 0 \quad (2.24)$$

For small volumes, the higher order functions are constrained to zero about the origin. Several orders of the spherical Bessel function are shown in Figure 2.10.

This can be shown to restrict the "information" in a finite region of space [176,177] expressed as a trade-off between precision and the number of terms required to model an arbitrary soundfield to that precision. The critical threshold, below which fields can be exponentially well modelled by increasing the number of terms from [177] is,

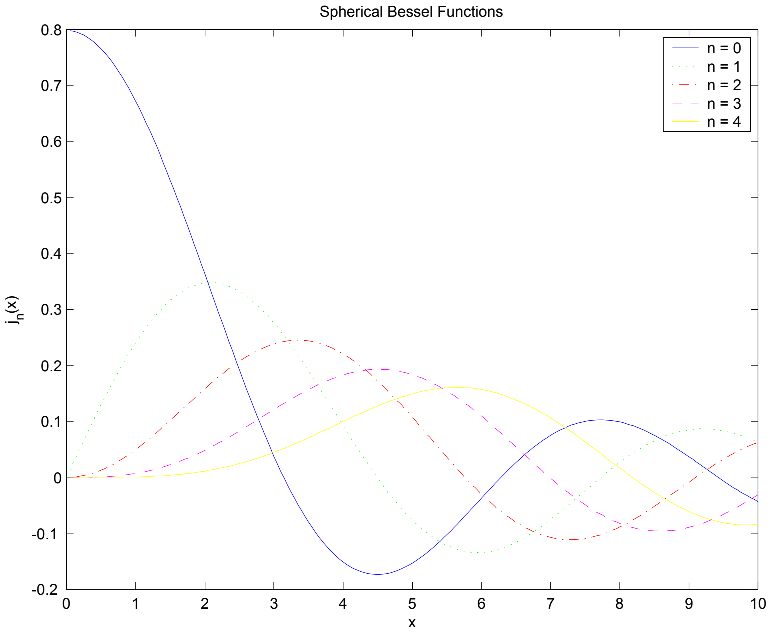

Figure 2.10: Spherical Bessel Functions of Order 0, 1, 2, 3 and 4 near the Origin

$$\begin{aligned} N &\triangleq \lceil e\pi Rk \rceil \\ \text{Terms} &\triangleq \left(\lceil e\pi Rk \rceil - 1\right)^2 \end{aligned} \tag{2.25}$$

where $\lceil\cdot\rceil$ represents the integer ceiling. Now consider that we construct our estimates of $A_n^{2m-1}(\omega)$ by considering measurements of the soundfield in the volume of the measurement device. Assume that the measurement device gives complete information on the soundfield with the exception of a limited precision. It can be seen that for a fixed precision, the number of significant terms varies with the square of the radius of the measurement region. This relationship is frequency dependent with lower frequencies requiring a proportionately larger region. Only infinite precision would give an infinite number of terms over a finite volume of space. A practical Ambisonics measurement device must occupy a volume of space set by the minimum frequency for which it can accurately measure the soundfield.

The Ambisonic decoding equation (2.6) does not depend on the radial distance of the speakers. This is a deficiency of Ambisonics and has been shown to introduce an error in the reconstruction [58]. Although the Ambisonic soundfield representation does not incorporate any measure of distance, only direction, this does not imply that it has a superior spatial performance. In fact Ambisonics is the special case of modal analysis, and the radial dependence terms in modal analysis allow us to fully analyze and understand the spatial performance of an Ambisonics system.

2.6 Chapter Summary

Four different approaches for representing a soundfield were detailed and discussed. It is evident that the most efficient representations, in terms of the number of signals required to represent an arbitrary soundfield, incorporate the constraint of the wave equation (2.7). Both Ambisonics and modal analysis have an angular structure formed from the spherical harmonics, (2.11), which are the natural orthogonal solutions to the separated angular wave equation. Modal analysis also incorporates the radial dependence of the spherical Bessel function, a result of the spatial coupling of the wave equation, and thus provides the most complete representation. Wave field synthesis techniques use the properties of the wave equation to record and reconstruct soundfields using specific transducer arrays. They are useful practical techniques, but do not provide an efficient general soundfield representation. In Chapter 3, the modal analysis representation will be used for analysis and discussion of the theoretical and practical issues for sound field reconstruction.

Since all the representations relate to the same physical object, a degree of equivalence would be expected. This is indeed the case and analytic transformations were derived to demonstrate this. Despite the apparent simplicity of Ambisonics and Taylor series expansion, they still have issues with respect to the region of measurement and reconstruction. No sound field representation provides for measurement at a single point, or reconstruction over an arbitrarily large volume. For higher orders, microphones must be larger, and the reconstruction will have a “sweet spot”.

Chapter 3

Soundfield Reconstruction

The next step beyond a framework for representing a soundfield is soundfield measurement and reconstruction. Suitably small microphone capsules can be positioned to sample the soundfield over a region of space and estimate the actual soundfield present. Given a soundfield representation, it is possible to construct a synthesis equation as seen in Chapter 2. This leads to the use of speakers to reconstruct a soundfield over a region of space.

Given that the desired soundfield is known, just how well can it be reconstructed in free space? How well can accurately can it be delivered to an audience? It is one thing to represent a sound field mathematically, but how practical is it to record and re-create. This chapter will work towards answering these questions.

In Section 3.1 the modal analysis framework is used to simulate and analyze the numerical complexity of soundfield. A requirement for the number of channels and the region over which the sound field reconstruction is accurate is obtained. The issues of practical implementation are addressed which further complicate the problem in Section 3.2.

In the context of the discussion and results, a review of some existing and experimental sound field systems is made to view the likely future performance of soundfield systems. Headphone delivery is discussed as an alternative. Finally the Chapter concludes with remarks about the likely performance bounds of current commercial and future spatial sound systems in terms of numerical soundfield accuracy.

3.1 Numerical Complexity

In Chapter 2 several soundfield representations were considered. Since they were shown to represent equivalent information, the modal analysis framework is selected for further analysis work. Given the representation of a soundfield, the interest lies in the relationship between the bandwidth, the number of signals required and the extent of the volume of reconstruction. To put this in context, we can consider a minimum requirement for a single person to be an accurate construction larger than the volume of the head. To be considered “high fidelity” the system would need to cover a frequency range from around 20Hz to 20kHz. Low frequencies present less of a challenge, but with a wavelength of under 0.04m at 10kHz we can see that our reconstruction sphere already represents more than 1,000 times the volume a single wavelength sphere[1].

Some indication of the required number of channels or the extent of the soundfield reconstruction region can be found in existing literature. Daniel et al [60] suggest that to achieve accurate (-20dB error) three-dimensional sound field reconstruction up to 2kHz over a volume around the size of a single listeners head, requires third order sound field representation and 16 audio channels with around 20 speakers. If only two-dimensional soundfield reconstruction is required, this reduces to around 9 channels [66]. Considering a suitable volume for multiple listeners, wave-field synthesis suggests the use of 200 channels to cover a 2m planar radius up to 2kHz. [58]. Further analysis of the two-dimensional case has been carried out [23] with the number of channels having a linear dependence on the system order rather than a quadratic dependence for the three-dimensional case.

Given that the installation of a 5.1 speaker array is already considered an undesirable intrusion into many domestic environments, the number of channels suggested is prohibitive. With a large number of channels and excessive data rates, practical capture of a complete soundfield is considered unlikely [132]. It is a central theme of this thesis that the level of accuracy suggested by numerical studies may not be necessary.

For the sake of completeness, some results are presented to confirm the review of literature. These results are based on Matlab simulations using the modal analysis framework. Consider the reconstruction of a three-dimensional soundfield using fourth order modal representation (25 signals or terms) and 27 speakers. Figure 3.1 shows some plots of the ideal and reconstructed soundfield for a far-field source, and a rich mixture of sources. While the reconstruction is three-dimensional, the soundfield is displayed for a planar cross-section at $z = 0$ with the color representing the magnitude of the real component of the soundfield. All dimensions are in metres and the speakers were placed at a radius of 2m.

For this and subsequent simulations, the speaker geometry was specified using putatively optimal sphere packings [178] – points placed to maximize the minimum angle between any two points. Note that it is a requirement of the reconstruction that $N_{\text{speakers}} \geq N_{\text{terms}}$.

[1] Based on the volume being the cube of the radius and considering the radius of reconstruction volume is 10 times half of the wavelength at 10kHz.

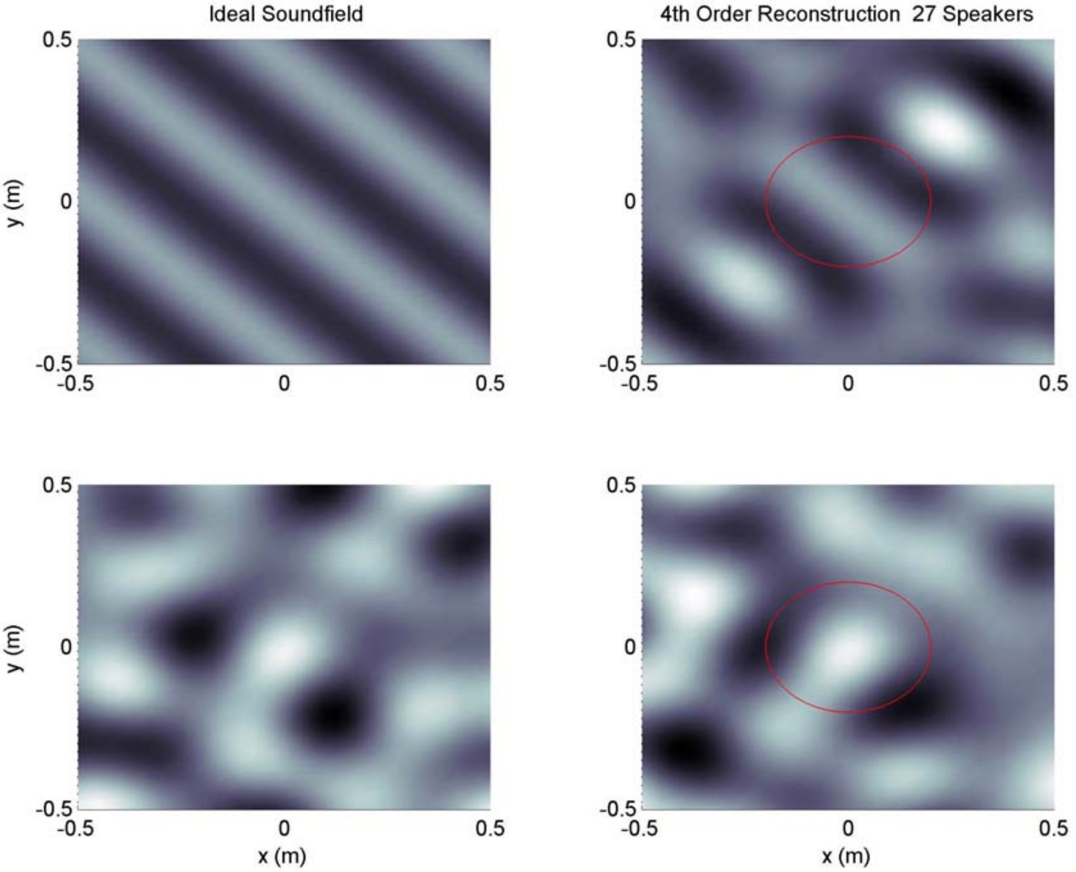

Figure 3.1: Comparison of Fourth Order 3D Soundfield Reconstruction with 27 Speakers – 1kHz Plane Wave and Complex Soundfield.

Since there is not always an exact solution for equally spacing points on a sphere [178], $N_{\text{speakers}} = N_{\text{terms}}$ can give an ill-conditioned inverse speaker matrix, particularly where the geometry of the speakers is not completely regular. An additional two speakers are used to eliminate this.

An error in the reconstruction can be calculated from the ratio of the energy in the difference of the actual and reconstructed soundfields to the energy in the actual soundfield over a particular volume,

$$\text{Error} = \frac{\iiint \left| p_{\text{actual}}\left(\mathbf{x}, \boldsymbol{\omega}\right) - p_{\text{reconstructed}}\left(\mathbf{x}, \boldsymbol{\omega}\right) \right|^2 d\mathbf{x}}{\iiint \left| p_{\text{actual}}\left(\mathbf{x}, \boldsymbol{\omega}\right) \right|^2 d\mathbf{x}} \tag{3.1}$$

Given a radius R we integrate over the sphere of volume $\frac{4}{3}\pi R^3$, however this factor is eliminated by the ratio. This error ratio can be approximated by numerical analysis with sufficient points distributed in the sphere of integration. For the simulations presented in subsequent plots, the errors are calculated using around 500 points to approximate the integral. Errors are plotted in dB where $\text{Error}_{dB} = 10 * \log_{10}(\text{Error})$.

Figure 3.1 also shows the radius in the reconstruction for which the error in the reconstructed soundfield is -20dB of the actual signal. This is an indication of where the reconstruction becomes decoherent. For the simulation shown this radius is just under 0.20m. In this region the soundfield is well modelled by the reconstruction. This results above are derived for an ideal numerical simulation – in practice this would be difficult to achieve for practical reasons.

Figure 3.2 shows a series volume error plots (in dB) against the radius or the region over which the soundfield is compared. These plots are shown for several combinations of parameters of the reconstruction. From this we can see that the figure from [60] of 20 channels or third-order reconstruction at 2kHz over the region of the head (0.08cm radius) having -20dB error is consistent with these results (-20dB at .15m for 1kHz with 18 channels). Note the sensitivity of the error to frequency. The volume of effective reconstruction scales inversely with the frequency. A significant improvement in the size of the region is observed when the order of representation is increased, however increasing the number of speakers beyond the necessary minimum does not make a large difference.

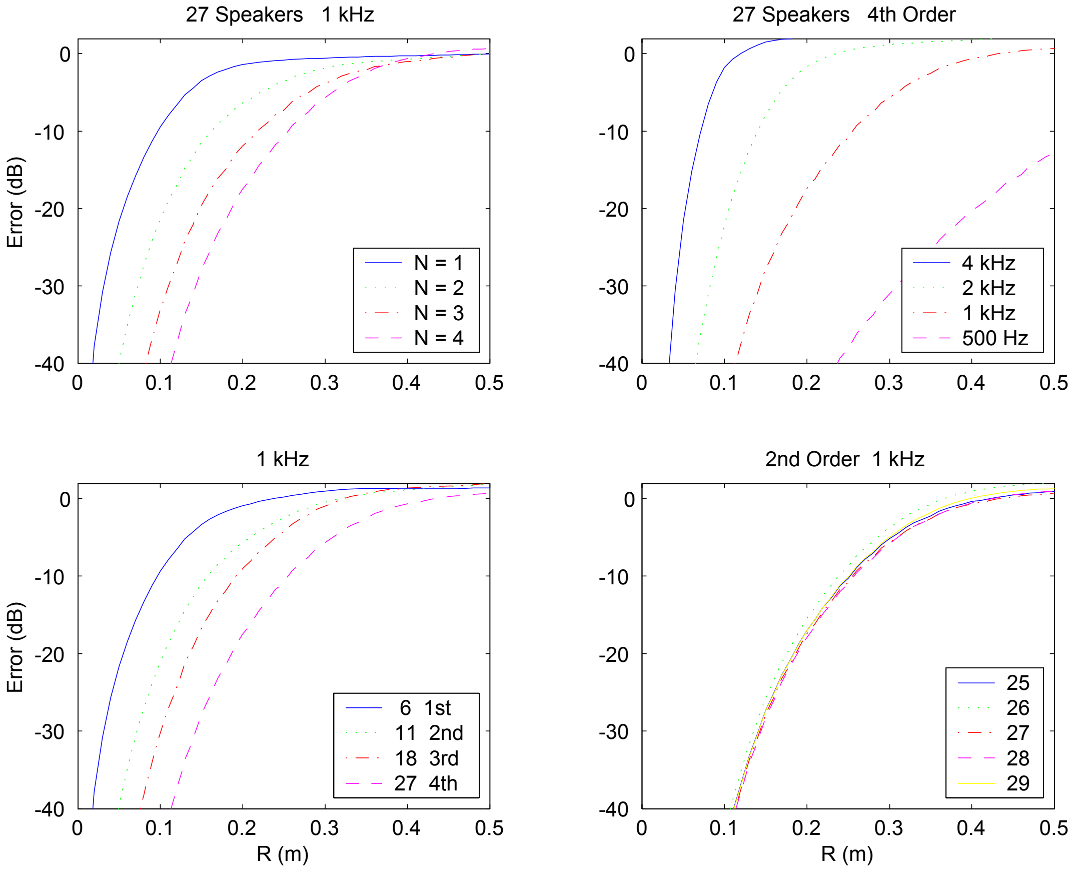

Figure 3.2: Analysis of Volume Error vs Radius for Soundfield Reconstruction – Comparison of the Effects of Order, Frequency and Number of Speakers.

We can consider the order of the modal representation as being the dominant factor in the accuracy of the reconstruction. From the plots in Figure 3.2 we can obtain a rough approximation for the size of the region of reasonably accurate reconstruction,

$$R_{-20\text{dB}} = \frac{50N}{f} \tag{3.2}$$

where R being the radius, N being the order of the representation and f being the frequency. Note that the number of channels of audio required will be $(N+1)^2$ and the number of speakers should be at least $(N+1)^2$. Compare this with the a similar equation obtained from the soundfield richness result [177] shown previously (2.25). This starts with the order of representation required to model a soundfield over a region of space and can be re-arranged to compare with (3.2),

$$\begin{aligned} N &\triangleq \lceil e\pi Rk \rceil \\ N &> e\pi Rk \\ R &< \frac{N}{e\pi k} = \frac{N.c}{2e\pi^2 f} = \frac{6N}{f} \end{aligned} \tag{3.3}$$

Note that the form is the same, although the equation from [177] predicts a much tighter bound. In practice the region of reasonable soundfield reconstruction will be larger than this. 20-30 channels can be considered a reasonable amount of audio data and channels to configure in an experimental environment. This would theoretically give a head-sized sphere of reconstruction up to 2kHz. In practice, the performance will fall below this due to experimental error.

Another issue to be considered is the nature of the filters that must be applied to the modal signals from the modal reconstruction equation (2.18). Considering that the relationship derived provided a frequency dependent relationship, for a practical real-time implementation we want to be able to implement this with reasonably low latency filters. This will be possible provided that the frequency dependence is sufficiently smooth, as the time domain representation will be time limited or reasonably concentrated in a finite duration. The average frequency dependent filters for the different modes are shown in Figure 3.3. The individual speaker gains vary in both magnitude and phase against the individual components for each mode and around the speaker array, however, the frequency dependent gain is similar for all gain coefficients of the same order. It can be seen that the frequency effect is to reduce the dominance of the higher order modes at lower frequencies. These frequency responses can be well modelled by reasonably short FIR filters or perhaps a suitable high-pass IIR filter with an order corresponding to the modal order and a cutoff frequency around 100Hz.

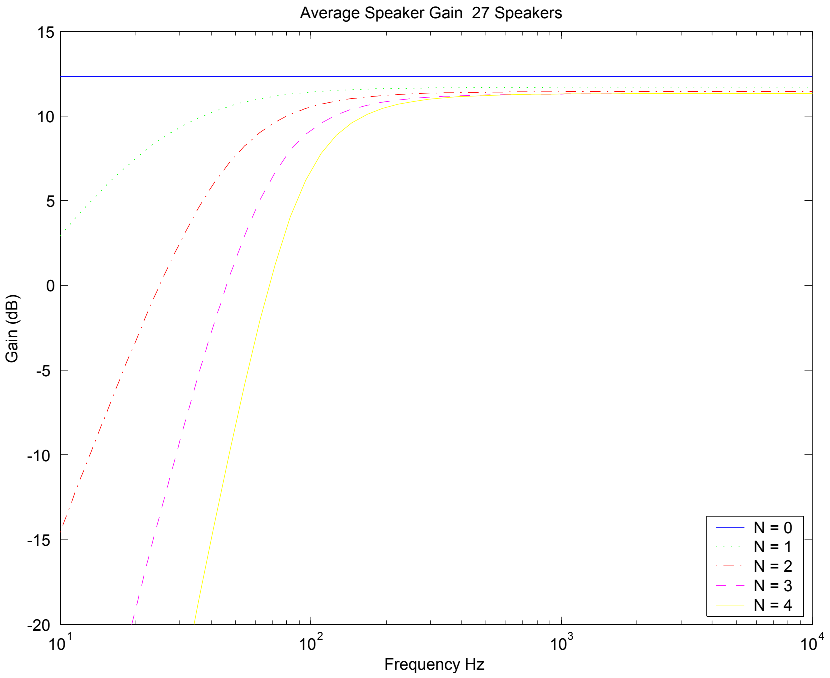

Figure 3.3: Average Speaker Gains vs Frequency for the Different Soundfield Reconstruction Orders

3.2 Practical Considerations

The use of real transducers and complex acoustical environments makes the problem of soundfield reconstruction extremely difficult.

- Placement of the speakers at the appropriate locations is not an easy task. Assuming that there is room for the full spherical array, it is then a challenge to accurately position the speakers.
- Actual speakers are less than ideal with frequency responses that can vary over direction, time and equivalent drivers. Ideally speakers should be small and have a single driver thus best approximating a point source, however there is a trade-off between the size and the general frequency response.
- The acoustics of the playback environment creates problems. Reverberant characteristics of rooms are complex and not easily inverted [125, 179–182]. Even where the actual room response can be measured, above 1kHz it will become very sensitive to any movement of objects in the room and also the changing room temperature and pressure.

Each one of these practical considerations will now be reviewed in isolation to determine the effect it has on the soundfield reconstruction. For consistency, we will use the same configuration as first presented in the simulations of Section 3.1. That is a fourth order, 27 speaker, three dimensional soundfield reconstruction at 1kHz with the speakers at a distance of 2m from the origin. Note that the soundfield plot is for a single far field source while the error plots are calculated using a suitably rich soundfield. Note also that since the random movement of speakers, adjustment of the gains and in particular the room reflections can cause an average gain error (which would not be a significant deterioration in a soundfield), the ideal and effected soundfields are normalized prior to comparison to eliminate this effect.

Firstly consider the placement error. To model this we will consider placement error having a normal distribution about the ideal location. An error with standard deviation as small as .01m starts to have a significant impact on the soundfield. At 5cm error the soundfield section is shown in Figure 3.4. The difference from the ideal reconstruction is noticeable. Error curves as a function of radius are plotted in Figure 3.5 for 0, 1, 2 and 5cm.

Now consider the frequency response error. Given that we are only considering the simulation at a single frequency, we can model this by a random gain adjustment to each speaker. The soundfield plot is shown for the gains having a normal distribution around unity with

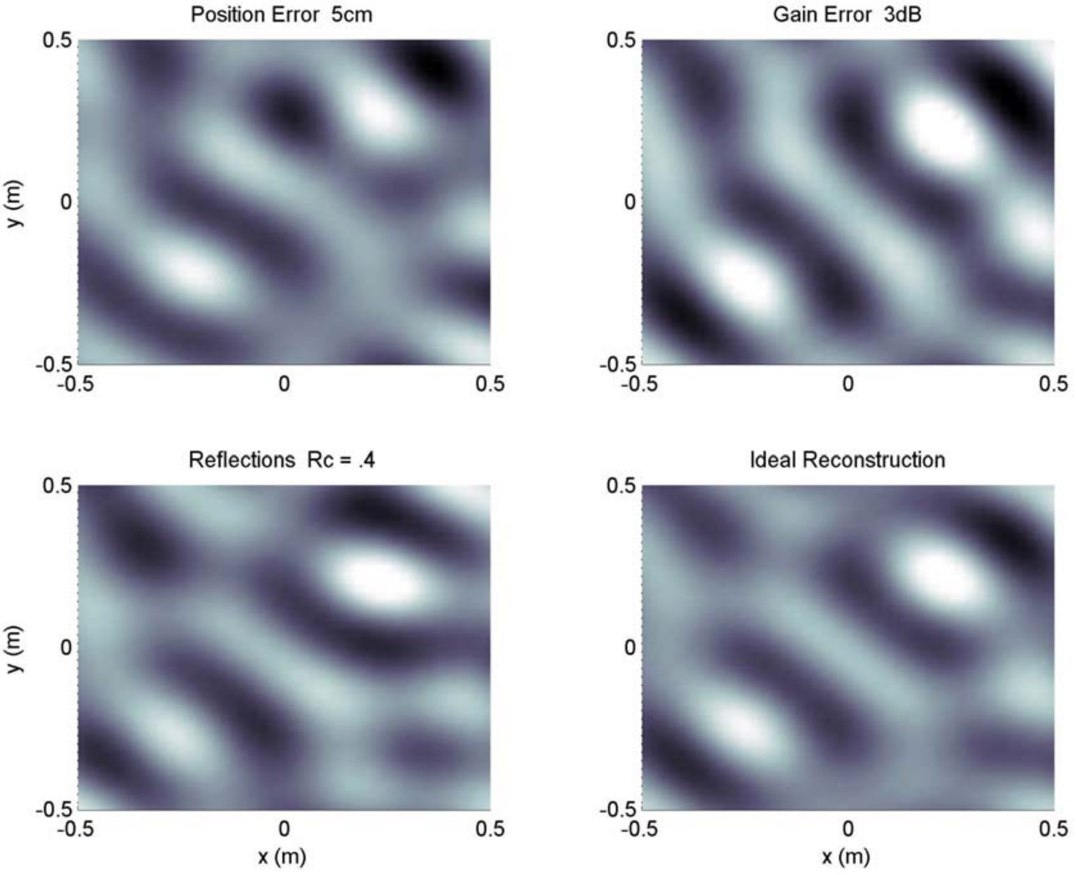

Figure 3.4: Effect of Practical Issues on Reconstruction – Position Error, Gain Error and Wall Reflections.

a 3dB standard deviation. As little as 1dB gain error will start to degrade the soundfield performance.

Finally consider the effect of reflections in the room. With the speaker radius at 2m, assume a shoebox room of dimensions 7m wide 9m deep and 5m tall. Model the first order reflections only using an image source method and adjust the wall reflection coefficients. The 27 speakers become effectively 189 sources. The wall reflection coefficient is varied through 0, 0.1, 0.2 and 0.5 for the error plots.

Interpreting the soundfield plots visually, the positioning error has the most significant effect. A 5cm position error on a 2m sphere corresponds to a 2.5% radial error and a 1.5° angle error. The general effect of any of the above practical issues is to eliminate the central tight region of accuracy and lower the overall performance. One way of describing this is that the “sweet spot” is not so sweet. This means the accuracy of the soundfield is more consistent over a given region, but the performance at any point is not as good as the ideal case.

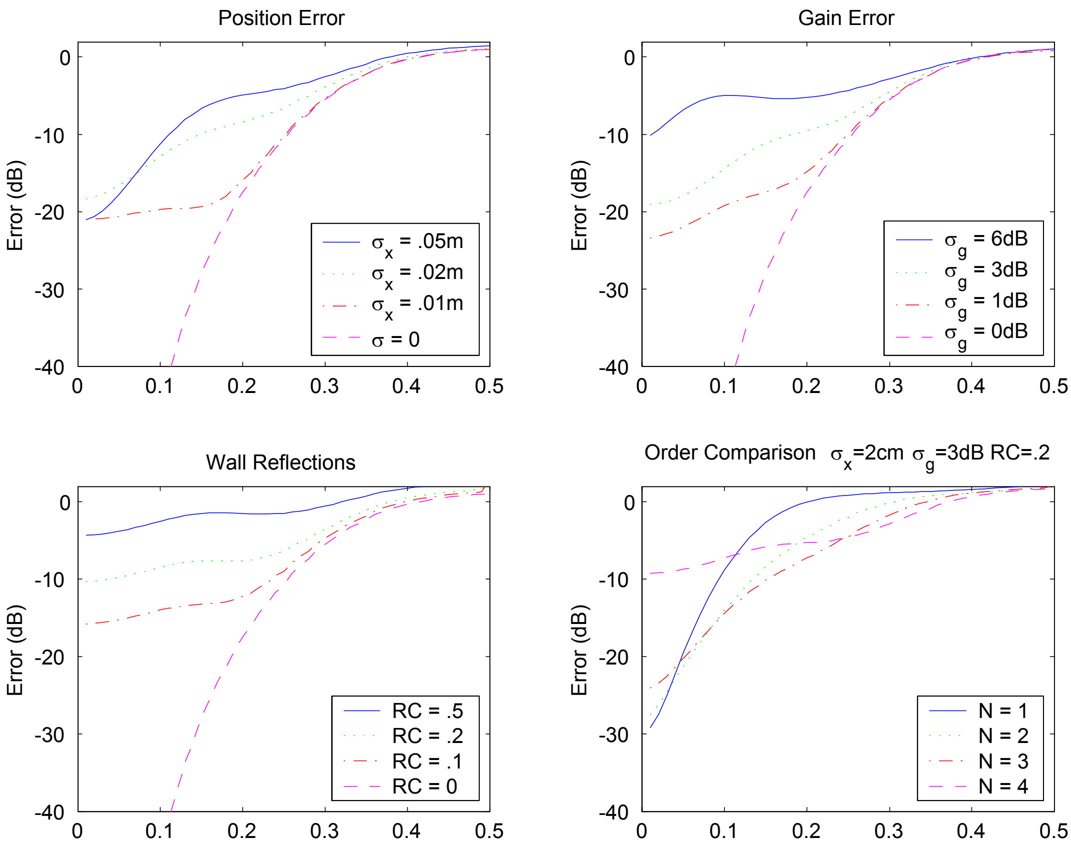

Figure 3.5: Error curves for Practical Soundfield Reconstruction – Effects of Position Error, Gain Error and Wall Reflections.

Note the fourth quarter of Figure 3.5. This figure shows the error curve for the combined practical deviations of $\sigma_x = .02\text{m}$ $\sigma_g = 3\text{dB}$ and reflection coefficient of 0.2. These are reasonable values for a fairly accurately setup soundfield system. The four curves represent different orders for the modal expansion. It is interesting to observe that even with these mild perturbations, the benefits of using a higher order representation are almost lost entirely. In fact, the first order system outperforms the higher order system for a small region around the origin. This can be explained intuitively by considering that in low order reconstruction, the 27 speakers are excited by only four independent signals. The error in placement and gain are thus averaged over a number of speakers and the low resolution information in the soundfield is reasonably well reconstructed.

From the simulations in this section, we could infer that the soundfield would be well constructed when a listener was present. Although this is a slightly different problem – introducing a listener into this region will have an effect on the soundfield – the appropriate soundfield should still be delivered. Ideally the comparison of the soundfield should be between the original soundfield with a head shaped object inserted, and the reconstructed soundfield with that same object in place[2]. This is an area of simulation that is an open problem and discussed further in Section 6.3. Where the speakers are fairly distant to the observation region, this effect will be secondary.

The numerical soundfield error is a measure of the reconstruction performance, however it is not necessarily related to the perceived performance. While errors of phase and amplitude have a significant effect on the numerical soundfield error, they may not always degrade the localization or directionality performance of the system. Another useful measure of system performance comes from reviewing the phase and intensity vectors for the reconstruction of a plane wave. Rather than achieving the exact wave-front, the goal is to maximize the "directionality" of the reconstruction. For a N of sources with gain g_n and position $\mathbf{x}_n$ the phase (velocity) and intensity (energy) vectors ratios [68] can be defined,

$$\begin{aligned} r_V &= \frac{\left\| \sum_{n=1}^{N} g_n \left\| \mathbf{x}_n \right\|^{-1} \widehat{\mathbf{x}_n} \right\|}{\left| \sum_{n=1}^{N} g_n \left\| \mathbf{x}_n \right\|^{-1} \right|} \\ &= \frac{\left\| \sum_{n=1}^{N} g_n \mathbf{x}_n \left\| \mathbf{x}_n \right\|^{-2} \right\|}{\left| \sum_{n=1}^{N} g_n \left\| \mathbf{x}_n \right\|^{-1} \right|} = \frac{\text{vector amplitude sum}}{\text{scalar amplitude sum}} \end{aligned} \tag{3.4}$$

[2] The Author would like to acknowledge Mohan Sondhi for stressing this point at a presentation at given by the Author at Bell Laboratories, NJ.

$$\begin{aligned} r_E &= \frac{\left\| \sum_{n=1}^{N} |g_n|^2 \left\|\mathbf{x}_n\right\|^{-2} \widehat{\mathbf{x}_n} \right\|}{\left| \sum_{n=1}^{N} |g_n|^2 \left\|\mathbf{x}_n\right\|^{-2} \right|} \\ &= \frac{\left\| \sum_{n=1}^{N} |g_n|^2 \mathbf{x}_n \left\|\mathbf{x}_n\right\|^{-3} \right\|}{\left| \sum_{n=1}^{N} |g_n|^2 \left\|\mathbf{x}_n\right\|^{-2} \right|} = \frac{\text{vector energy sum}}{\text{scalar energy sum}} \end{aligned} \tag{3.5}$$

A single point source will give $r_V = r_E = 1$. Ideally a soundfield system should be able to reconstruct a sound from an arbitrary direction with $r_V = r_E = 1$. This is not numerically possible without an infinite number of speakers. The amplitude or phase ratio, r_V accuracy is important for low frequencies (below 700Hz [68]) where the auditory localization is mostly sensitive to interaural phase differences. The intensity or energy ratio, r_E is more relevant to the localization of high frequencies where auditory localization is mostly sensitive to interaural intensity differences [183]. Note that r_V is equivalent to the first order modal or Ambisonic formulation and thus will be solved exactly around the sphere in the ideal modal or Ambisonic system. Extensive work on the analysis of r_V and r_E for planar Ambisonic and 5 channel speaker arrays is presented in the thesis by Bamford [65].

Generally, r_E cannot be solved exactly and a desirable solution would be to obtain a constant r_E against direction [52]. A further simulation exercise would be to model the effect of practical issues on these parameters. This form of analysis is also suited to reviewing the performance of the soundfield system away from the central location. Although the soundfield numerically becomes decoherent, the parameters r_V and r_E provide a measure of how "directional" the system will remain. Further analysis can also review the error in the notional direction,

$$\begin{aligned} D_V &= \frac{\sum_{n=1}^{N} g_n \left\|\mathbf{x}_n\right\|^{-1} \widehat{\mathbf{x}_n}}{\left\| \sum_{n=1}^{N} g_n \left\|\mathbf{x}_n\right\|^{-1} \widehat{\mathbf{x}_n} \right\|} \cdot \widehat{\mathbf{d}} \\ &= \begin{pmatrix} \text{vector amplitude} \\ \text{sum unit vector} \end{pmatrix} \cdot \begin{pmatrix} \text{desired direction} \\ \text{unit vector} \end{pmatrix} \end{aligned} \tag{3.6}$$

$$\begin{aligned} D_E &= \frac{\sum_{n=1}^{N} |g_n|^2 \left\|\mathbf{x}_n\right\|^{-2} \widehat{\mathbf{x}_n}}{\left\| \sum_{n=1}^{N} |g_n|^2 \left\|\mathbf{x}_n\right\|^{-2} \widehat{\mathbf{x}_n} \right\|} \cdot \widehat{\mathbf{d}} \\ &= \begin{pmatrix} \text{vector energy} \\ \text{sum unit vector} \end{pmatrix} \cdot \begin{pmatrix} \text{desired direction} \\ \text{unit vector} \end{pmatrix} \end{aligned} \tag{3.7}$$

Again ideally $D_V = D_E = 1$ for a single direction source. Review of r_V, r_E, D_V and D_E against direction for a soundfield system provides an indication of how "directional" the system is around the sphere and also how correct the reconstructed “direction” is. These models of first and second order are often referred to as localization models, and are based on experimental evidence in human localization experiments [183,184]. Although they are geometric relationships, they are more appropriate for performance assessment than perhaps the numerical soundfield error, as they incorporate an aspect of what is likely to be perceived by the listener. As will be discussed further in this thesis, the subject of localization perception is far more complex and not readily assessed by any numerical means.

Another possible method of soundfield reconstruction performance is to review the recreated inter-aural time differences and inter-aural intensity differences, which are the actual first order perceptual cues. These basic sound field techniques demonstrate reasonable success in recreating the first order cues [67].

3.3 Review of Results

Following on from the results regarding the likely requirements and practical issues with soundfield reconstruction, this section reviews some other results from the literature.

Ambisonics has been used as a method of soundfield reconstruction, largely for recorded and performance art. A well configured Ambisonic system, though only first order, creates an excellent spatial sense though the volume of numerically correct area is quite small (Section 2.5.3 and 3.1). Ambisonics is only truly "holographic" for this small central region [50]. This does not mean that an Ambisonic array will not sound "good" for a large volume – it is just not numerically accurate. Ambisonics has a strong following and is believed to work over large volumes. It is more likely that Ambisonics is perceived to work well since it spreads the energy out over many channels and creates a spatially pleasing sound, though it may only be spatially correct at the array centre. Extensions to Ambisonics with higher orders and two-dimensional windowing functions have produced good perceptual results [50]. The quality of the soundfield reproduced at the array centre was found to significantly improve with correct speaker distance and gain alignment [33] however in practice, an unaligned array was preferred by BT since the soundfield quality was more consistent.

Wave Field Synthesis was initially set out by Berkhout [45] and later refined and implemented in practice by Berkhout and others [3, 29, 40, 41, 47, 175]. With speakers positioned at 10cm spacing it is possible to reproduce wavefronts up to 3.4kHz [40], however the work is largely theoretical and speaker numbers quickly exceed 100, even for small listening areas. Resent experiments are using up to 200 channels for reconstruction [21].

Microphone arrays have been used to measure soundfields using the modal analysis techniques known as near-field holography [31, 185] and using the a spherical microphone array [20]. The modal or spherical harmonic theory can be used for orthogonal representation and reconstruction.

Another area of soundfield reconstuction and representation is the area of multiple control points using a least mean squares formulation. This is essentially a linear algebraic approach to solving the control of a sound field at a number of points in space [37, 39, 42, 186] or on the surface enclosing a volume [44]. This approach still has basic physical limitations which manifest predominantly as the size of the control region and the angular separation of the speakers. The results are similar in scale to that presented previously. One advantage of this technique is the use of microphones at the control points to identify the actual reconstruction impulse responses from the speakers. In this way it automatically adapts or compensates for speaker placements, imperfections and room responses (to the extent that they are invertible [179]).

If the position of the listener is known, the problem can be considered as only reconstructing the sound field at or near each ear. This can be considered for one or more listeners [48]. This is effectively transaural as described in Section 1.3.1. The general approach is one of multi-input multi-output type inversion using measured impulse responses or adaptive filter design to minimize cross talk. This is a reasonably successful technique for a single listener [25] and can be well suited to applications where the position of the listener is fairly well known [57]. Transaural has had some experimental success [22,25,62] but can be overly sensitive to listener position. Some variants of sound field control and binaural recording have demonstrated some success [56, 145].

3.4 Headphone Delivery

Recreating a complete sound field may seem excessive when in the end, the perception of sound is carried out through two small regions relative to the listener – the ears. Provided the transfer functions from a source to the ears, and also from a set of headphones to the ears are known, it is possible to control the sound at the ears with a high numerical accuracy [149]. Headphones are considered an "optimal" method for the reproduction of 3-D sound [15] and show significantly better localization performance than small speaker arrays [8] for a similar complexity. With less uncertainty in the physical path from transducer to the listener, practically, headphones are far more tractable.

The assumption could be made that if the acoustic waveforms at a listener's eardrums were the same under headphones as in free field, then the listener's experience would also be the same. Under specific conditions, experimental work has shown that this assumption is reasonable [101][3]. Several studies have shown that headphones can produce localization results equivalent to that observed with free field sound sources [17, 133, 140], with subjects unable to discriminate real sources from virtual sources [147].

The Head Related Transfer Function (HRTF) is a linear model of the propagation of sound from a point in space to the listeners eardrums. It is suggested that the HRTF is sufficient to explain all of the phenomena observed for sound localization [73], however bone-conduction and acoustic skin pressure can have a undoubtable effect for certain sounds [183]. HRTFs are fairly complex in their nature and position dependence [38] many methods have been proposed for efficient decomposition and synthesis of HRTFs to generate spatial sound [32, 43, 55, 64, 142]. Real-time synthesis and application of HRTFs is feasible with current desktop computer technology.

[3] Just how reasonable is the topic of discussion in Chapter 4.

Head related transfer functions vary from person to person. There is some debate about the need for individualized HRTFs when using headphone delivery of 3-D audio. Azimuthal position is the easiest to simulated with generic filters [138]. Elevation localization tends to improve with individualized HRTFs [140] along with a reduction in the confusion of forward and rearward sounds [133, 187]. Other studies indicate that little is gained from individualization [188]. Head tracking has been shown to reduce the dependence on individualized HRTFs [137, 144] by significantly reducing elevation error and front-back confusion. Distance accuracy and externalization is relatively insensitive to individualized HRTFs [136]. Some evidence relates the loss of pinna cues to increased front back reversals [189]. In general headphone experiments, the extent to which the presence of the headphones does not have a perceptual effect is not always documented. The degree of acoustic transparency and tactile intrusion of the headphones used would have an experimental effect that is difficult to account for.

It is important to note though that free field localization is not perfect, with 6% or so front back image reversals for speech, and up to 20% reversals for band limited noise [17, 56, 187]. Although performance of headphones is typically worse than the free field, it is not significantly so with front-back confusion around twice as likely with headphones as in the free field [17].

Where possible individualized HRTFs should be used, however many applications require a generalized HRTF. In general, many listeners can obtain useful static directional information from an auditory display, particularly for the horizontal dimension, without requiring the use of individually tailored HRTFs [187]. Recent evidence suggests that the listener will even learn to hear through an alternate set of ears much the same way as we learn a second language [143]. This issue is discussed in Chapter 4 on the psychology of hearing.

Although headphones present less problems than non-ideal speakers in an echoic acoustic environment (see Section 3.2), they are not ideal. The matching of transducer to the ear canal varies across subjects. Spectral peak differences of up to 35dB from the "normal ear" have been shown [27] and variances of more than 10dB occur above 6kHz for most subjects [34]. Ideally the individual headphone related transfer function should be taken into account in any headphone simulation. Even with accurately characterized headphones, the response will still vary with headphone positioning [24]. The process of HRTF and headphone response measurement for customization is also very sensitive to the consistent placement of the microphone within the listener's ear [101], especially above 5kHz. Sealed ear canal responses can be used as an alternative [36]. Headphones also create an acoustic loading on the ear, which has associated perceptual effects [26]. For best results headphones should be selected to minimize variance and intrusion to the subject.

If the listener is not constrained, the use of headphones to render a sound field adds the complexity of measuring and compensating for the orientation of the listener's head. In sound field reconstruction techniques this is not required. This adds to system complexity and is difficult to achieve at low latencies. Extreme latency in the update of the binaural rendering can lead to perceived motion of the sound sources dependent on the listener's head movements. However, studies have shown that this effect is not too significant for latencies under 100ms [137] which is achievable in practice [190].

The Binaural approach to sound reproduction need not rely on headphones. Cross talk cancelling filters based on the listener's head position, relative to the speakers, can be used to deliver a binaural signal from speakers. This technique is very effective for virtual sound sources presented 60 degrees either side of the forward direction of the listener [145]. Sound sources to the side or behind tend to be erroneously located either forward or mirrored from their intended location. Delivery is very sensitive to the listener's head location [63] and speaker layout [28, 59].

3.5 Chapter Summary

We have seen that soundfield reconstruction, over a suitable volume for a single listener, requires a large number of channels. The empirical relationship developed is $R_{-20\text{dB}} = 50N/f$ or expressed as the number of channels $N_{\text{channels}} = \left(\frac{1}{50}R_{-20\text{dB}}f + 1\right)^2$. This provides a 10cm radius at 2kHz given 25 channels. Practical issues create complications, however this goal should still be attainable. In practice, the reconstructed soundfield may be perceived as being plausibly correct over a much larger volume, however the numerically correct volume is quite small. This is not suitable for high fidelity audio with such a small bandwidth, and the required size of the soundfield array (2m radius) is a lot to justify for a single listener. If only a planar soundfield is considered the listening region can be expanded but is still not very large (2m radius with 200 channels [58]).

Considering the common setup of a 5 channel (5.1) speaker array, the true soundfield reconstruction theoretically is very small (2cm at 2kHz) and the error in placement, gains and room response of the average setup would mean that there is no coherent soundfield. Understandably, the design of these systems and the multi-channel mixing techniques currently used are not based on soundfield principles, rather they are based on vector-based intensity panning [52, 53]. Where sound is used with movies, we have learned to interpret more into the soundfield than is physically present. We are never troubled that the dialog does not originate spatially from the same point as an actors face on the screen. The concept of “cartoonification” suggests that it is a very much reduced data description of sound that is important [157] – is a sound near, is it far, is it moving, is it coming from behind or in front, is it dangerous? Because of association and context, the perception of the sound track to a movie can be a rich, appropriate and well localized symphony of spatial sound – however we have shown that nothing even close to a numerically accurate soundfield is created. Observations, such as these demonstrate the power of perception against system numerical performance. This is the theme of the remaining chapters of the thesis.

If the goal is delivering accurate spatial sound, headphones are the natural solution. The numerical problem is far more tractable than significant volume soundfield reconstruction. Still, even with headphones, the pursuit of numerical accuracy may be misdirected effort. Even when everything is correct you will get some strange results. Hearing a real sound source in the free field can sometimes lead to erroneous location. Auditory illusions, ventriloquism and front-back confusion are just some examples of where a perceived location can be incorrect with a real sound source. It would be misguided to expect any more from headphones than from free field listening.

In working to commercialize a headphone spatial sound algorithm, much experimental work was carried out with both individualized and general headphone simulations [167]. It became very apparent that with inappropriate listening conditions, an individualized headphone simulation performed worse than a general headphone simulation in the right conditions. There appeared more benefit in structuring the environment, than in customizing the individuals spatial sound presentation. An accurate simulation of a huge concert hall with distant sound sources simply does not sound right when standing near the wall of a small room. It has also been observed that the perceived quality of a soundfield system can far exceed the theoretical numerical accuracy [190].

What is it that can make the best numerical headphone simulation fail? How can a crude speaker array create the perception of an immersive and compelling soundfield? There are no easy answers, however a review of the literature and experimental work in psychology and auditory perception provides a basis from which we can assess the perceptual uncertainty against the numerical performance. This theme will be explored in Chapters 4, 5 and 6.

Chapter 4

Psychology of Spatial Hearing

As set out in Section 1.4, goal of a spatial sound system is to optimize the **experience** delivered to the listener. The experience should be accurate, appropriate, consistent and convincing. Given that the experience relies on perception, we need to examine the psychology of spatial hearing to determine what influences and controls there are for the perception of spatial sound. This chapter is set out in several sections covering various aspects of the psychology of spatial hearing. The last two sections deal with estimating the general accuracy of spatial hearing, and the implications this has for virtual auditory displays. To begin, an overview of the issues in the field of the psychology of spatial hearing follows.

Sensory perception is complicated, and hearing is no exception. Even considering stimulus of a single sense, the segregation and perception of auditory input is very complex. A good analogy is that made by Darwin and Carlyon in a recent chapter on Auditory Grouping [92],

> “Imagine that you are walking along one of the enclosing arms of a harbor on a calm day. Could you, by looking at the waves entering the harbour, describe the events happening out at sea? ... The computational problem of the auditory system is to interpret this complex waveform as sound producing events. Each event must be assigned the appropriate instantaneous properties such as location, timbre and pitch, and their variation over time tracked to obtain such properties as melodic line, speech articulation or spatial trajectory.”

A simple serial hierachical model of signal processing is not appropriate - the processing leading up to perception is parallel and complex with many interaction between the processing levels [157]. Even just considering the processing of the audio sense in isolation, prior

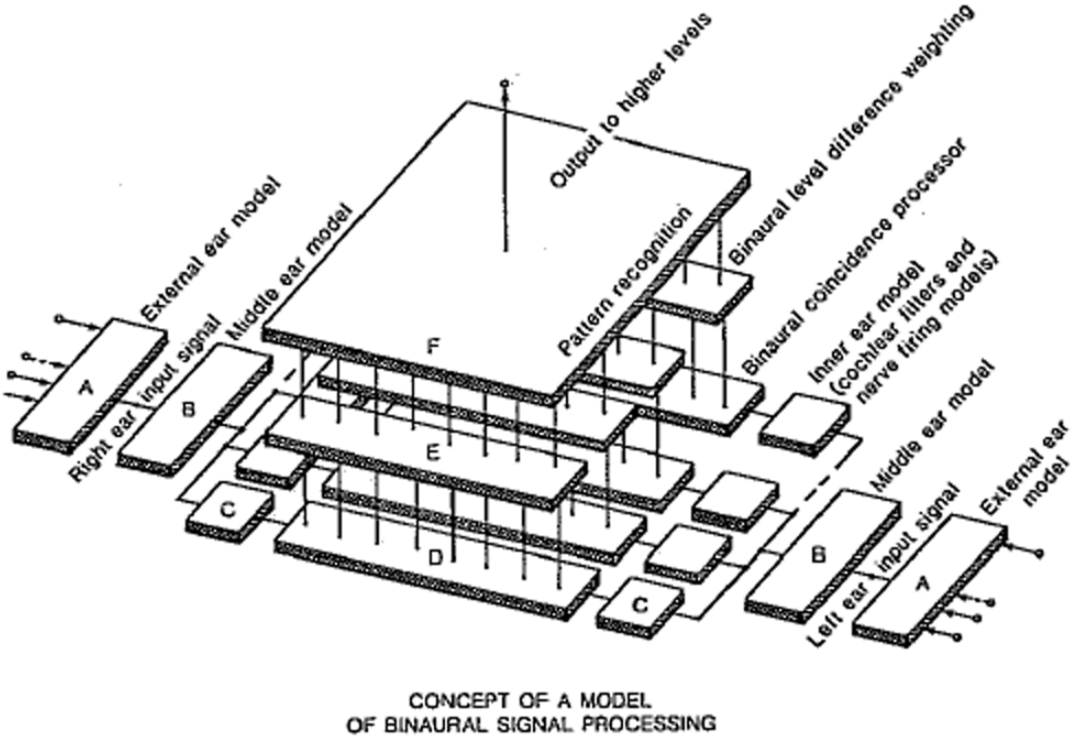

Figure 4.1: Concept Model of a Model of Binaural Signal Processing.

to recognition and classification of a sound, a complex array of signal processing and pattern matching is performed [191]. Figure 4.1 from [191] shows a model for this complexity.

Much of the existing research provides an overview of spatial hearing as observed from Lateralization[1] type of experiments. The dual process model of Lord Rayleigh [184] is still dominant with interaural phase differences dominating low frequency localization and interaural intensity differences dominating high frequency localization. More recent overviews of the psychophysics of auditory localization are available [80, 84, 89, 91, 93, 98, 99, 102]. Most research to date deals with experiments performed with simplified stimulus [99]. Laboratory studies for auditory localization tend to deal with only one or two simultaneous localization cues simultaneously [89]. This is scientific practice to attempt to break something down and deduce rules. However, this forces us into an unusual perception and thus questionable experimental results – stabilizing and simplifying stimulus may complicate and confuse perception [83].

Research shows that the localization cues from auditory events in the every-day environment

[1] Lateralization refers to using headphones or sound sources close to the ears to control the phase and intensity differences of simple sound sources. It is different to localization in that the sound sources are typically perceived within the head – they are not localized in external space. These experiments simplify the study of hearing mechanisms by eliminating the effects of reflections and reverberance.

are often ambiguous or in conflict [80]. Binaural localization cues are ambiguous [134] and can sometimes only resolve localization to a cone or torus. Despite the errors and complexity of the auditory localization cues, our perception of the acoustical world is very stable [99]. In perceiving this complex auditory environment, our perceptual system performs in a robust manner with the "assumption of a coherent environment in which sounds should come from the places that are occupied by the events that made them". Thus a large amount of auditory streaming, perceptual grouping and interpretation occurs before the perception of sound source and location occurs [96]. This grouping forms from principles of sound sources that relate to the reality around us – such as unrelated sounds seldom start or stop at the same time. Very few studies deal with the complexities of a typical real-world auditory environment and thus are incomplete and could be misleading [192, 193]. Some experimental work has shown that the ability to localize sounds degrades quite significantly with multiple sound sources, particularly if they overlap spectrally [81]. Localization in this environment is quite different from the experiments upon which most of spatial hearing theory is based.

Current work in psycho-acoustics and perceptual coding covers the issues of hearing thresholds and masking [84]. These already are complex models but perform well in systems such as MPEG audio encoding. Experiments in the area of temporal induction [83, 194], phenomic restoration [195] and linguistic organization [196] show that the perception is sensitive to context, meaning, memory and structure [113]. Certain aspects of audio signals can be "heard" even when not present – we can be convinced of hearing more than the actual sound when expected signals are not present or obscured by noise. It is evident that there is a level of association and perceptual effects beyond masking and thresholds – familiar sounds such as our own names can be perceived in pure noise [96]. The auditory system tries to fit any stimulus to a causative perception – the link from sound to perception is not direct and not always causal. Echoes are not individually perceived [90], rather they have a complex effect on audio perception [102]. Despite a complex set of sound reflections a single sound image is perceived at a well defined distance with relatively constant timbre. The echoes reveal information about the source distance and acoustic space. Again this demonstrates a great degree of sophistication in processing and perception of spatial audio. Some success has been made with computational models that demonstrate single source localization performance, similar to human subjects [197, 198]. This is, however, a much simpler task compared to spatial perception in a complex audio field combined with other stimulus.

Given the complexity of appropriate perceptual models, there is further evidence that they are not static. There is debate over what extent of localization is present at birth and what is learnt from tactile and visual experience feedback [193]. The ability to relearn localization exists well into our adult life – when our ears are artificially modified we can learn to hear again in a relatively short time [127]. Neurological studies have shown that separate neural centres are responsive to sounds in particular locations in space. Further, the neural centre excited by a sound at a given location can be dependent on the direction of eye-gaze [100, 199] and consequently perceived location is also effected by gaze direction [88, 107, 200]. The eye position and sound processing interact at a very early neural level.

Hearing is an important sense for communication. Psychology research on communication has shown the complexity and interrelationship of the models for human communication. It is not possible to consider hearing and communication as a hierachical of independent processing stages – detection, characterization, identification and perception. From the preface of the book, Human Communication: A Unified View [201],

> "Human communication is, therefore, very much a matter of interrelated events on numerous levels of activity. In fact, the effect of one level on others is profound, not secondary as was once thought."

The text "Auditory Perception" by Richard M. Warren [83] is a good summary of research into the nature of auditory perception. It describes many experiments and auditory illusions that demonstrate the potentially large differences between the actual physical stimulus and what we perceive. This is a continuing and very challenging area of research. Some general rules for perception have been stated [83]: -

1. Sensory input is interpreted in terms of familiar causative agents or events and not in terms of the manner and nature of sensory stimulation.
2. Perceptual changes occur during exposure to an unchanging stimulus pattern.
3. Prior stimulation influences perceptual criteria.

Perceiving sound is more complex than the task of memory-based association [94, 99]. We use many auditory cues to recognize a sound object. These cues may be correlated and redundant, with no predominant cue uniquely determining perception. A listener will make use of whatever cues lead to best performance for a specific set of events [94] and "there is no pressure-variation that will always lead to one and only one perception." An identical stimulus will create different perceptions across different individuals, and even within the same individual at different times [90]. We can focus our attention on a particular sound object [202] in an auditory scene and in effect alter the information we perceive. Listening is an active process, it allows age, experience, expectation and expertise to influence perception [99].

Despite all of this complexity, arguably the perceived soundfield cues represent a small part of the total soundfield. Less information than that of the complete sound field representation is utilized [132]. We extract a description of the sound environment with emphasis on the aspects that are important to us [157].

The remainder of this chapter outlines the important areas of hearing psychology relating to the problem of what information is important and appropriate in a soundfield, for controlling perception. Firstly in Section 4.1we will look at the extent to which spatial sound perception is adaptive and can be biased. In Section 4.2 the effect of conflict and integration of the different senses is reviewed. Section 4.3 looks at the premise of continuity of perception and how this effects perception. These three areas address the major components of the psychology of spatial hearing and lead on to a comparison with the numerical results and conclusions in the following chapters.

4.1 Hearing Adaption

There is an increasing amount of experimental evidence to support the notion that the capacity for recalibrating auditory localization continues well into adult life [127] – hearing and our ability to localize sound sources is an adaptive process. Results suggest that auditory spatial perception is governed by an internal representation of external space that can be re-tuned by sensory experience, even in adult animals [135, 155]. Hearing adaption and associate "memory" occurs on both short-term and long-term scales [203].

At the smallest time interval, we quickly adapt to ignore constant information presented in the form of early echo arrivals after the direct sound [85–87, 90, 97, 204]. This adaption creates an impression of the acoustic space and any sudden changes or new echoes are readily perceived [104]. In a reverberant environment, we learn the properties of the sound source and acoustic space to better estimate location [83, 154] distances [18, 82] and interpret spatial relationships over time [103]. Previous sounds or noises can effect the localization of subsequent sources [105, 108]. Experiments show that we adapt to the re-arrangement of auditory cues [205], and must do so as we grow and our head shape changes [83, 206]. We can recalibrate the way that our internal representation is mapped to auditory cues [128]. Our internal representation of auditory space is not fixed, it is constantly changing with feedback and stimulus to allow us to efficiently perceive our surroundings in a dynamic environment – it is adaptive.

Binaural adaption has been demonstrated for simple stimuli [89] and it has been shown experimentally that general localization ability can be enhanced by practice [193]. Even when the sound delivery is synthetic, localization performance can improve with exposure. Localization in binaural recordings can be facilitated by a short training session [129] and this effect has been shown to last as long as four days between sessions [130]. There is a strong suggestion that we can learn to hear through imperfect binaural reproduction provided that such a reproduction in physically plausible and consistent [207]. Some evidence suggests that the adaption is not always complete [144, 208] and is only able to match a linear remapping of audio cues [135]. This could then predict that alternating between a virtual and real auditory display could dull the accuracy of spatial hearing in both situations. However, results suggest that adaption to pinnae changes are possible and complete with the new "earprint" learnt in a similar way to a second language [141, 143]. This adaption occurs over a fairly long time period (several days) and persists after the exposure, suggesting that a long-term memory is built up of an alternate pinnae interpretation. Classic psychology experiments with psuedophones to rotate [206] or even swap [126] the ears demonstrates the extent to which our hearing can adapt.

The nature of the adaption is not easy isolated. Adaption takes place to form congruity of the senses. Distortion of the visual field has been shown to have an after-effect on auditory localization [118]. Measuring the after-effect displacement of spatial hearing subsequent to exposing a subject to a displaced image and sound source [114], provides some indication of the characteristics of adaption. With the image and audio of a person speaking displaced by 20 degrees, using a small TV monitor and separate speaker, after only 12 minutes of exposure and after effect was observed of around 4 degrees. Interestingly, this is a similar angle to the detection threshold of a discrepancy between visual and sound location [209]. Feedback from the other senses, particularly vision, modifies the way we perceive sound locations. This is not just a bias that exists only when the stimulus are inconsistent – it is an adaption that modifies the process of spatial sound localization and persists after a discrepant exposure. In extreme cases, this can lead to "double images" [126] and has a definite effect in between exposures to alternate HRTFs [206].

At a higher perceptual level, distance perception is also effected by context and association and can be considered to be adaptive, or learnt. Whispers suffer a perceived distance bias towards the listener and shouts away from the listener [99, 121, 210]. The subjective impression of distance seems to dominate distance localization [124]. Distance localization can be very poor with an unfamiliar sound, and increases quickly with experience [211]. We can estimate distance better in an environment we are familiar with and have had previous exposure to [82].

From the compelling experimental evidence, we can concluded that the auditory system is adaptive and recalibrates at many levels. Evidence exists to suggest that this occurs over both short and long time frames, within a single exposure and between separate exposures and from the smallest level to extreme changes in HRTFs. There is little semantic difference between learning to localize through different ears or learning to localize better in a familiar acoustical environment. Spatial hearing will improve over time and compensate for imperfections, provided they are regular and deterministic. It has even been suggested that learnt localization through a suitable alternate earprint can exceed our own localization ability [187]. An observation of the author is that a sound system becomes more enjoyable in a familiar acoustical environment. Given the evidence above that we learn to adapt and essentially isolate more information as we become familiar with an acoustical environment, or listening condition, this observation is not surprising.

4.2 Multi-Modal Sensory Interaction

4.2.1 Sensory Conflict

Spatial cues from other sensory modalities effect sound localization [144] and we do not localize sounds in isolation from out other senses [193]. Integrating multiple sources of information is a natural function of human endeavour [212]. We cannot even intentionally ignore the influence of other senses and alternate cues. Three senses dominate in the perception of spatial localization – vision, proprioception and audition. To the extent that each sense can bias the perception of the other, audition is shown to be the weakest of the localization senses [123].

What happens when visual or proprioception cues are in conflict with the auditory stimulus? Within certain limits of plausibility, the audio perception is usually dominated by the other senses and the sound is perceived as originating from a suitable location, different from that which would otherwise be perceived by the experience of the soundfield in isolation. Although the auditory sense has a faster response time [213], vision has a higher spatial accuracy and tends to be dominant in most situations [212]. The perceived location of the non-dominant sensory mode will tend to shift towards the location from the dominant mode [15]. This is generally known as the ventriloquist effect. Where there is a strong association between the nature and movements of the visual cues, sounds can be displaced up to 30 degrees in azimuth, and even greater angles in elevation [119]. Visual dominance can be so strong that subjects are perceptually unaware of the audio stimulus [116]. For events in 15-25 degrees from the front of the listener, the visual sense tends to be the dominant modality. Tactile senses are particularly dominant where a moving sound source can be located by the listener's hand [112,214]. It is intersensory interaction that drives hearing adaption – Moore suggests a plasticity in the relations between the senses, where a recalibration of auditory space can occur on the basis of information from the other senses [102].

The highest level of spatial representation is drawn from all senses, but visual input has the most power to influence it [102]. Perhaps this is best summarized by Tom Holman in his recent book [12].

> "Vision is obviously also important for localization and can overwhelm aural impression. Sight dominates sound for localization. Nevertheless, mismatches between the position of a sound source visually and aurally do cause cognitive dissonance, which tends toward limiting the suspension of disbelief usually sought."

The extreme interpretation of this is that near enough is good enough for spatial audio where other sensory cues are provided. Immediately this suggests that the pursuit of numerical accuracy in spatial sound delivery is not warranted.

There are some similarities between the early processing that occurs in the senses [152]. Aspects of grouping, repetition, figure-background separation and complex patterns can be characterized in similar ways. However, the differences appear to outweigh the similarities. This is largely due to the way we use particular sources of information [120]. Hearing is very much the continually alert sense that does always require conscious attention direction to be aware of our surroundings.

> "Animals, from the simplest to the human, use their eyes to recognize objects, to distinguish the edible from the poisonous, to distinguish the friend from the foe. Similarly, most animals use their ears for detecting warnings of distant and as yet unrecognized things."

The conflict, bias and association of auditory events and visual or tactile objects is dependent on the plausibility of their association. Unrelated sensory input is not integrated. A good example is given by Marks [113],

> "When stimuli presented to different senses bear no meaningful relation to each other, interaction often seems to be small or non-existent... But meaningfully related stimuli are quite a different matter. The voice of a good ventriloquist sounds displaced in space, away from the ventriloquist's mouth, which does not move, toward the dummy's, which does."

The effect of bias, where senses are in conflict, is bidirectional. Visual location of an object can bias the heard location and the acoustic location can bias the perceived visual location in a simple visual field [110,117]. However, the effect of auditory stimulus on visual perception is of a smaller magnitude.

Studies have shown the importance of visual cues in the shaping of the auditory space map in the brain [156]. The level of interaction in the development of the neural pathways of the brain is responsible for the strong relationship between visual and auditory perception. Wherever spatial audio is present along with other sensory cues, the influence of any conflict in sensory input cannot be ignored. Sensory input is not restricted to simultaneous stimulus with the spatial audio. For example, if a subject has seen the dimensions of a room prior to dimming the lights, this will have a lasting effect on the acceptance of a simulation of an acoustic space very different to that which was visually observed. Rooms do no suddenly transform around us.

4.2.2 Sensory Integration

Where the senses are simultaneously stimulated, but not necessarily in conflict, a large degree of sensory integration occurs to form our perception. This means we can perceive elements that are not necessarily present, but are expected to be there. This interaction cannot be stated simply. In perceiving reality we attempt to predict the current state of the world around us from sensory input of the past. For this to be robust to errors in the sensory detection, identification and localization cues, a large amount of sensory cooperation occurs and our hypothesis of the world state are largely restricted to rules of possibility learnt from experience [80, 96, 113, 157]. An excellent account of intersensory interaction is given by Welch and Warren [213]. Where we are exposed to familiar or typical situations the assumption of unity is strong [112] and related input from the different senses is perceived as a single event. Strong interaction between the senses is a corollary of evolution for survival [83]. When we extrapolate our perception back to what we think we have sensed, there can be significant error and bias. For example, the perceived quality of video material can be improved by improving the quality of the associated audio [15, 215]. In simple experiments, a voiced syllable is heard as a combination of the actual audio and the appropriate audio which would accompany the visual cues. When watching a person speak, what is heard is strongly influenced by what is seen [115] – what we believe we heard may never have transpired. Speech in a noise environment is actually heard louder in the noise and detected better where visual cues are present that are correlated with the articulation of the speech [216]. Simply having a textured visual field improves the accuracy of auditory localization [213]. The organization or grouping of stimulus by one modality can effect the perception of another [217]. Overall there is a strong indication that visual and auditory processing are not independent [153], and share many common neural pathways.

There is experimental evidence that a distortion of one sense, though initially may disrupt the congruence of spatial perception of a single object, can be reduced over time [218]. Late in the 19th century the psychological experiments pioneered by Stratton involving inverting or displacing the visual field of view demonstrate in the commentary that over time, objects are seen and heard as one [219]. The experiment of reversing the ears was mentioned previously in relation to hearing adaption [126]. This experiment, and others involving auditory rearrangement are discussed in more recent works [205]. They suggests a process of reinterpretation of the sensory aspects of objects in respect to their locality significance. This process is typically observed over a time frame of several days. It is not simply that spatial sound localization adapts to match the visual input. Rather, over time, the senses recalibrate to form an integrated and consistent spatial representation of the observed world. Neither sense remains unaffected by the disruption.

Carr suggests that the accuracy of the auditory system requires constant aid from more accurate sensory localization [218]. Without the integration with other senses, the accuracy of spatial audio perception would perhaps degrade.

> “Localization of auditory objects is continually being supplemented by visual and inferential knowledge. Without such aid, auditory localization would be of little practical value.”

Note that this is not in reference to vision alone, otherwise there would be the obvious contradiction of a blind person. However, in all cases it could be argued that position from hearing localization is learnt and calibrated against other senses or against further information gathering activities. The observation of a blind person on localizing distant sound sources is very different from that of a sighted person, indicating the effect of the limited range or proprioception in calibrating spatial sound. Further, blind individuals have far more acute acoustic detection within the reach of their bodies. Experiments have shown the perception of blind people feeling objects with their face as a result of near acoustic object perception.

Simply put by Moore “What we hear is influenced by what we see” [89]. This observation is much stronger than just the fact that a conflict of the senses alters the perception. Generally, observers are unaware of the conflict and are surprised that the perception of the auditory event changes once the visual cue is removed. Experimental work in this area is largely in the case of speech perception. We use all of our senses together to gain one perception of the world around us [113]. Sense perception gives meaning and this is inextricably linked in with organization and memory. With advances in equipment to create synthetic stimulus of the senses, this is an area of research that is growing in interest [107,200]. Unless the goal is for a spatial sound system to operate in the absence of other sensory input – the comfortable seat in the dark analogy – we cannot ignore the effect of intersensory interactions. In fact, from practical experience [167] and a thorough review of existing literature, the perceived performance of a spatial sound system can be more efficiently improved by effecting the other senses.

4.3 Premise of Continuity of Perception

The concept of the effect of past sensory input has been previously mentioned. This is an area worth elaborating on as it is of particular interest in the context of a spatial audio system where for complete immersion there is some "discontinuity of reality" that needs to be overcome. In our familiar experience, acoustical properties do not change suddenly or without reason. The premise of hearing adaption has been discussed in Section 4.1. Where changes are slow, this adaption tracks the changes so that perceptual confusion does not occur. Events such as walking through a door or changing our position create a cognitive expectation for change. Putting on headphones, or starting a simulation creates an event that violates our expected premise of continuity or transition event.

Our modern culture has incorporates some methods to handle this exception. Movies rarely begin with immediate plot content – the title or opening sequence of a movie typically has a recognizable context and places in a state of expecting a transition. We accept the discontinuity of a two dimensional image projection and see through a photo or screen without concern for the discontinuity at the border. Is there such an analogy for audio? Considering this question shows that spatial audio simulation has a unique challenge in that there is little or no precedent for acceptance of the transition. An audiophile, to fully enjoy the experience of a high fidelity audio system will voluntarily welcome this transition – dimming the lights, shutting the eyes, assuming a comfortable position, relaxing and projecting into the soundfield. Can we ask the same of any-one who enters a 3D simulation and expects the audio to be immediately compelling and immersive?

We "filter" our perception based on an assumption of continuity. Anything conflicting this assumption will lead a sense of disbelief or a lower sense of immersion or telepresence. In order to be able to function in a complex environment with broad and conflicting sensory input, perceptual processing must occur to predict the present reality based on previous sensory input. This element of prediction adds significant complexity and is made robust by combining information from all senses [157]. The compellingness of an auditory demonstration is linked to how well the presented audio matches the acoustical environment of where the listener is or was previously [146]. A smooth transition can make an imperfect spatial sound experience seem very realistic.

With headphones, particular attention must be paid to the issue of continuity of acoustical environment. Headphones upset the normal adaption and learning process of acoustics in different rooms [103]. When we enter a new room, the short term memory of echoes and acoustic geometry is reset – this is not necessarily the case when headphones are first worn. Interpreting the acoustical cues presented over headphones with the assumed spatial geometry of the previous (and perhaps still visible) room is bound to create conflict or ambiguity in localization.

The importance of hearing as a pervasive sense for environmental awareness has already been mentioned. It is our only long range panoramic sense [127] and ingrained in our survival [220]. Hearing is largely a defense sense for warnings of unknown [120]. Consider the effect on hearing when our other senses are deprived or impaired – wearing a straight jacket in the dark with a suppressed sense of balance, you would certainly be listening out for any clues to your surroundings. Now consider the situation in a typical VR simulation. Vision is impaired through the use of shutter goggles or a VR headset, visual cues may conflict with balance and proprioceptive cues. Our sense of hearing will operate in a defensive mode. Rather than being led into the VR simulation we know to be artificial, our hearing will endeavour to give some indication of the surroundings of our true physical body. This is a much understated challenge for the deployment of compelling spatial audio.

A very compelling experiment was performed by Gilkey et al [2], which the author has personally experienced. The subject is led into a room, in which there is the host and a set of familiar objects in the room – a table, some scissors, a phone, a blackboard, a newspaper. The host introduces herself and asks the subject to look around and the put on the headphones and shut their eyes. A previous binaural recording made in the exact same room with the same objects and same host is then replayed. The apparent accuracy and realism of the spatial sound has to be experienced to be fully understood. Goose-bumps as the host scratches a fingernail on the board while writing, a flinch as the scissors cut near the ear, a persistent image of the newspaper as it is ruffled and torn in front of the subject. With simple spatial sound technology – a Kemar dummy head and a DAT recorder, no headphone or microphone compensation – the spatial audio presentation is flawless. The striking feature separating this from any other experiment is the extent of the sensory and perceptual continuity.

Many experiments do not consider the perceptual effect of a discontinuity of perception on the experiment. A recent work by Zahorik [18] listeners were presented synthetic binaural signals measured from their own ears in a real space. The test was carried out in a sound proof booth with the listeners being asked to close their eyes only for the audition of the simulated sound. The simulated acoustic environment was that of a much larger echoic auditorium. Listeners consistently underestimated the distance to the source. Being asked to discard the knowledge of the room they are physically present in, ignore the absence of any noise floor appropriate for the room size being simulated and the absence of appropriate self perception is a large implicit perceptual stretch. Given the premise of perceptual continuity, it is not surprising the distances were underestimated. Despite this, very little discussion in the work is made other than the actual acoustical cues being created by the simulation. Though perceptual discontinuity is difficult to avoid in a practical experiment, it seems convenient to discard in any discussions or analysis of the results.

4.4 Chapter Summary

The goal of a spatial sound system is to optimize the **experience** delivered to the listener. Experience is subjective and not readily observable. Therefore it is not possible to address the goal of delivering a specific experience without reference to the psychology of spatial sound perception. This chapter has brought together key research ideas and results in this area from the last century. It is clearly evident that there is a significant amount of bias and error in human spatial sound perception. This cannot be ignored in the design and evaluation of spatial sound systems. The effects of improved numerical accuracy or design comparison may be obscured by psychological influences.

Taken in isolation, spatial hearing is not always accurate or unambiguous. When combined with additional sensory information, a knowledge of our surroundings, expectations and causal interpretation, we generally perceive very accurate spatial location of acoustical events – we are immediately convinced that we heard a sound in the location of the object perceived. Since reality rarely deceives, this creates an exaggerated expectation of accurate spatial sound perception from a synthetic spatial sound system. When spatial sound is presented without causal plausibility it would not be surprising that our perception of spatial audio accuracy is lower.

With the issues of perceptual uncertainly and unrealistic expectations, any research or experiment into spatial sound delivery that does not address the psychological issues is potentially flawed or of marginal benefit. Chapter 5 continues to analyze the significance of the perceptual error compared to the numerical and practical issues of spatial sound delivery.

Chapter 5

Reconstruction versus Perception – A Comparison

In Chapter 3 a literature review and some simulations were discussed to demonstrate the numerical error in a soundfield reconstruction system. The empirical relationship $N_{\text{channels}} = \left(\frac{1}{50}R_{-20\text{dB}}f + 1\right)^2$ was developed with a reasonable configuration providing a 10cm radius at 2kHz using 25 channels. This Chapter looks at a comparison of the numerical soundfield reconstruction errors with the errors that may be introduced through perceptual uncertainties. This is not an easy comparison to make. In pursuing a numerical approach we can review the equivalent numerical error to an error in the direction of a sound source. This is reviewed for both a soundfield and a binaural reconstruction providing an indication of how an incorrect position relates to numerical error.

Chapter 4, Psychology of Spatial Hearing, gave an indication of the errors that can be expected due to auditory adaption, bias and sensory interaction. Given this as an angular error, we can use the numerical error graphs for sound source displacement to achieve an equivalent volume error for the soundfield. Although this comparisons ignores any perceptual effects and is insensitive to the nature of the error, it is a means to establish a comparison of the reconstruction and perceptual errors.

5.1 Soundfield Error

Consider a far field point source. The soundfield created will be given by $\frac{1}{4\pi}e^{ik\|\mathbf{x}-\mathbf{y}\|}/\,\|\mathbf{x}-\mathbf{y}\|$ with $\mathbf{y}$, the position of the sound source being suitably distant. Now consider this sound source displaced by an angle. Within the soundfield region of observation this will cause a perturbation. This error is expressed as a power ratio as before (3.1). Figure 5.1 shows the error as a function of the displaced angle for several volume radii.

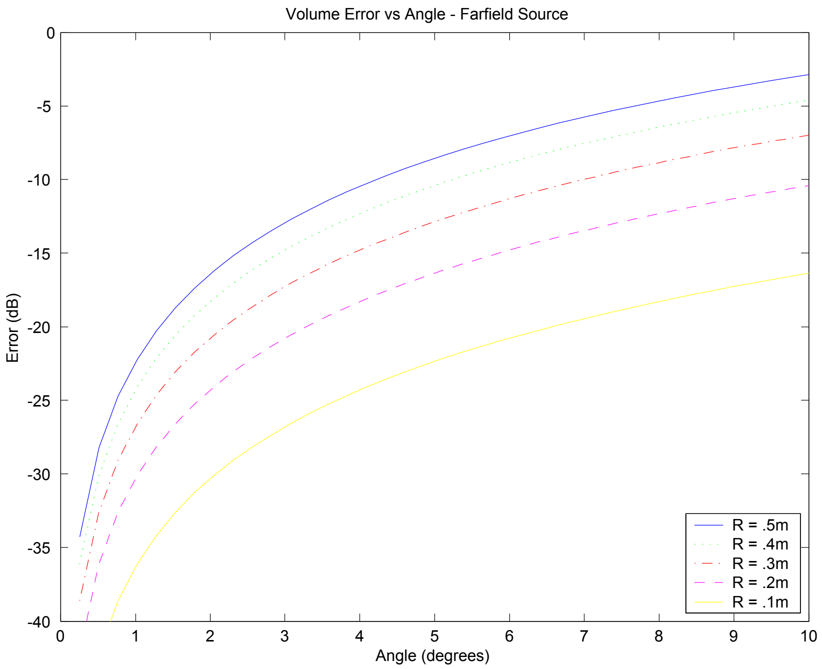

Figure 5.1: Soundfield Error vs Angular Source Displacement for a Far-field Point Source.

Previously, a reasonable bound had been considered an error of -20dB over a 20cm radius sphere using 27 speakers and 25 channels. This compares to an angular displacement of around 3.5°.

Note though that the numerical errors being compared are of a different nature. For a single far-field source, the error in reconstruction can be considered to be a "dispersion" of the source with the reconstructed soundfield more indicative of a spread sound source than a single sound source. In the displacement case, the error is obviously a displacement. However, extending this analogy to a complex soundfield that incorporates many sources with different dispersion (width) and distance properties, it becomes more reasonable to compare the numerical errors.

5.2 Binaural Error

The discussion of headphone spatial sound delivery in Section 3.4 demonstrated that it is possible to achieve accurate headphone delivery. Accurate measurements of HRTFs or adequate computational models can be combined with careful placement of headphones to achieve errors of the order of 1dB.

Consider the equivalent numerical error for binaural spatial sound reconstruction as a sound source is displaced by an angle. Figure 5.2 from [221] shows the level differences for several source positions as a function of frequency based on a simple sphere model for the head. At 1kHz, the binaural level difference is around 2dB for a source displaced by 10° and increases to 3dB as the source moves to a 45° displacement. We could interpolate a binaural gain difference of the order of 1dB corresponding to an angular displacement of around 5°. This is a simple approximation since HRTFs are very complex and it is the relative power spectral densities that provide spatial location cues [38]. The significance of individualized or non-individualized HRTFs has already been discussed. Here we are reviewing the practical errors in the delivery of binaural spatial sound and relating this to an angular source displacement.

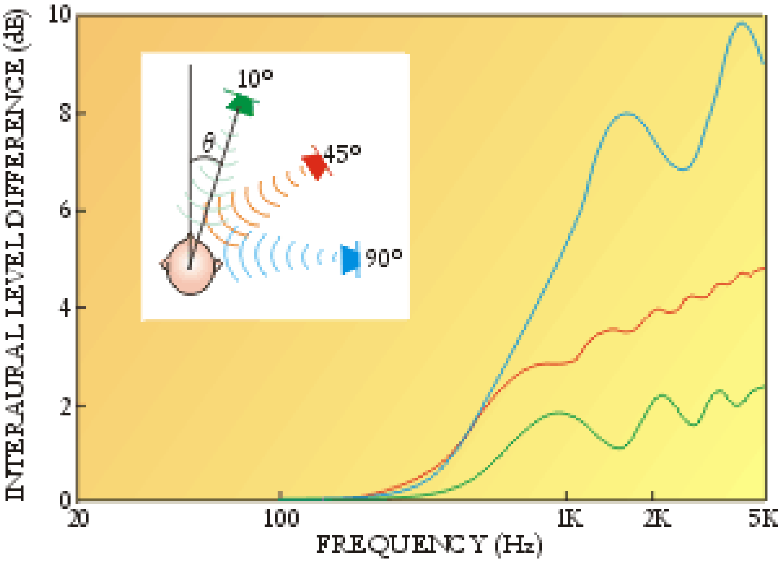

Figure 5.2: Interaural Level Difference for Binaural Signals for Different Source Positions

5.3 Perceptual Error

Listening tests in the absence of visual cues and in an anechoic chamber establish the azimuthal accuracy of spatial hearing at between 1 and 4° for the forward direction and dropping off to 12-16° for the rear [222]. Although such studies are stated as the performance of spatial hearing in “normal” situations, the experimental situation is typically in an anechoic environment with a single source and no visual cues (a blindfold or darkness). This is hardly a normal situation. Little is known about the acuity of sound localization (in isolation) in a echoic environment with many sources [106], but it would be expected to be inferior to that in the simple single source echoic environment.

We **perceive** spatial sound localization as being superior to this in a typical real environment. This is the result of interaction and combination of multiple sensory inputs and assumptions of continuity and expectations. The lay person can tolerate a discrepancy of 15 degrees in an audio source before noticing, while the professional sound designer can notice around four degrees of offset [209]. However, the suggestion of this degree of error in a subjects spatial hearing would be disputed.

Psychological experiments have demonstrated that the true performance of the spatial sound localization (not the overall perceptual performance) in more complex environments is in fact inferior to that in an anechoic, single sense environment. Previous sound in the form of broadband noise, has been shown to shift directional perception by 3-4° for forward sound sources [105]. After effects of misaligned visual and auditory cues can cause a bias of around 4° [114]. Where visual cues dominate, the perceived audio location can be shifted by as much as 30 degrees in azimuth, and even greater angles in elevation [119]. From the evidence in the existing literature it is reasonable to expect a perceptual bias or error of the order of 4°. This is not the worst case. Where strong alternate sensory cues are present, especially visual, the perceptual errors are likely to far exceed this.

In Sections 5.1 and 5.2 we have already seen that the numerical error or presently achievable soundfield and binaural systems is of a comparable magnitude to a positional error of around 3.5°. It is evident that the current numerical systems for a single listener are at least good enough to provide information to a sufficient numerical accuracy. Any further improvement in the numerical accuracy will arguably go unnoticed and have minimal effect on the perception of the listener.

This is not strictly the case for multiple listeners, where the accuracy of a soundfield over a large region could be considerably improved to give an appropriate stimulus to each listener. However, the problem of large region soundfield reconstruction remains a challenge for the practical reasons stated in Section 46. For a single listener, increased numerical accuracy is not warranted for improved perception. For multiple listeners, any improvement in the theoretical numerical performance is likely to be countered by additional errors introduced by the practical issues.

5.4 Expectations – A Comparison to Visual Simulation

In this section, a comparison is made to visual simulation. One issue which becomes apparent is the very different level of expectation that exists for visual and audio simulation. No matter how good the color, resolution or realism of a portrait is, we rarely expect it to speak to us. We do not expect visual representations and simulations to be perfect – we are quite tolerant to imperfections and actually tend to "see through them" rather than focusing on them. This is not the case with audio.

Much of this relates to the premise of the continuity of perception (Section 4.3) discussed earlier. For many reasons we are more accustomed to discontinuities in the visual field – we saccade, we blink, we constantly interpret images representing reality and we watch video images with abrupt scene changes. We are able to quickly compensate for significant distortions of perspective and scaling in still and moving visual imagery. Until recently video games had particularly coarse graphics and imagery but still delivered an extremely engaging experience. In driving we routinely create a reward spatial awareness through the occasional glance at a small mirror. The skill of a movie director is in the use of discontinuities, imagery, perspective, framing, organization and association to deliver to the audience a perception rather than to create an accurate simulation. The visual media and visual arts are far more advanced in the understanding and implementation of perceptual delivery, and we as an audience are far more experienced. We can allow video to cut and chop and change but still have a strong perception of some virtual world through which we have had only glimpses We cannot accept the "God's ear"[1] perspective of audio [15].

Through modern technology we are "trained" to assume an alternate point of view through our visual sense – we construct an parallel reality through a photograph with little concern for issues of absolute scaling or orientation as we assume the position of the camera in the original image space. We can focus our attention on this parallel reality (static or dynamic) and suppress conflicting sensory input from our immediate reality that would otherwise de-

[1] The "God's ear" perspective refers to the discontinuous audio perspective created when the point of view changes to follow the visual scenes. It is an analogy to being able to "listen in" from a convenient location without having to move or physically be present.

tract from the interpretation of the image or video. The network logos now overlaid on TV broadcasts can be distracting but are invariably ignored most of the time. Few people are even aware of the sequence of circles at the upper right of film (and video transfers) that are typically present to indicate the timing for a reel changeover. Compare this to the distraction of an occasional minor static or hum in an audio presentation.

The sensitivity and immediacy of the audio sense is reinforced in many ways. As mentioned previously, it provides a constant surveillance of our surroundings. Sounds can wake us, draw our attention and provoke reflex responses. Even in modern society where the unexpected threats have been reduced, there is a significant benefit in the ability to be able to detect minor aberration in sounds – changes in characteristic sounds can indicate early fault warnings or aberrant behaviour of most machinery.

Thus, in comparison, when we begin to project through a simulated audio presentation an alternate reality, it is difficult to disconnect from the actual reality (Section 4.3) and we are acutely sensitive to errors and irregularities in the simulated presentation. In this sense we expect comparatively more from a spatial audio system than we would of a spatial video system, however as shown in Section 5.3 this is not a simple expectation or requirement in the sense of numerical accuracy. It is a greater expectation in the sense of continuity, plausibility and absence of distraction.

5.5 An Experiment in Spatial Auditory Perception

In working on this thesis and reviewing personal experience, existing literature and thought experiments, a strong asymmetry has been identified in the experimental field of spatial audio. This asymmetry is best stated as **"The visual cue of an absence is very different to the absence of a visual cue."**

This asymmetry is prevalent in all experimental work predominantly because of the degree of difficulty in creating a compelling visual cue of an absence. To see the implications of this issue it will be discussed in greater detail.

As was shown in Chapter 3 and reviewed in Section 5.3, it is feasible to create sufficiently numerically accurate soundfield or binaural simulations to represent and arbitrary sound source positioned in space. It is trivial to conduct such an experiment in an environment where the possibility of any conflicting visual or other sensory cues are removed – the dark comfortable chair. Thus it becomes quite possible to construct an experiment where this simulation is perceptually accepted as being the real sound source. Even where visual or other cues are present this is still possible [167]. Real visual cues and other stimulus can be positioned in the place of the appropriate sound generating object, however they can be made mute. Thus it is easy to create a situation in which there is little suspicion that the soundfield being perceived is artificial.

Following on from this it is trivial to create an acoustical simulation where the visual and other sensory cues, including the premise of continuity, are in conflict with the situation. Arguably, wherever visual or other sensory cues are not suppressed, this is the normal experimental situation. It is nearly impossible to create and experimental situation which provides the presence of visual and other sensory cues that conflict with a real soundfield.

The requirement of an invisible sound producing source is rather prohibitive. Any attempt to create synthetic visual or other sensory cues, although they in themselves may be compelling, is not without the knowledge of the subject that the cues are artificial. Although we are trained to project and construct a reality an alternate representation in the visual sense (Section 5.4) we none the less know it to be an artificial construct – the portrait analogy. Thus any present visual simulation of reality will be known to be a simulation. We know when we are wearing a headset, shutter goggles or looking through a restricted field of view screen. We know when our visual sense is impaired – even though we chose to interpret an alternate reality this does not mean we expect to be in it and experience the appropriate auditory stimulus. Thus an artificial simulation of the absence of a visual cue with current technology cannot be considered a true visual cue of an absence.

The use of mirrors and appropriate optics presents the possibility of creating a discrepancy between the visual and auditory stimulus with creating an artificial visual scene. This has been applied in some recent experiments and in support of this argument, the adaption was weakened when the subject was aware that their vision was being distorted [112, 203, 205]. Generally though, the physical and acoustical effect of prisms or mirrors will provide an indication that the visual scene is altered.

A typical criticism of headphones is that the spatial sound imaging failure is the result of numerical accuracy. However, there is no experimental design that easily replaces headphones with the actual soundfield while retaining the visual and other sensory cues of the environment in which the headphones were tested. This amounts to instantly changing the acoustical environment around the subject with no changes to the visual environment.

Since this experiment cannot be performed directly, a search was carried out for an experiment where the visual and other sensory environment provided a real and convincing set of cues that were in direct conflict with a real soundfield. That is to say, an experiment where an actual sound source was present and producing sound along with a set of sensory cues that mandated such a sound source could not possibly be present – not simply that a sound source could not be seen.

The following experiment was performed over 60 years ago by Wallach [223] and fulfills the above criteria. A stationary listener is positioned inside a striped cylinder. The cylinder is constructed from a thin acoustically transparent but visually opaque material. On commencing this experiment the cylinder begins to rotate slowly. If this rotation is below the threshold of detection for the sense of balance or rotational inertia, the subject soon perceive himself or herself as rotating within a stationary enclosure. A sound source is place directly in front of the listener on the other side of the cylinder (about equidistant as the listener from the cylinder) as shown in Figure 5.3. In this situation, the supposition of a sound source rotating in sync with the subject on the opposite side of a stationary enclosure is not particularly plausible, however this is the only perception that would be consistent with the exact real sound field arriving at the subject from the stationary sound source on the opposite side of the cylinder.

The sound source is regularly perceived by the subject (12 out of 15) as being stationary either directly above or directly below the listener. This experiment is referenced recently by Moore with regard ti the interaction of the senses ".. the interpretation of auditory spatial cues is strongly influenced by perceived visual orientation." [89]. A stronger interpretation is that an exact soundfield can be incorrectly perceived when there are strong visual cues that conflict with the auditory reality – the presence of the visual cue of an absence! Strictly speaking, the cue is not just visual but also a combination of ego-motion and plausibility. If the subject perceived himself or herself as stationary in a stationary enclosure, the premise of an acoustically transparent cylinder becomes more plausible and the effect collapses – the motion of the cylinder and time for the subject to falsely perceive ego-motion is obligatory.

Interestingly, one of the harshest criticisms of headphone simulation is the elevation of forward sound sources. This criticism remains directed at the numerical accuracy of a headphone system. The experiment referenced above is the closest to simulating the situation of headphones simulating a sound source in the forward visual range that is not present while using an actual soundfield. Interestingly, in this case the perceived elevation of a sound source is complete and very much in conflict with the actual audio stimulus presented. The numerical difference between the numerical and perceived soundfield is almost maximized.

It is true that any sound source simulated within the visual field will have the strongest contradictory visual cues and a solution for this is to perceive the sound source at the nearest plausible location out of the visual field. It is also true that in an evolutionary sense, an attack is far more likely from above than from below. Since in the forward direction, the visual sense dominates [89, 212], elevation of frontal sound sources away from the visual fovea is

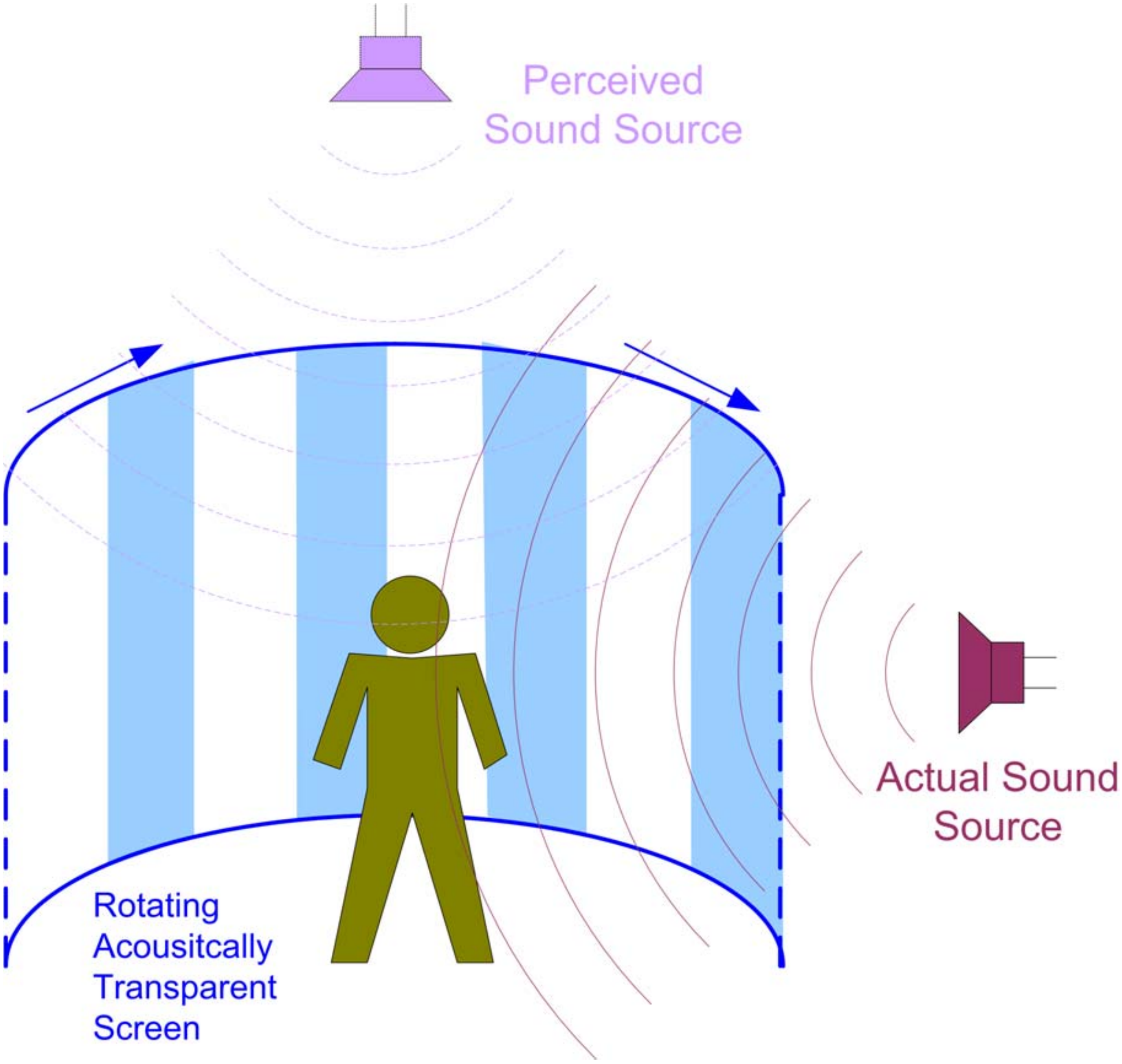

Figure 5.3: Schematic of an Experiment to Resolve the Asymmetry of Visual and Auditory Cues in Soundfield Perception

consistent with a model of the safest and smallest perturbation interpretation of conflicting cues. It is observed in this experiment with an exact soundfield. Thus it should be **expected** with a headphone simulation. Avoiding this perceptual phenomena with headphone listening may not be possible without removing or altering the conflicting visual cues.

5.6 Chapter Summary

Chapter 3 set out the numerical and practical obstacles for accurate soundfield reconstruction. Chapter 4 introduced the various issues in the psychology of spatial sound perception that will cause uncertainly and error in the spatial sound experience of the listener. This chapter provided a comparative analysis of these two concerns. Arguably, a second or third order sound field system provides a numerical accuracy, for a single listener, that is matched to the uncertainty and error in perception. Any further effort in the soundfield accuracy will not improve the perceived accuracy. Similarly, present headphone simulations provide a numerical accuracy that surpasses the notional numerical uncertainly related to perceptual errors. By comparing sound to vision, some examples were given that show we are not conditioned to "hear through" a virtual audio system in the same way we see through a picture or video. It is probably something we can learn to do better – that is, learn to focus on the spatial cues present within the spatial sound simulation rather than the cues that conflict with our actual or previous environment. Where conflicting sensory or contextual cues are present, the perceptual uncertainty can overwhelm any numerical uncertainty. To illustrate just how significant perceptual error can be, an experiment was discussed that shows the perceived location of a sound source can be displaced by as much as 90°.

Chapter 6

Conclusions

6.1 Overview

For accurate volumetric sound field reconstruction, mathematical constraints dictate large numbers of channels and speakers. Practical issues limit the performance we can expect from a spatial sound system. However, we do not just hear, we perceive. The correct ultimate goal of any spatial sound system is to effect a perception in the listener. Perception is not a deterministic result from experiencing a numerically accurate soundfield – the psychological aspects of spatial sound perception cannot be ignored. Without addressing the psychological and sense integration, numerical accuracy is not an efficient solution to the actual problem of spatial sound delivery.

It is easy to ignore the other senses and treat hearing alone. It is easy to forget the impact of impression, memory and understanding on what we believe we are hearing. This is the failing of many spatial sound systems and experiments. Even with perfect audio reconstruction, conflicting information or sensory input is present (e.g., sensation of headphones, removal of other senses, lack of continuity with previous sensory input) then the perception of the spatial sound may be incorrect. If the subject is willing to suspend his or her disbelief, even a low resolution audio simulation can be compelling. The psychology of spatial perception can work both for and against us.

As introduced in Section 1.4, this thesis set out to answer two main questions:

> **“What is more important – accurate soundfield or binaural reconstruction or the management and control of perceptual influences?”**
>
> **“Is headphone presentation of binaural spatial sound good enough?”**

The literature review, mathematical considering, simulation results and psychology discussion have covered the material that is appropriate to addressing these questions. In short, the conclusions can be stated in the following paragraphs.

Importance of Numerical Accuracy versus Perception

Perceptual errors can be expected of the order of a 4° angle (Section 5.3). The numerical error in soundfield or binaural reconstruction with present practical systems is of the order of a 3.5° to 5° angle (Sections 5.1 and 5.2). Current numerical systems for a single listener are at least good enough to provide information to a sufficient numerical accuracy. Any further improvement in the numerical accuracy will arguably go unnoticed and have minimal effect on the perception of the listener. Though this may not be the case for multiple listeners, with soundfield reconstruction over a large volume the practical issues begin to outweigh any improvements in numerical accuracy that can be achieved.

Are Headphones Good Enough?

The answer to this is two fold. Where the audio sense is considered in isolation and the other senses are "deprived" the answer is perhaps not. They can be made to be extremely close with suitable individualization of the responses and the use of headphones that do not impose a physical presence can provide convincing spatial audio. However, the numerical uncertainty in the headphones is perhaps slightly larger than that equivalent error in the perception of spatial audio in the absence of other senses.

In the case where other sensory input is present, the answer is a fairly confident yes. Incorrect perception of headphone spatial audio will still occur, however, this would occur even if the acoustical delivery was perfect. It is the typical conflict and integration of other sensory input when using headphones that causes this effect. To improve the perceived system performance, effort should be directed to suitable control or alteration of the other sensory inputs to improve the overall simulation or environment consistency with the audio presentation. Further effort in improving the individualized responses of headphones and concerns for the numerical accuracy are not warranted if they are to be used in a rich sensory context. The asymmetry of experiments has been discussed to demonstrate the difficult of experimental validation of this result. However the Wallach experiment, Section 5.5, provides strong evidence that even if headphone or soundfield simulation was exact, if the visual cues are in conflict the perceived result is not deterministic.

The following section looks at the implications of the results and discussions to the specific problem of creating a virtual audio display. Finally, Section 6.3 looks at several open problems that have been revealed in this work and sets out some ideas for how they may be pursued. A general comment from the work is that in the field of numerical soundfield representation and reconstruction, the mathematics is already present and adequate. Although further results are no-doubt available to create incremental improvements in the accuracy and efficiency of a soundfield system, the heart of the problem now lies with issues of perception. A "better sound" in not confirmation of a successful practical realization of a numerical result, neither will an improved numerical result necessarily imply an improved perception. Without understanding the uncertainties and psychological issues in this area, soundfield reconstruction is an area likely to stagnate with the numerical research divergent from any perceptual progress – it is math for math's sake and does not address the fundamental problem.

6.2 Implications and Suggestions for Spatial Audio

It has been stated that perceptual design is useful for reducing spatial audio simulation complexity [166]. Further, from the results of this thesis, it is apparent that perceptual design is essential to audio displays and the best way forward for improved spatial sound systems. It is not as important to get things numerically accurate as it is to match the complexity and consistency of a real environment. Spatial audio needs to be plausible and consistent with the listeners environment and expectations. Computational effort should be directed towards the aspects of the virtual auditory display that are most perceptually significant [148]. Delivery of spatial audio is a broad systems engineering area with a need to cover issues from implementation through to perceptual psychology [19].

The concepts of auditory scene analysis and auditory streaming [80, 99] are useful in the design of spatial audio for VR situations. Organization theories such as that of Gestalt provide a framework to assist in achieving the desired perception of an auditory scene. Some results are available discussing what aspects of an audio environment can be left out, what must be included and how to exaggerate the desired grouping of sounds to encourage association with a particular object or event. It is not always necessary to get exact spatialization for all sound components, just a strong majority to allow auditory grouping to facilitate compelling localization. There is a tendency a listener to experience a perception that is consonant with a normal stimulus situation [112]. Familiar sounds, environments, objects and behaviours will help to immerse a listener in a spatial audio simulation.

It has been shown that the auditory system is capable of adapting to modified spatial sound delivery. To enable reenforced learning, the system should incorporate some form of feedback [109] and remain consistent in the numerical characteristics of the spatial sound presented [135]. For fastest adaption, the subject should be under the impression that the sound and matching sensory stimulus refer to the same event or object [205, 214]. The goal should be to stimulate the as many sense modalities as possible in a way that has a plausible congruent interpretation. Where there is a high degree of temporal and associative connection between video and audio material, the perception will be of a single event located closest to the visual stimulus [111]. Appropriate synchronization of the sensory stimuli is very important for proper association [114, 119].

In the design of a spatial audio system, it is not as important to get the localization correct as it is to create a consistent acoustic environment. Listeners are sensitive to abnormal changes in the relationship of direct sound and echoes [95, 129]. Any characteristics of the simulated acoustic environment that do not match expectations or plausibility criteria will significantly detract from the spatial sound perception. The required level of complexity for acoustical modelling is an important is an area for further study. Use of the masked reconstruction effect can aid in increasing the quality of the perception [83] – as used in the concept of comfort noise in telephony, a suitable masking noise is more appropriate than a complete absence or unusual characteristic of any audio presentation.

Individual HRTFs are not required for most spatial audio environments. It is evident that we can learn to hear through some-one else's ears [127]. Spatial perception will improve as we learn the ear-prints and acoustic space of a simulation [129, 130]. It is important to use the HRTFs of a good localizer [187] as the HRTFs of a poor localizer can be in some way degenerate. We can learn the linear remap of head size and azimuth location [135] in a fairly short time. We can learn the nuances of other pinnae over time [141, 143] and this knowledge persists as would a second language. Compensating for head movement of the listener can accelerate the learning of the new pinnae and aid in resolving front-back confusion in the absence of string visual cues [139].

Headphones should be selected to reduce the variability in their transfer function [34] and also for comfort and minimal perceptual loading [26]. The smoother the transition to headphone listening, the more compelling any subsequent spatial audio presentation will be. Often the lack of continuity in the acoustic environment and detecting the presence of headphones is far more destructive to the sense or realism than poorly matched HRTFs. With headphones, any sound source not rendered with significant complexity or any noise introduced after binaural synthesis will not be externalized [8]. This provides further indication of an artificial spatial audio presentation and significantly reduces the "realism" and compellingness. For any spatial sound over headphones to be convincing, all other sounds over headphones must also be spatialized.

There is evidence to support that we can learn to “hear into” virtual reality systems. Increased exposure to spatial audio will help people learn to use the information presented to project into an alternate reality in the same way that we have learnt to engage with video. No video system is “indistinguishable” from reality, yet we readily place ourselves in the frame of the picture without being instructed to do so. We are trained to spawn a local reality in which we perceive the picture a similar skill may be developed with audio.

There is an argument that Virtual Reality will never be instantaneously compelling. Without the exact replica of the senses, there will be a time of learning the expectancy and continuity rules of the new environment. Much as we are not born into reality with immediate conscious perception, the same may be true of virtual reality. However, there are strong arguments that the learning and exposure is cumulative, even when it is not continuous. A definite goal is to improve the perceptual quality and sense of presence. Audio is important in achieving this [1], however it is not necessary [2, 224] and from this thesis not useful to strive for numerically correct reconstruction.

In spatial soundfield reconstruction, a lot of work is carried out to overcome the complex acoustic problems, practical setup errors and uncertainties in the listening environment – these efforts may be futile. For headphone delivery of spatial audio, errors in the transfer functions and individualization of Head Related Transfer Functions (HRTFs) is often the main cause for concern, however this may not be necessary. Experiments in perception of real sound sources show that even if the soundfield is exact, perception may differ with time and environmental factors. These results should not be put aside or forgotten, but incorporated into virtual audio display design to set sensible expectations and increase perceptual performance through the control of other factors.

The magician carefully manages the attention and focus of their subjects to produce the illusions of disappearance and transformation. Virtual Reality in the future should evolve to use techniques to draw our focus and attention, delivering appropriate sensory cues to create a desired perception. Mathematical elegance and correctness may not be as important as timing, congruity and continuity.

6.3 Open Problems

6.3.1 Optimized Decoding

The decoding equations of Chapter 2 are based on a angular invariant optimization. When the speaker array geometry is spherically symmetric, they create a spherical region of reconstruction with uniform error vs. angle distribution. This is appropriate where a spherical region is suitable for the application. In practice, rather than a spherically symmetric array, it is likely that the speaker array will be non-uniform with speaker distributed to match the desired directions of accurate sound localization. It is easy to take advantage of this when using directional panning laws [52], but becomes a non-trivial problem for soundfield representation and decoding.

Consider the case where the usual modal soundfield representation is used, but the speaker geometry is not uniform and possibly even degenerate or ill-conditioned on the sphere. The decoding equation (2.18) repeated again here for easy reference,

$$\begin{aligned}
\mathbf{S}_s(\omega) &= \begin{bmatrix} \alpha_0^0(\mathbf{x}_s, \boldsymbol{\omega}) & \alpha_1^{-1}(\mathbf{x}_s, \boldsymbol{\omega}) & \alpha_1^0(\mathbf{x}_s, \boldsymbol{\omega}) & \alpha_1^1(\mathbf{x}_s, \boldsymbol{\omega}) & \ldots & \alpha_P^P(\mathbf{x}_s, \boldsymbol{\omega}) \end{bmatrix}^T \qquad (6.1) \\
\mathbb{S}(\omega) &= \begin{bmatrix} \mathbf{S}_1 & \mathbf{S}_2 & \mathbf{S}_3 & \ldots & \mathbf{S}_N \end{bmatrix} \\
\mathbf{U}(\omega) &= \begin{bmatrix} U_1(\omega) \\ U_2(\omega) \\ \ldots \\ U_N(\omega) \end{bmatrix} = \mathbb{S}(\omega)^T \left(\mathbb{S}(\omega) . \mathbb{S}(\omega)^T\right)^{-1} \begin{bmatrix} B_0^0(\omega) \\ B_1^{-1}(\omega) \\ \ldots \\ B_P^P(\omega) \end{bmatrix}
\end{aligned}$$

may construct a speaker contribution matrix that is not easily inverted. As seen in Figure 3.3, the gains for a uniformly distributed array are regular. For an arbitrary speaker array, this may not be the case.

Consider further, the case where the region for which accurate soundfield reconstruction is required is not spherical. An ellipsoid may be more suitable for a listener who is unlikely to rotate their head significantly about the forward facing direction. Perhaps even a toroid that covers the expected location of the ears through azimuthal head rotation and some degree of translation. It must be possible to optimize the speaker decoding coefficients to reduce the reconstruction error over an arbitrary region. This is an open problem for further investigation. The required number of channels and speakers would be reduced by constraining the effective region of reconstruction. This is notable in the two-dimensional case where the number of terms for a given order representation is $2N+1$ as opposed to $(N+1)^2$ [23]. However, in the case of planar soundfield reconstruction, all elevation directional information is lost. It would be more appropriate to still reproduce the elevation effects but perhaps with a precision that relates to the relative spatial hearing sensitivity of elevation versus azimuth.

The following equations set out this problem as an optimization. The speaker gain coefficients are selected to optimize the error in reconstruction over the desired volume and for a suitable class of soundfields that is representative of the desired soundfield to be reconstructed. Rather than the speaker gain coefficients being $\mathbb{S}\left(\omega\right)^T\left(\mathbb{S}\left(\omega\right).\mathbb{S}\left(\omega\right)^T\right)^{-1}$ which minimizes, in a least squares sense, the matching error in the modal domain, the speaker gain coefficients will minimize the error in the spatial domain. Representing the decode with a general gain matrix $\mathbb{G}\left(\omega\right)$ of size $N \times (P+1)^2$,

$$\mathbf{U}(\omega) = \begin{bmatrix} U_1(\omega) \\ U_2(\omega) \\ \cdots \\ U_N(\omega) \end{bmatrix} = \mathbb{G}\left(\omega\right) \begin{bmatrix} B_0^0(\omega) \\ B_1^{-1}(\omega) \\ \cdots \\ B_P^P(\omega) \end{bmatrix} \tag{6.2}$$

Remaining in the frequency domain, we can write the equation for the actual spatial pressure field generated by these speaker signals using the speaker positions $\mathbf{x}_i$ and the point source radiation equation $\frac{1}{4\pi}e^{ik\|\mathbf{x}_n-\mathbf{y}\|}/\left\|\mathbf{x}_n-\mathbf{y}\right\|$,

$$p'(\mathbf{y},\omega) = \sum_{n=1}^{N} U_n\left(\omega\right)\frac{e^{ik\|\mathbf{x}_n-\mathbf{y}\|}}{4\pi\left\|\mathbf{x}_n-\mathbf{y}\right\|} \tag{6.3}$$

Now consider the ideal soundfield that would be reconstructed from the given modal signals $B_n^m(\omega)$. Substituting these into the synthesis equation (2.10),

$$p(\mathbf{y},\omega) = \sum_{n=0}^{P}\sum_{m=-n}^{n}\left(4\pi j^n B_n^m\left(\omega\right)\right) j_n\left(k\left\|\mathbf{y}\right\|\right) Y_n^m\left(\widehat{\mathbf{y}}\right) \tag{6.4}$$

Now to optimize the reconstruction over a particular volume introduce a weighting vector $W\left(\mathbf{y}\right)$ to effect this such that $\iiint W\left(\mathbf{y}\right).d\mathbf{y} < \infty$ then the error in the reconstruction based on a set of speaker gains $\mathbb{G}\left(\omega\right)$ and a particular soundfield $\mathbf{B}\left(\omega\right) = B_n^m\left(\omega\right)$ will be given by,

$$\begin{aligned} E_{\mathbb{G}(\omega)\mathbf{B}(\omega)} &= \iiint W\left(\mathbf{y}\right)\left|p(\mathbf{y},\omega) - p(\mathbf{y},\omega)\right|^2 .d\mathbf{y} \qquad (6.5) \\ &= \iiint W\left(\mathbf{y}\right)\left|\begin{array}{c} \sum_{n=0}^{P}\sum_{m=-n}^{n}\left(4\pi j^n B_n^m\left(\omega\right)\right) j_n\left(k\left\|\mathbf{y}\right\|\right) Y_n^m\left(\widehat{\mathbf{y}}\right) \\ -\sum_{q=1}^{N} U_q\left(\omega\right)\frac{e^{ik\|\mathbf{x}_q-\mathbf{y}\|}}{4\pi\|\mathbf{x}_q-\mathbf{y}\|} \end{array}\right|^2 d\mathbf{y} \end{aligned}$$

To find a suitable decoder for a class of soundfields, the optimization problem is expressed on the expectation value of $E_{\mathbb{G}(\omega)\mathbf{B}(\omega)}$ for $\mathbf{B}\left(\omega\right)$ having a suitable distribution for the desired soundfield. In the absence of any other information, consider $\mathbf{B}\left(\omega\right)$ to have a normalized white distribution. The optimization is then given as,

$$\mathbb{G}\left(\omega\right) = \arg \min_{\mathbf{B}(\omega)\in\{B_n^m(\omega)=\mathbf{N}(0,\sigma)\}} \left\{\mathbf{E}\left[E_{\mathbb{G}(\omega)\mathbf{B}(\omega)}\right]\right\} \tag{6.6}$$

With the gain obtained for a suitable set of frequencies (ω), the appropriate decoding filters can be obtained.

This is a complex calculation and an analytic solution would be difficult to obtain for an arbitrary weighting function $W\left(\mathbf{y}\right)$. However, constructed as a numerical optimization problem, with a suitable finite element dimensionality over the spatial integral (100-1000 would be sufficient) the problem could be solved numerically with a standard numerical processing package such as Matlab on a desktop PC. Note that the solution need only be found once for a specific speaker array geometry. The computation required during reconstruction of the soundfield is equivalent to that for the standard decoding equation.

6.3.2 Effect of the Listener on the Soundfield

In Section 3.2 the effect of the presence of the listener on the soundfield was discussed. Ideally, we are trying to reconstruct the same soundfield around a listener that would have been present had that listener been present in the same relative position in the original soundfield. This is obviously not possible without restricting the position of the listener, in which case a binaural approach would be better suited. A general assumption is made that if the soundfield is captured with the listener absent, and then reconstructed over a suitably large volume with the listener present, the effect of the listener will not destroy the reconstruction. This assumption is based on the extension of the Kirchoff-Helmholtz integral (2.20), which shows that by controlling or knowing the soundfield only on the surface of a volume we control or know it throughout the interior.

The problem with this assumption can be stated in two conjugate ways :-

- It is generally understood that the Kirchoff-Helmholtz integral is based on the volume enclosure being devoid of any sound sources. However, in the strict sense, the Kirchoff-Helmholtz integral is only correct where the enclosed volume is free of any acoustically active or reactive elements.

- If it we consider the need only to control the surface of the volume enclosure, then the appropriate field to excite on the surface in reconstruction would be the field that existed in the original soundfield if an equivalent object to the listener was present in the appropriate relative position. Further, it would not be possible to use the Kirchoff reconstruction integral to control the pressure over the entire region. A form of actual control would be required with some form of feedback used so that the speaker actuation was modified to control the pressure field over the surface, compensating for the reactive properties of any object within the region.

The effect of the listener is due to the scattering and obstruction of the sounds being produced by the reconstruction speakers. Since the speakers are not coincident with the original sound source location, the scattering and obstruction effects will be different in the case of the ideal field and the reconstructed field. This effect becomes smaller as more speakers are used at greater distances from the listener's location.

The open problem is how we can analyze and assess the impact of this incorrect assumption on the reconstruction of a soundfield.

As a first step, we can consider using in the speaker decoding formulation the soundfield created by a speaker with the listener present. That is, instead of considering $\frac{1}{4\pi}e^{ik\|\mathbf{x}_i-\mathbf{y}\|}/\ \|\mathbf{x}_i-\mathbf{y}\|$ as speaker contribution, we can consider the net soundfield from the incident and scattered wavefronts (Figure 6.1). Given an model of a listener (perhaps a simple sphere) we can determine a formulation for this combined soundfield. This could be used in the optimization speaker gain determination framework (6.6). Alternately, by representing this combined soundfield in the modal domain, the simple inverse decoding formulation could be used (2.18).

Ultimately the problem would be to consider the soundfield created around the scattering object when the actual sound sources were present, compared with the soundfield created around the same scattering object when the reconstruction sound sources are present. This is shown schematically in Figure 6.2. This could be evaluated for different positions of the scattering object (with the same relative position in both the actual and reconstructed situations).

This approach would allow an estimate of the significance of this effect, and by using the modified decoding equations it would be a way to introduce a first order correction factor to the problem.

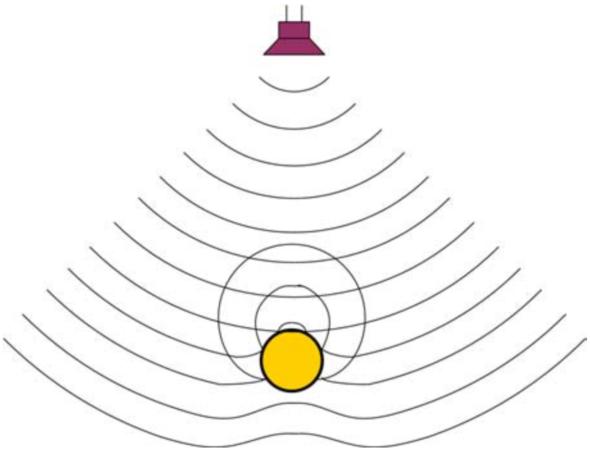

Figure 6.1: Point Source modified for Listener Scattering and Obstruction

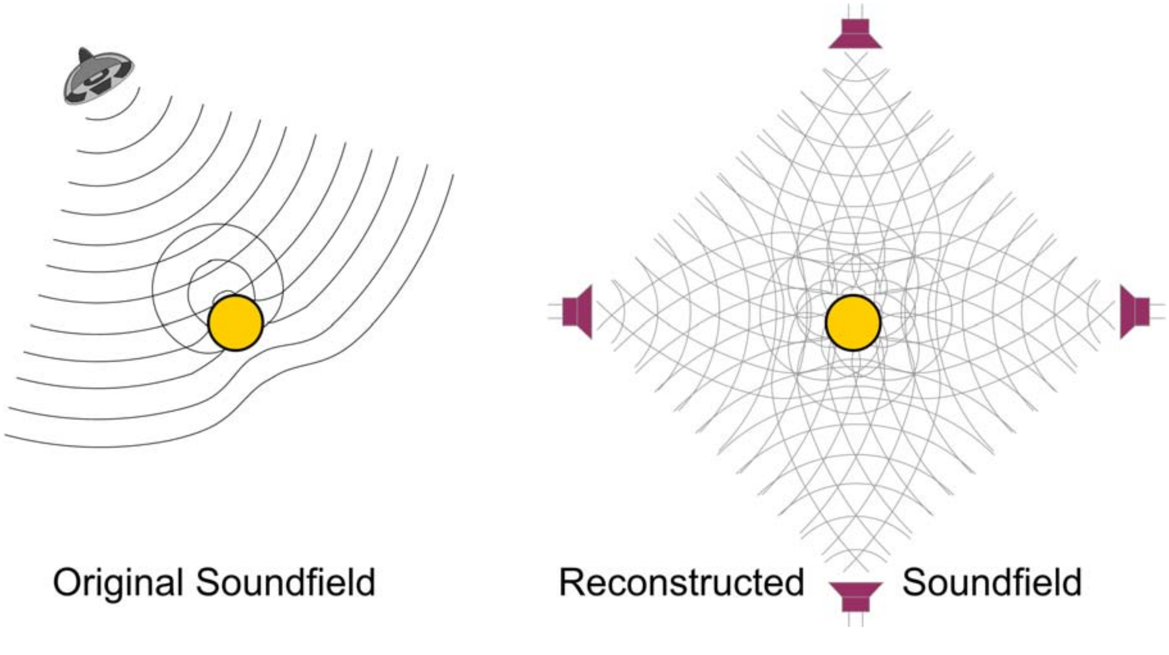

Figure 6.2: Schematic of Comparison of Actual Sound Source and Reconstructed Sound Source using Modified Sphere Diffraction Soundfield

6.3.3 Perceptual Design of Spatial Audio

The thesis has presented a comparison of the numerical, practical and perceptual issues in the delivery of spatial sound. From the discussion it is apparent that the perceptual issues and uncertainty dominate the perceived performance of a spatial sound system. Given this the approach to designing the content and delivery of spatial audio should be largely based on perceptual issues. This concept of audio design or “cartoonification” [157] is a considerably challenging and open problem. Audio signals would be selected based on the likely perceptual response rather than necessarily the physical correspondence. The choice of sounds will be effected by the issues of context and association as well as societal and cultural elements. This is an art that is well practiced in movie sound tracks and sound effects. For example, we rarely question the many toolkit sounds of the Foley stage such as the creaking door, paper rustle, sandbox and water bowl. Other sounds such as a gun ricochet, car tyre screech, the whoosh of air from a hand manoeuvre or the thud of a connected punch are so ingrained and expected that we associate them with the actions and they add to the immersion of a movie, even though they would rarely occur in reality.

An extreme example of this is the sound of space-ships and explosions in space familiar from the Star Wars epic. A physically accurate simulation would be completely silent, yet this would not match the expectations of a typical audience and would detract from the perceptual experience. In many ways, beyond just the numerical accuracy, spatial sound to achieve a desired perception will be divergent from a physically realistic simulation. As this is a “soft science”, it does not appeal to a broad section of the academic community, however with the discussions of this thesis it is only too apparent that if the goal is truly to optimize the **experience** of a spatial sound system then these issues are more pertinent than numerical analysis. It is also apparent that improved techniques for forming quantitative measures of perceptual efficiency are required to track progress in this area.

Hearing has evolved to improve our survival [157, 193]. Sound localization has evolved to perform optimally in certain conditions; generally these conditions relate to the survival of the organism. To really understand what the mechanisms of sound localization really are, it would be necessary to study sound localization in this environment [192]. One of the most basic features of the auditory system to evolve would have been the ability to separate and analyze an auditory scene [220], giving attention to the important source in a complex auditory environment. Since most studies deal with a single sound source, this suggests that hearing research to date can be misleading and incomplete when it comes to the natural perception of complex auditory environments [83, 93, 96, 193]. Perceptual research and the design of spatial audio systems utilizing these ideas would improve the ability to anticipate and control a listener’s perception in a spatial audio environment.

6.3.4 Experimental Design

The Wallach experiment detailed in Section 5.5 was provided from a review of the literature of an experiment to simulate the perception of a real auditory stimulus with the presence of the visual cue indicating and absence or impossibility of a sound source in the appropriate location. This is an experiment that would be worth repeating for the explicit goal of practical validation of the premise of this thesis. There are also other avenues of exploration for experimental design to achieve this goal. These approaches are listed in order of increasing technical requirements. Note that in all of these cases, the result is not direct and would have to be interpreted appropriately. Either the sound source is slightly modified, or the visual field is slight modified, or both. However they do represent progress in the direction of resolving the experimental asymmetry when compared with the predominant spatial sound experiments reviewed in the literature.

- The use of association to create objects that make inappropriate sounds. For example a real experimental room where a telephone boils and a kettle rings. Careful design of physical objects incorporating the actual sound making elements (boiling water, telephone bell) could eliminate any synthetic sound source or soundfield. The use of synthetic but compelling visual stimulus with modern technology.
- The use of visually invisible (transparent) acoustical sources, or acoustically transparent reflective surfaces.
- The use of high frequency acoustical demodulation to create a sound source that emanates from “thin air” with appropriate directionality properties toward the listener.
- The use of other sensory stimulus to provide a compelling sensory evidence of the absence of a sound source.
- The appropriate design of an experiment in a completely immersive environment where the auditory cues are shifted from the visual cues and the results analyzed. This approach attempts to break the asymmetry by making both visual and audio sensory input synthetic. It would require a fairly encompassing simulation environment.

The design, performance and analysis of results from experiments such as this would be an interesting area of work and provide an experimental validation and bound for the expectations for spatial audio perception where conflicting visual cues are present.

6.3.5 Improved Problem Formulation

It is evident that the simple analysis of the numerical error in soundfield over a reconstruction volume is not a useful measure of soundfield system performance. Perceptual issues cause problems for deterministic observations and are not easily incorporated into an optimization. However, different measures of the soundfield characteristics may provide a numerical foundation that is better suited to optimizing the perceived result. The ratios of phase and intensity directionality and direction have already been discussed (3.4, 3.6) and other published works consider extended geometries such as these [68]. A more general analysis of the soundfield properties could be formulated to optimize the desirable characteristics of a soundfield over an extended volume. Such desirable properties would include stable sound positions when the listener moves, minimal frequency dispersion and minimal motion artefacts (frequency response) as a sound source is moved.

Rather than discarding numerical analysis, a suitable problem formulation could be developed where numerical techniques could better predict the perceptual qualities of a soundfield and then the reconstruction problem could be solved to address these criteria.

Appendix A

Equivalence of Ambisonics and Taylor Series

This Appendix sets out the information equivalence and mapping between the Ambisonic and Taylor series representation of sound fields as set out in Chapter 2.

A.1 Taylor Series Redundancy

The Taylor series has $\frac{1}{6}\left(N^3+6N^2+11N+6\right)$ terms for representation to order N and Ambisonics provides $\left(N+1\right)^2$ terms for representation to order N. This section demonstrates the redundancy in the Taylor series coefficients, as a result of the Helmholtz equation constraint (2.7) on the pressure sound field. If we restrict the Taylor series representation to the set of scalar time-varying fields satisfying the wave equation (i.e., physical sound-fields), then it can be shown that the same amount of information is present in both the Taylor and Ambisonic representations.

Consider the Helmholtz spatial wave equation,

$$\nabla^2 p\left(x,y,z,t\right)=\frac{1}{c^2}\frac{\partial^2}{\partial t^2}p\left(x,y,z,t\right) \qquad \text{(A.1)}$$

Taking the LHS and substituting the Taylor series expansion about the origin (2.4) gives,

$$\nabla^2 p\left(\mathbf{x}\right) \;=\; \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\left[\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}p_{l,m,n}\left(t\right).x^l y^m z^n\right] \qquad \text{(A.2)}$$

$$= \sum_{l=2}^{\infty}\sum_{m=2}^{\infty}\sum_{n=2}^{\infty} p_{l,m,n}\left(t\right) \begin{bmatrix} l(l-1)x^{l-2}y^m z^n + \\ m(m-1)x^l y^{m-2} z^n + \\ n(n-1)x^l y^m z^{n-2} \end{bmatrix}$$

For simplicity of this proof, the Taylor series expansion around the origin is used. A similar redundancy can be derived for an arbitrary point. Substituting this back into (A.1),

$$\sum_{l=2}^{\infty}\sum_{m=2}^{\infty}\sum_{n=2}^{\infty} p_{l,m,n}\left(t\right) \begin{bmatrix} l(l-1)x^{l-2}y^m z^n + \\ m(m-1)x^l y^{m-2} z^n + \\ n(n-1)x^l y^m z^{n-2} \end{bmatrix} = \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} p_{l,m,n}\left(t\right) x^l y^m z^n \tag{A.3}$$

From this, we can equate the like terms of the polynomials in (x, y, z) to obtain a recurrence relation,

$$\begin{matrix} \left(l+2\right)\left(l+1\right)p_{l+2,m,n}(t) + \\ \left(m+2\right)\left(m+1\right)p_{l,m+2,n}(t) + \\ \left(n+2\right)\left(n+1\right)p_{l,m,n+2}(t) \end{matrix} = \frac{1}{c^2}\frac{\partial^2}{\partial t^2}p_{l,m,n}\left(t\right) \tag{A.4}$$

This shows that for any order N there exists a set of $\frac{1}{2}\left(N^2+3N+2\right)$ equations relating derivative of terms $p_{l,m,n}\left(t\right)$ of order N with the $\frac{1}{2}\left(N^2+7N+12\right)$ terms of order $N+2$. Since each equation has three left had side terms, there are $\frac{1}{2}\left(3N^2+9N+6\right)$ terms of the Taylor coefficients of order $N+2$. Thus for N>=2 there are more terms in the degeneracy equation set than there are terms in the higher order. Hence the terms are not uniquely represented.

In order to determine the magnitude of the degeneracy, we need to determine the number of linearly independent left hand side combinations of the order $N+2$ coefficients. However, since the right hand side $\frac{1}{c^2}\frac{\partial^2}{\partial t^2}p_{l,m,n}\left(t\right)$ represents a set of $\frac{1}{2}\left(N^2+3N+2\right)$ free variables, if there was any redundancy in the degeneracy equation, a contradiction would occur, and there would be no solution of the order $N+2$ coefficients in the recurrence relation that satisfied the $\frac{1}{2}\left(N^2+3N+2\right)$ equations. We know this cannot be the case, as the wave equation supports soundfields that are continuous and differentiable to any order. Thus we can determine that there are $\frac{1}{2}\left(N^2+3N+2\right)$ dependent terms $p_{l,m,n}\left(t\right)$ for order $N+2$.

This implies that from second order, the Taylor coefficients are over specified. We can calculate the number of unique coefficients required being the number of coefficients in the Taylor expression of order N less the number of coefficients in the Taylor expression of order $N-2$

,

$$\begin{aligned} & \frac{N^3+6N^2+11N+6}{6}-\frac{(N-2)^3+6(N-2)^2+11(N-2)+6}{6} \\ = \ & (N+1)^2 \end{aligned} \tag{A.5}$$

This is the same number of terms as required by the Ambisonic representation. Since the Ambisonics representation is based on a plane wave basis function (for direction of arrival) and plane wave form and entire solution to the wave equation, it is to be expected that the constraint of the wave equation on the Taylor series expansion is equivalent to the constraint of the spherical harmonics (direction of arrival) in Ambisonics.

A.2 Mapping from Taylor Series to Ambisonics.

This section demonstrates the equivalence between the Ambisonic and Taylor representations of a sound field. The previous section demonstrated that with the wave equation constraint, the Ambisonic and Taylor representations contain an equivalent amount of information.

Consider a plane wave travelling in the direction of the unit vector $|(x_1, y_1, z_1)| = 1$ with radian frequency ω,

$$p\left(x,y,z,t\right) = W_{x_1,y_1,z_1,\omega}\left(x,y,z,t\right) = e^{i\omega\left(t-\frac{xx_1+yy_1+zz_1}{c}\right)} \tag{A.6}$$

This forms a basis set for all acoustical sound-fields with the direction and frequency as the basis parameters [168]. By demonstrating the equivalence of Ambisonic and Taylor representations for this general basis function, the two representations are shown to be equivalent. Firstly, consider the Taylor series expansion of this plane wave at the origin. Consider the arbitrary partial derivative of the plane wave,

$$\begin{aligned} p_{l,m,n}\left(t\right) &= \left.\frac{\partial^l}{\partial x^l}\frac{\partial^m}{\partial x^m}\frac{\partial^n}{\partial x^n}p\left(x,y,z,t\right)\right|_{x=0,y=0,z=0} \\ &= \left.\frac{\partial^l}{\partial x^l}\frac{\partial^m}{\partial x^m}\frac{\partial^n}{\partial x^n}W_{x_1,y_1,z_1,\omega}\left(x,y,z,t\right)\right|_{x=0,y=0,z=0} \\ &= \left.\left(\frac{-i\omega}{c}\right)^{l+m+n} x_1^l y_1^m z_1^n . W_{x_1,y_1,z_1,\omega}\left(x,y,z,t\right)\right|_{x=0,y=0,z=0} \end{aligned} \tag{A.7}$$

$$= \left(\frac{-i\omega}{c}\right)^{l+m+n} x_1^l y_1^m z_1^n . e^{j\omega t}$$

By rearranging this we can write an arbitrary term of a polynomial in (x_1, y_1, z_1) as

$$x_1^l y_1^m z_1^n = \left(\frac{c}{-i\omega}\right)^{l+m+n} e^{-i\omega t} p_{l,m,n}\left(t\right) \tag{A.8}$$

Now the Ambisonic representation is obtained by considering the sensitivity of each of the channels to a plane wave of the incident direction. Assuming a theoretical Ambisonic system where the phase and frequency response is ideal, each of the Ambisonic signals will be a real scalar multiple of the fundamental tone. Define a signal $A_n(t)$ as the time varying Ambisonic signal associated with a directionality pattern $f_n(x, y, z)$,

$$f_n\left(x, y, z\right) = \sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\sum_{s=0}^{\infty} \gamma_{p,q,s} x^p y^q z^s \tag{A.9}$$

The directional polynomials for representation of up to third-order Ambisonics are set out in Table 2.2. For a plane wave (A.6), this signal ($A(t)$) will be a multiple of the time phasor with magnitude set by evaluating the directionality pattern polynomial at the coordinates of the incident plane wave direction. This scalar multiple will be a polynomial in the unit vector components of the plane wave direction,

$$\begin{aligned} A_n\left(t\right) &= f_n\left(x_1, y_1, z_1\right).e^{i\omega t} \qquad \text{(A.10)} \\ &= \sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\sum_{s=0}^{\infty} \gamma_{p,q,s} x_1^p y_1^q z_1^s e^{i\omega t} \end{aligned}$$

By substitution of (A.8) into (A.10), we obtain a frequency dependant relationship between the Taylor series signals and the Ambisonic signals,

$$A_n\left(t\right) = \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \gamma_{l,m,n} \left(\frac{c}{-i\omega}\right)^{l+m+n} p_{l,m,n}\left(t\right) \tag{A.11}$$

This relationship is independent of time and the plane wave direction, however it is dependent on the plane wave frequency. Taking the Fourier transform of both sides of (A.11) gives a transfer function mapping from the Taylor expansion sound field signals to an Ambisonic signal. This is equivalent to a convolutional transfer function in the time domain. Note that the $\left(\frac{1}{\omega}\right)^{l+m+n}$ is in effect an integral, adding poles at the origin ($\omega = 0$). The Taylor series representation, based on the derivatives of the soundfield at the origin is very sensitive to high frequencies. This dependence is suppressed in Ambisonics through the integral, and conceptually through Ambisonics reliance on frequency independent directional sensitivities (something difficult to achieve in practice).

This mapping is invertible for the class of fields represented by $p_{l,m,n}\left(t\right)$ that satisfy the wave equation. The proof for this is not included, however the previous section proved the two representations have the same size signal space and since the basis functions are linearly independent (Taylor) and orthogonal (Ambisonic) the mapping is invertible.

Appendix B

Equivalence of Ambisonics and Modal Analysis

Consider the Ambisonics polynomials as set out in Table 2.2 and the spherical harmonic angular dependency (2.11). The following identities can be used to generate the Ambisonic polynomials from the modal spherical harmonics. From the spherical coordinates (Figure 2.1), we can define the following for the unit vector $\widehat{\mathbf{x}} = (\widehat{x}, \widehat{y}, \widehat{z}) = \mathbf{x}/\left\|\mathbf{x}\right\|$,

$$\begin{aligned} \widehat{x} &= \sin\theta\cos\phi \\ \widehat{y} &= \sin\theta\sin\phi \\ \widehat{z} &= \cos\theta \end{aligned} \tag{B.1}$$

This then leads to the generator function for the $2N+1$ new Ambisonic polynomials at order n, which here we will name $f_n^0 \dots f_n^{2n}$,

$$\begin{aligned} f_n^0 &= \sqrt{4\pi} Y_n^0 \\ f_n^1 &= -\sqrt{4\pi}\left(\frac{Y_n^1 + Y_n^{-1}}{\sqrt{2}}\right) \\ f_n^2 &= -\sqrt{4\pi}\left(\frac{Y_n^1 - Y_n^{-1}}{\sqrt{2}}\right) \\ f_n^{2m-1} &= (-1)^m\sqrt{4\pi}\left(\frac{Y_n^m + Y_n^{-m}}{\sqrt{2}}\right) \quad 0 < m \le n \\ f_n^{2m} &= (-1)^m\sqrt{4\pi}\left(\frac{Y_n^m - Y_n^{-m}}{\sqrt{2}}\right) \quad 0 < m \le n \end{aligned} \tag{B.2}$$

This identity can be seen by expanding the first few spherical harmonics,

$$\begin{aligned}
Y_0^0(\theta,\phi) &= \sqrt{\frac{1}{4\pi}} \\
Y_1^0(\theta,\phi) &= \sqrt{\frac{1}{4\pi}}\sqrt{3}\cos\theta = \sqrt{\frac{1}{4\pi}}\sqrt{3}z \\
Y_1^1(\theta,\phi) &= -\sqrt{\frac{1}{4\pi}}\sqrt{\frac{3}{2}}(\sin\theta)(\cos(\phi)+i\sin(\phi)) = \sqrt{\frac{1}{4\pi}}\sqrt{3}\frac{x+iy}{\sqrt{2}} \\
Y_1^{-1}(\theta,\phi) &= -\sqrt{\frac{1}{4\pi}}\sqrt{\frac{3}{2}}(\sin\theta)(\cos(\phi)+i\sin(\phi)) = \sqrt{\frac{1}{4\pi}}\sqrt{3}\frac{x-iy}{\sqrt{2}} \\
Y_2^0(\theta,\phi) &= \sqrt{\frac{5}{4\pi}}\frac{1}{2}\left(3\cos(\theta)^2-1\right) = \sqrt{\frac{1}{4\pi}}\frac{\sqrt{5}}{2}\left(3z^2-1\right) \\
Y_2^1(\theta,\phi) &= -\sqrt{\frac{5}{4\pi}}\sqrt{\frac{1}{6}}\left(3\cos\theta\sin\theta\right)\left(\cos(\phi)+i\sin(\phi)\right) = \sqrt{\frac{1}{4\pi}}\sqrt{15}z\frac{x+iy}{\sqrt{2}}
\end{aligned} \tag{B.3}$$

Consider the spatial dependence of a plane wave,

$$W_{\widehat{\mathbf{y}},\omega}(\mathbf{x}) = e^{ik\mathbf{x}\widehat{\mathbf{y}}} \tag{B.4}$$

Over the entire space, plane waves form an orthonormal basis function set,

$$\frac{1}{V}\iiint W_{\widehat{\mathbf{y}_1},\omega_1}(\mathbf{x}).\overline{W_{\widehat{\mathbf{y}_2},\omega_2}(\mathbf{x})}.d\mathbf{x} = \boldsymbol{\delta}\left(\widehat{\mathbf{y}}_1,\widehat{\mathbf{y}}_2\right)\boldsymbol{\delta}\left(\widehat{\boldsymbol{\omega}_1},\widehat{\boldsymbol{\omega}_2}\right) \tag{B.5}$$

as $V \to \mathbb{R}^3$. Now, consider the Ambisonic signal $A_n^l(t)$. This will be generated by a microphone with directional sensitivity f_n^l placed at the origin. This is the projection of the actual sound field $p(\mathbf{x})$ onto a spatial function being the combination of plane waves according to the directional sensitivity f_n^l. This is easily expressed in the frequency domain with $A_n^l(\omega)$ being the Fourier transform of $A_n^l(t)$,

$$\begin{aligned}
A_n^l(\omega) &= \frac{1}{V}\iiint p(\mathbf{x})\iint f_n^l(\widehat{\mathbf{y}})\, W_{\widehat{\mathbf{y}},\omega}(\mathbf{x})\, d\widehat{\mathbf{y}}d\mathbf{x} \\
A_n^l(t) &= IFFT\left\{A_n^l(\omega)\right\}
\end{aligned} \tag{B.6}$$

Now by reference to the Ambisonic generator equation (B.2) we can expand the Ambisonic function f_n^l in terms of the spherical harmonics (2.11). Note that the the function f_n^l has in fact three expansion variants f_n^0 and f_n^{2m-1} or f_n^{2m} for $0 < m \leq n$. The expansion is shown for the f_n^{2m-1} variant,

$$\begin{aligned} A_n^{2m-1}(\omega) &= \frac{1}{V}\iiint p(\mathbf{x})\iint f_n^{2m-1}(\widehat{\mathbf{y}})\,W_{\widehat{\mathbf{y}},\omega}(\mathbf{x})\,d\widehat{\mathbf{y}}d\mathbf{x} \\ &= \frac{1}{V}\iiint p(\mathbf{x})\iint (-1)^m\sqrt{4\pi}\left(\frac{Y_n^m(\widehat{\mathbf{y}})+Y_n^{-m}(\widehat{\mathbf{y}})}{\sqrt{2}}\right)W_{\widehat{\mathbf{y}},\omega}(\mathbf{x})\,d\widehat{\mathbf{y}}d\mathbf{x} \end{aligned} \tag{B.7}$$

It is trivial to show a similar relationship for f_n^0 and f_n^{2m}. Following on from this expand the plane wave as a sum of spherical harmonics and the spherical Bessel function [168],

$$\begin{aligned} W_{\widehat{\mathbf{y}},\omega}(\mathbf{x}) &= e^{ik\mathbf{x}.\widehat{\mathbf{y}}} \\ &= 4\pi\sum_{q=0}^{\infty}\sum_{p=-q}^{q}(i)^q j_q(k\|\mathbf{x}\|).Y_q^p(\widehat{\mathbf{x}}).\overline{Y_q^p(\widehat{\mathbf{y}})} \end{aligned} \tag{B.8}$$

Substituting this into (B.7),

$$\begin{aligned} A_n^{2m-1}(\omega) &= \frac{4\pi}{V}\iiint p(\mathbf{x}).\iint (-1)^m\sqrt{4\pi}\left(\frac{Y_n^m(\widehat{\mathbf{y}})+Y_n^{-m}(\widehat{\mathbf{y}})}{\sqrt{2}}\right) \\ &\qquad \sum_{q=0}^{\infty}\sum_{p=-q}^{q}(i)^q j_q(k\|\mathbf{x}\|)\,Y_q^p(\widehat{\mathbf{x}})\,\overline{Y_q^p(\widehat{\mathbf{y}})}d\widehat{\mathbf{y}}d\mathbf{x} \end{aligned} \tag{B.9}$$

And from the orthogonality of the Spherical Harmonic functions $Y_n^m(\widehat{\mathbf{y}})$, the summation over p, q and the integral over $\widehat{\mathbf{y}}$ collapses to give

$$A_n^{2m-1}(\omega) = \frac{(4\pi)^{\frac{3}{2}}(-1)^m(i)^n}{V}\iiint p(\mathbf{x})\,j_n(k\|\mathbf{x}\|)\left(\frac{Y_n^m(\widehat{\mathbf{x}})+Y_n^{-m}(\widehat{\mathbf{x}})}{\sqrt{2}}\right)d\mathbf{x} \tag{B.10}$$

Similar relationships can be found for A_n^0 and A_n^{2m}. With the exception of some scaling factors, and the "shuffled" basis set ($Y_n^m(\widehat{\mathbf{x}})+Y_n^{-m}(\widehat{\mathbf{x}}), Y_n^m(\widehat{\mathbf{x}})-Y_n^{-m}(\widehat{\mathbf{x}})$ compared with $Y_n^m(\widehat{\mathbf{x}}), Y_n^{-m}(\widehat{\mathbf{x}})$), this shows that the Ambisonics signal is equivalent to the projection of the soundfield onto the modal basis function over a large volume.

Consider the integration over a sphere of radius R,

$$A_n^{2m-1}(\omega) = \frac{\sqrt{2\pi}(-1)^m(i)^n}{R^3}\iiint p(\mathbf{x})\,j_n(k\|\mathbf{x}\|)\left(Y_n^m(\widehat{\mathbf{x}})+Y_n^{-m}(\widehat{\mathbf{x}})\right)d\mathbf{x} \tag{B.11}$$

Interestingly, although Ambisonics is defined as the directional sensitivity at a point, it only has complete information when the integral is extended across a large (ultimately infinite) region. This relates to the impossible problem of determining direction of arrival of an arbitrary plane wave to infinite precision by measurement at a single point.

Ambisonics and modal analysis are closely related. The math of Ambisonics is simpler in nature by assuming a infinitesimal measurement volume with infinite precision. Although this is not a practical possibility, it eliminates the frequency and radial dependence from the Ambisonic sound field representation. With a complete radial dependence through the spherical Bessel functions, modal analysis is a more complete representation and it incorporates the radial and frequency dependence of the sound field. Thus, using the modal soundfield synthesis and reconstruction equations will yield more accurate results where the reconstruction is over a finite volume and the reconstruction transducers are at a finite distance. Ambisonics assumes an infinitesimal reconstruction volume and speakers at an infinite distance.

Bibliography

[1] R. H. Gilkey and J. M. Weisenberger, "The sense of presence for the suddenly-deafened adult: Implications for virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 4, no. 4, pp. 357–363, 1995.

[2] R. H. Gilkey, B. D. Simpson, S. K. Isabelle, A. J. Kordik, and J. M. Weisenberger, "Audition and the sense of presence in virtual environments," *Journal of the Acoustical Society of America*, vol. 105, no. 2, pp. 1163–1164, February 1999, Conference Abstract.

[3] M. M. Boone, E. N. G. Verheijen, and P. F. Van Tol, "Spatial sound-field reproduction by wave-field synthesis," *Journal of the Audio Engineering Society*, vol. 43, no. 12, pp. 1003–1012, December 1995.

[4] R. M. Held and N. I. Durlach, "Telepresence," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 1, pp. 102–112, 1992.

[5] J. V. Draper, D. B. Kaber, and J. M. Usher, "Telepresence," *Human Factors*, vol. 40, no. 3, pp. 354–375, September 1998.

[6] "Telepresence," *BT Technology Journal*, vol. 15, no. 4, October 1997, Special Issue.

[7] M. P. Hollier, A. N. Rimell, and D. Burraston, "Spatial audio technology for telepresence," *BT Technology Journal*, vol. 15, no. 4, pp. 33 – 41, October 1997.

[8] W. Krebber, H. Gierlich, and K. Genuit, "Auditory virtual environments: Basics and applications for interactive simulations," *Signal Processing*, vol. 80, pp. 2307–2322, 2000.

[9] W. Gaver and R. Smith, "Auditory icons in large-scale collaborative environments," in *Human-Computer Interaction - INTERACT '90*, D. Diaper et Al., Ed. 1990, pp. 735–740, Elsevier Science Publishers B.V. (North-Holland).

[10] E. Jovanov, K. Wegner, V. Radivojevic, D. Starcevic, M. S. Quinn, and D. B. Karron, "Tactical audio and acoustic rendering in biomedical applications," *IEEE Transactions on Information Technology*, vol. 3, no. 2, pp. 109–118, June 1999.

[11] G. Eckel, "Immersive audio-augmented environments - the LISTEN project," in *Proceedings of the 5th International Conference on Information Visualization (IV2001)*, B. Banissi, F. Khosrowshahi, M. Sarfraz, and A. Ursyn, Eds., Los Alamitos, CA, USA, 2001, IEEE Computer Society Press.

[12] T. Holman, *Sound for Film and Television*, Focal Press, Boston, 1997.

[13] C. Kyriakakis, “Virtual loudspeakers and virtual microphones for multichannel audio,” in *IEEE International Conference on Consumer Electronics*, Los Angeles, June 2000, pp. 404–405, IEEE.

[14] M. A. Gerzon, “Practical periphony: The reproduction of full sphere sound,” in *AES 65th Convention*, London, 1980, Audio Engineering Society, Preprint 1571.

[15] D. R. Begault, “Auditory and non-auditory factors that potentially influence virtual acoustic imagery,” in *AES 16th International Conference on Spatial Sound Reproduction*. 1999, Audio Engineering Society.

[16] G. Dickins, D. McGrath, A. McKeag, A. Reilly, L. Layton, B. Conolly, R. Cartright, R. Buttler, and S. Bartlett Et. Al., “Binaural simulation studies and experiments,” Tech. Rep., Lake Technology, Sydney, 1996-2000, Internal Discussion Papers - Unpublished.

[17] F. L. Wightman and D. J. Kistler, “Headphone simulation of free-field listening. II: Psychophysical validation,” *Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 686–878, February 1989.

[18] P. Zahorik, “Assessing auditory distance perception using virtual acoustics,” *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1832–1846, April 2002.

[19] E. Wenzel, “Issues in the development of virtual acoustic environments,” *Journal of the Acoustical Society of America*, vol. 92, no. 4, pp. 2332, October 1991, Conference Abstract.

[20] J. Meyer and G. W. Elko, “A spherical microphone array for spatial sound recording,” *Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2346, May 2002.

[21] M. M. Boone, “Acoustic rendering with wave field synthesis,” in *ACM SIGGraph and EUROGraphics Campfire: Acoustic Rendering for Virtual Environments*, Snowbird, Utah, May 2001, ACM.

[22] D. B. Ward, “On the performance of acoustic crosstalk cancellation in a reverberant environment,” *Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 1195–1197, August 2001.

[23] D. B. Ward and T. D. Abhayapala, “Reproduction of a plane-wave sound field using an array of loudspeakers,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 607–707, September 2001.

[24] A. Kulkarni and H. S. Colburn, “Variability in the characterization of the headphone transfer function,” *Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 1071–1074, February 2000.

[25] A. Mouchtaris, P. Reveliotis, and C. Kyriakakis, “Inverse filter design for immersive audio rendering over loudspeakers,” *IEEE Transactions on Multimedia*, vol. 2, no. 2, pp. 77–87, June 2000.

[26] M. Vorlander, “Acoustic load on the ear caused by headphones,” *Journal of the Acoustical Society of America*, vol. 107, no. 4, pp. 2082–2088, April 2000.

[27] S. E. Voss, J. J. Rosowski, C. A. Shera, and W. T. Peake, “Acoustic mechanisms that determine the ear-canal sound pressures generated by earphones,” *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1548–1565, March 2000.

[28] D. B. Ward and G. W. Elko, “A new robust system for 3d audio using loudspeakers,” in *ICASSP 2000*, Istanbul, 2000, pp. 781–784, IEEE.

[29] D. de Vries and M. M. Boone, “Wave field synthesis and analysis using array technology,” in *Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, October 1999, pp. 15–18, IEEE.

[30] G. Dickins and R. Kennedy, “Towards optimal sound field representation,” in *106th Convention of the Audio Engineering Society*, Munich, May 1999, Preprint 4925.

[31] E. G. Williams, “Reconstruction and projection of interior sound fields using a spherical measurement array,” in *ASA / ICA Congress, Seattle*, 1998.

[32] N. Chateau and A. W. Bronkhorst, “Efficient representation of head-related transfer functions,” *Journal of the Acoustical Society of America*, vol. 101, no. 5, pp. 3084, November 1997, Conference Abstract.

[33] G. Dickins, “Automated time alignment and equalization of a speaker array for sound-field reproduction,” in *6th Australian Regional Convention*. Audio Engineering Society, September 1996, Audio Engineering Society, Preprint 4317.

[34] D. Pralong and S. Carlile, “The role of individualized headphone calibration for the generation of high fidelity virtual auditory space,” *Journal of the Acoustical Society of America*, vol. 100, no. 6, pp. 3785–3793, December 1996.

[35] D. Stanzial, N. Prodi, and G. Schiffrer, “Reactive acoustic intensity for general fields and energy polarization,” *Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 1868–1876, April 1996.

[36] H. Moller, Hammershoi D, C. B. Jensen, and M. F. Sorensen, “Transfer characteristics of headphones measured on human ears,” *Journal of the Audio Engineering Society*, vol. 43, pp. 203–217, 1995.

[37] P. A. Nelson, F. Orduna-Bustamante, and H. Hamada, “Inverse filter design and equalization zones in MultiChannel sound reproduction,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 3, pp. 185–192, May 1995.

[38] H. L. Han, “Measuring a dummy head in search of pinnae cues,” *Journal of the Audio Engineering Society*, vol. 42, pp. 15–37, 1994.

[39] P. A. Nelson, “Active control of acoustic fields and the reproduction of sound,” *Journal of Sound and Vibration*, vol. 177, no. 4, pp. 447–477, 1994.

[40] A. J. Berkhout, D. de Vries, and P. Vogel, “Acoustic control by wave field synthesis,” *Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2780, May 1993.

[41] M. M. Boone and E. N. G. Verheijen, “Multi-channel sound reproduction based on wave field synthesis,” 1993, Audio Engineering Society, Preprint 3719.

[42] O. Kirkeby and P. A. Nelson, “Reproduction of plane wave sound fields,” *Journal of the Acoustical Society of America*, vol. 94, pp. 2992–3000, 1993.

[43] D. J. Kistler and F. L. Wightman, “A model of head-related transfer functions based on principle component analysis and minimum phase reconstruction,” *Journal of the Acoustical Sociery of America*, vol. 91, pp. 1637–1647, 1992.

[44] P. A. Nelson and S. J. Elliot, *Active Control of Sound*, Academic Press, New York, 1992.

[45] A. J. Berkhout, “A holographic approach to acoustic control,” *Journal of the Audio Engineering Society*, vol. 36, no. 12, pp. 977–995, 1988.

[46] P. M. Morse and K. Ingard, *Theoretical Acoustics*, McGraw Hill, New York, 1968.

[47] E. Hulsebos, D. DeVries, and E. Bourdillat, “Improved microphone array configurations for auralization of sound fields by wave-field synthesis,” *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 779–790, October 2000.

[48] J. Lim and C. Kyriakakis, “Virtual loudspeaker rendering for multiple listeners,” in *109th Convention, Los Angeles*. 2000, Audio Engineering Society, Preprint 5183.

[49] J. J. Lopez, F. Orduna, and A. Gonzalez, “Modelling and measurement of cross-talk cancellation zones for small displacements of the listener in transaural sound reproduction,” in *109th Convention, Los Angeles*. 2000, Audio Engineering Society, Preprint 5267.

[50] M. A. Poletti, “A unified theory of horizontal holographic sound systems,” *Journal of the Audio Engineering Society*, vol. 48, no. 12, pp. 1155–1182, December 2000.

[51] G. Dickins, P. Flanagan, and L. Layton, “Real-time virtual acoustics for 5.1,” in *Proceedings of the AES 16th International Conference*, Ravoniemi, Finland, April 1999, pp. 136–140, Audio Engineering Society.

[52] G. Dickins, M. Flax, A. McKeag, and D. McGrath, “Optimal 3d speaker panning,” in *The Proceedings of the AES 16th International Conference*, Ravoniemi, Finland, April 1999, pp. 421–426, Audio Engineering Society.

[53] G. Dickins, P. Flanagan, and L. Layton, “The SP1 - acoustic simulation for post production,” in *Proceedings of the 106th Convenrion of the Audio Engineering Society*, Munich, May 1999, Audio Engineering Society.

[54] U. Horbach and M. M. Boone, “Future transmission and rendering formats for multichannel sound,” in *AES 16th International Conference on Spatial Sound Reproduction*. 1999, Audio Engineering Society.

[55] J. Jot, V. Larcher, and J. Pernaux, "A comparative study of 3-D audio encoding and rendering techniques," in *AES 16th International Conference on Spatial Sound Reproduction*. 1999, Audio Engineering Society.

[56] Y. Kahana, P. A. Nelson, O. Kirkeby, and H. Hamada, "A multiple microphone recording technique for the generation of virtual acoustic images," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1503–1516, March 1999.

[57] C. Kyriakakis, P. Tsakalides, and T. Holman, "Surrounded by sound: Acquisition and rendering methods for immersive audio," *IEEE Signal Processing Magazine*, pp. 55–66, January 1999.

[58] R. Nicol and M. Emerit, "3d-sound reproduction over and extensive listening area: A hybrid method derived from holophony and ambisonics," in *AES 16th International Conference on Spatial Sound Reproduction*. 1999, Audio Engineering Society.

[59] D. B. Ward and G. W. Elko, "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," *IEEE Signal Processing Letters*, vol. 6, no. 5, pp. 106–108, May 1999.

[60] J. Daniel and J. Rault, "Ambisonics encoding of other audio formats for multiple listening conditions," in *105th Convention*. 1998, Audio Engineering Society, Preprint 4795.

[61] C. Kyriakakis and T. Holman, "Immersive audio for desktop systems," *Journal of the Acoustical Society of America*, vol. 103, pp. Volume 6, Page 3753, May 1998.

[62] C. Kyriakakis, "Fundamental and technological limitations of immersive audio systems," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 941–951, May 1998.

[63] D. B. Ward and G. W. Elko, "Optimum loudspeaker spacing for robust crosstalk cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-98)*, Seattle, WA, May 1998, pp. 3541–3544, IEEE Press, Vol 6.

[64] J. Mackenzie, J. Huopaniemi, V. Valimaki, and I. Kale, "Low-order modelling of head related transfer functions using balanced model truncation," *IEEE Signal Processing Letters*, vol. 4, no. 2, pp. 39–41, February 1997.

[65] J. Bamford, "An analysis of ambisonic sound systems of first and second order," M.S. thesis, Audio Laboratory, University of Waterloo, 1996.

[66] M. Poletti, "The design of encoding functions for stereophonic and polyphonic sound systems," *Journal of the Audio Engineering Society*, vol. 44, no. 11, pp. 948–963, November 1996.

[67] C. J. MacCabe and D. J. Furlong, "Virtual imaging capabilities of surround sound systems," *Journal of the Audio Engineering Society*, vol. 42, no. 1/2, pp. 38–49, Jan/Feb 1994.

[68] M. A. Gerzon, "General metatheory of auditory localization," in *AES 92nd Convention*, Vienna, 1992, Audio Engineering Society, Preprint 3306.

[69] M. A. Gerzon, “Optimal matrices for multispeaker stereo,” *Journal of the Audio Engineering Society*, vol. 40, no. 7/8, pp. 571–589, July/August 1992.

[70] D. H. Cooper, “Comments on distinction between stereophonic and binaural sound,” *Journal of the Audio Engineering Society*, vol. 39, pp. 261–266, 1991.

[71] D. H. Cooper and J. L. Bauck, “Prospects for transaural recording,” *Journal of the Audio Engineering Society*, vol. 37, pp. 3–19, 1989.

[72] D. J. Furlong, “Compariative study of effective soundfield reconstruction,” in *87th Convention of the Audio Engineering Society*, New York, October 1989, Preprint 2842.

[73] T. Gotoh, “Can the acoustic head related transfer function explain every phenomenon in sound localization,” in *Localization of Sound: Theory and Applications*, R. W. Gatehouse, Ed., Connecticut, 1982, pp. 244–249, Amphora Press.

[74] K. Farrar, “Soundfield microphone,” *Wireless World*, pp. 48–103, October 1979.

[75] J. H. Smith, “The sound field microphone,” *db*, pp. 34–37, July 1978.

[76] M. A. Gerzon, “The optimum choice of surround sound encoding specification,” in *AES 56th Convention*, Paris, 1977, Audio Engineering Society, Preprint 1199.

[77] M. A. Gerzon, “Surround sound psychoacoustics,” *Wireless World*, vol. 80, no. 12, pp. 483–486, December 1974.

[78] M. A. Gerzon, “Periphony: With height sound reproduction,” *Journal of the Audio Engineering Society*, vol. 21, pp. 2–10, 1973.

[79] G. C. Stecker and E. R. Hafter, “Localization judgments are sensitive to late-arriving sound,” *Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2355, May 2002.

[80] A. S. Bregman, *Auditory Scene Analysis : The Perceptual Orginization of Sound*, The MIT Press, Massachusetts, 1999.

[81] B. C. J. Moore, “Controversies and mysteries in spatial hearing,” in *AES 16th International Conference on Spatial Sound Reproduction*. 1999, Audio Engineering Society.

[82] D. R. Moore and A. J. King, “Auditory perception: The near and far of sound localization,” *Current Biology*, vol. 9, no. 10, pp. 361–363, May 1999.

[83] R. M. Warren, *Auditory Perception: A New Analysis and Synthesis*, Cambridge, Cambridge UK, 1999.

[84] E. Zwicker and H. Fastl, *Psychoacoustics : Facts and Models*, Springer, Germany, 1999.

[85] R. K. Clifton and R. L. Freyman, “The precedence effect: Beyond echo suppression,” in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds., New Jersey, 1997, pp. 233–256, Lawrence Erlbaum.

[86] E. R. Hafter, “Binaural adaptation and the effectiveness of a stimulus beyond its onset,” in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H. Gilkey and T. R. Anderson, Eds., New Jersey, 1997, pp. 211–232, Lawrence Erlbaum.

[87] W. M. Hartmann, “Listening in a room and the precedence effect,” in *Binaural and Spatial Hearing in Real and Virtual Auditory Environments*, R. H. Gilkey and T. R. Anderson, Eds., Mahwah, New Jersey, 1997, pp. 191–210, Lawrence Erlbaum Associates.

[88] J. Lewald, “Eye-position effects in directional hearing,” *Behavioural Brain Research*, vol. 87, pp. 35–48, 1997.

[89] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, London, 1997 (4th Ed).

[90] R. O. Duda, “Auditory localization demonstrations,” *Acustica*, vol. 82, pp. 346–357, 1996.

[91] D. M. Howard and J. Angus, *Acoustics and Psychoacoustics*, Focal Press, Oxford, 1996.

[92] C. J. Darwin and R. P. Carlyon, “Auditory grouping,” in *Hearing : Handbook of Perception and Cognition - 2nd Edition*, B. C. J. Moore, Ed., pp. 387–420. Academic Press, San Diego, 1995.

[93] D. W. Grantham, “Spatial hearing and related phenomena,” in *Hearing : Handbook of Perception and Cognition - 2nd Edition*, B. C. J. Moore, Ed., pp. 297–339. Academic Press, San Diego, 1995.

[94] S. Handel, “Timbre perception and auditory object identification,” in *Hearing : Handbook of Perception and Cognition - 2nd Edition*, B. C. J. Moore, Ed., pp. 425–462. Academic Press, San Diego, 1995.

[95] R. K. Clifton, R. L. Freyman, R. Y. Litovsky, and D. McCall, “Listener’s expectations about echoes can raise or lower echo threshold,” *Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1525–1533, March 1994.

[96] A. S. Bregman, “Auditory scene analysis: Hearing in complex environments,” in *Thinking in Sound : The Cognitive Psychology of Human Audition*, S. McAdams and E. Bigand, Eds., chapter 2, pp. 10–36. Oxford University Press, Oxford, 1993.

[97] R. L. Freyman, R. K. Clifton, and R. Y. Litovsky, “Dynamic processes in the precedence effect,” *Journal of the Acoustical Society of America*, vol. 90, no. 2, pp. 874–883, August 1991.

[98] S. A. Gelfand, *Hearing: An Intorduction to Psychological and Physiological Acoustics*, chapter 13 - Binaural Hearing, pp. 419–455, Marcel Dekker Inc., New York, 1990.

[99] S. Handel, *Listening: An Introduction to the Perception of Auditory Events*, The MIT Press, Massachusetts, 1989.

[100] S. Kuwada and T. C. T. Yin, “Physiological studies of directional hearing,” in *Directional Hearing*, W. A. Yost and G. Gourevitch, Eds., chapter 6, pp. 146–167. Springer-Verlag, New York, 1987.

[101] F. L. Wightman, D. J Kistler, and M. E. Perkins, “A new approach to the study of human sound localization,” in *Directional Hearing*, W. A. Yost and G. Gourevitch, Eds., chapter 2, pp. 26–48. Springer-Verlag, New York, 1987.

[102] B. C. J. Moore, *Introduction to the Psychology of Hearing*, Macmillan Press Ltd, London, 1977 (1st Ed).

[103] G. Plenge, “On the differences between localization and lateralization,” *Journal of the Acoustical Society of America*, vol. 56, no. 3, pp. 944–951, September 1974.

[104] R. Clifton, R. Freyman, and J. Meo, “What the precedence effect tells us about room acoustics,” *Perception and Psychophysics*, vol. 64, no. 2, pp. 180–188, 2002.

[105] S. Carlile, S. Hyams, and S. Delaney, “Systematic distortions of auditory space perception following prolonged exposure to broadband noise,” *Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 416–424, July 2001.

[106] D. L. Guettler, R. S. Boila, and W. T. Nelson, “Monitoring and localizing simultaneous real-world sounds: Implications for the design of spatial audio displays,” *Journal of the Acoustical Society of America*, vol. 108, no. 5, pp. 2573, November 2000, Conference Abstract.

[107] J. Lewald and W. H. Ehrenstein, “Auditory-visual spatial integration: A new psychophysical approach using laser pointing to acoustical targets,” *Journal of the Acoustical Society of America*, vol. 104, no. 3, pp. 1586–1597, September 1998.

[108] M. Kashino, “Auditory after-effects revealing adaptive representation of auditory space,” *Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2680, October 1996, Conference Abstract.

[109] W. M. Hartmann, “Location of sound in rooms,” *Journal of the Acoustical Society of America*, vol. 74, no. 5, pp. 1380–1391, November 1983.

[110] P. Bertelson and M. Radeau, “Cross-modal bias and perceptual fusion with auditory-visual spatial discordance,” *Perception and Psychophysics*, vol. 29, no. 6, pp. 578–584, 1981.

[111] D. H. Warren, R. B. Welch, and T. J. McCarthy, “The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses,” *Perception and Psychophysics*, vol. 30, no. 6, pp. 557–564, 1981.

[112] R. B. Welch and D. H. Warren, “Immediate perceptual response to intersensory discrepancy,” *Psychological Bulletin*, vol. 88, no. 3, pp. 638–667, 1980.

[113] L. E. Marks, *The Unity of the Senses: Interrelations Among the Modalities*, Academic Press, New York, 1978.

[114] M. Radeau and P. Bertelson, “Adaption to auditory-visual discordance and ventriloquism in semi-realistic situations,” *Perception and Psychophysics*, vol. 22, no. 2, pp. 137–146, 1977.

[115] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, pp. 746–748, December 1976.

[116] M. I. Posner, M. J. Nissen, and R. M. Klein, “Visual dominance: An information-processing account of its origins and significance,” *Psychological Review*, vol. 83, no. 2, pp. 157–171, 1976.

[117] M. Radeau and P. Bertelson, “The effect of a textured visual field on modality dominance in a ventriloquist situation,” *Perception and Psychophysics*, vol. 20, no. 4, pp. 227–235, 1976.

[118] M. Cohen, “Changes in auditory localization following prismatic exposure under continuous and terminal visual feedback,” *Perceptual and Motor Skills*, vol. 38, pp. 1202, 1974.

[119] C. E. Jack and W. R. Thurlow, “Effects of degree of visual assoaciation and angle of displacement on the "ventriloquism" effect,” *Perceptual and Motor Skills*, vol. 37, pp. 967–979, 1973.

[120] B. Julesz and I. J. Hirsh, “Visual and auditory perception - an essay of comparison,” in *Human Communication: A Unified View*, E. E. David and P. B. Denes, Eds., chapter Eight, pp. 283–341. Mc-Graw-Hill, New York, 1972.

[121] M. B. Gardner, “Distance estimation of 0degree or apparent 0degree oriented speech signals in anechoic space,” *Journal of the Acoustical Society of America*, vol. 45, no. 1, pp. 47–53, 1969.

[122] M. B. Gardner, “Image fusion, broadening and displacement in sound localization,” *Journal of the Acoustical Society of America*, vol. 46, no. 2, pp. 339–349, 1969.

[123] Herbert L. Pick, David H. Warren, and John C. Hay, “Sensory conflict in judgements of spatial direction,” *Perception and Psychophysics*, vol. 6, no. 4, pp. 203–205, 1969.

[124] M. B. Gardner, “Proximity image effect in sound localization,” *Journal of the Acoustical Society of America*, vol. 43, pp. 163, 1968.

[125] D. Lubman, “Fluctuations of sound with position in a reverberant room,” *Journal of the Acoustical Society of America*, vol. 44, no. 6, pp. 1491–1502, 1968.

[126] P. T. Young, “Auditory localization with acoustical transposition of the ears,” *Journal of Experimental Psychology*, vol. 11, no. 6, pp. 399–429, December 1928.

[127] A. J. King, J. W. H. Schnupp, and T. P. Doubell, “The shape of ears to come: Dynamic coding of auditory space,” *Trends in Cognitive Sciences*, vol. 5, no. 6, pp. 261–269, June 2001.

[128] T. Streeter, B. Shinn-Cunningham, and A. Brughera, "Short-term adaption to novel combinations of acoustical spatial cues," *Journal of the Acoustical Sociery of America*, vol. 109, no. 5, pp. 2376, May 2001.

[129] D. Trapenskas and O. Johansson, "Localization performance of binaurally recorded sounds with and without training," *International Journal of Industrial Ergonomics*, vol. 27, pp. 405–410, 2001.

[130] P. Zahorik, "Effects of visual-feedback training in 3-D sound displays," *Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2487, May 2001, Conference Abstract.

[131] R. Duraiswami, L. Davis, S. A. Shamma, H. C. Elman, R. O. Duda, V. R. Algazi, Q. Liu, and S. T. Raveendra, "Individualized HRTFs using computer vision and computational acoustics," *Journal of the Acoustical Society of America*, vol. 108, no. 5, pp. 2597, November 2000, Conference Abstract.

[132] J. D. Johnston and Y. H. Lam, "Perceptual soundfield reconstruction," in *109th Convention, Los Angeles*. 2000, Audio Engineering Society, Preprint 5202.

[133] E. H. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 528–537, January 2000.

[134] B. Shinn-Cunningham, S. Santarelli, and N. Kopco, "Tori of confusion: Binaural localization cues for sources within reach of a listener," *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1627–1636, March 2000.

[135] B. Shinn-Cunningham, "Adapting to remapped auditory localization cues: A decision theory model," *Perception and Psychophysics*, vol. 61, no. 2, pp. 33–47, 2000.

[136] P. Zahorik, "Distance localization using non-individualized head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 108, no. 5, pp. 2597, November 2000, Conference Abstract.

[137] E. M. Wenzel, "Effect of increasing system latency on localization of virtual sounds," in *AES 16th International Conference on Spatial Sound Reproduction*. 1999, Audio Engineering Society.

[138] E. M. Wenzel and D. R. Begault, "Are individualized head-related transfer functions required for auditory information displays," *Journal of the Acoustical Society of America*, vol. 105, no. 2, pp. 1035, February 1999, Conference Abstract.

[139] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2841–2853, 1999.

[140] H. Heuermann and H. Colonius, "Localization experiments with saccadic responses in virtual auditory environments," in *Psychophysics, Physiology and Models of Hearing*, T. Dau, V. Hohmann, and B. Killmeier, Eds., pp. 89–92. World Scientific, Singapore, 1998.

[141] P. M. Hofman, J. G. A. Van Riswick, and A. J. Van Opstal, “Relearning sound localization with new ears,” *Nature Neuroscience*, vol. 1, no. 5, pp. 417–421, September 1998.

[142] J. Jot, S. Wardle, and V. Larcher, “Approaches to binaural synthesis,” in *105th Convention*. 1998, Audio Engineering Society, Preprint 4861.

[143] F. Wightman and D. Kistler, “Of vulcan ears, human ears and earprints,” *Nature Neuroscience*, vol. 1, no. 5, pp. 337–339, September 1998.

[144] B. Shinn-Cunningham, H. Lehnert, G. Kramer, E. Wenzel, and N. Durlach, “Auditory displays,” in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. H Gilkey and T. R. Anderson, Eds., New Jersey, 1997, pp. 611–663, Lawrence Erlbaum.

[145] P. A. Nelson, F. Orduna-Bustamante, and D. Engler, “Experiments on a system for the synthesis of virtual acoustic sources,” *Journal of the Audio Engineering Society*, vol. 44, no. 11, pp. 990–1006, November 1996.

[146] B. D. Simpson, D. W. Hale, S. K. Isabelle, and Robert H. Gilkey, “The experiences of untrained subjects listening to virtual sounds,” *The Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2633, October 1996, Conference Abstract.

[147] P. Zahorik, F. L. Wightman, and D. J. Kistler, “The fidelity of virtual auditory displays,” *The Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 2596, April 1996, Conference Abstract.

[148] F. Wightman, D. Kistler, and M. Arruda, “Perceptual consequences of engineering compromises in synthesis of virtual auditory objects,” *The Journal of the Acoustical Society of America*, vol. 92, no. 4, pp. 2332, October 1992, Conference Abstract.

[149] F. L. Wightman and D. J. Kistler, “Headphone simulation of free-field listening I: Stimulus synthesis,” *Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 858–867, February 1989.

[150] P. Chueng, “Minimal ecological sound design for a sense of presence in digital virtual environments,” in *Proceedings of the Sixth Human Centred Technology Postgraduate Workshop*, UK, 2002, University of Sussex.

[151] P. Chueng, “Designing sound canvas: The role of expectation and discrimination,” in *Extended Abstracts of CHI 2002 Conference on Human Factors in Computing Systems*. 2002, ACM - SIGCHI.

[152] A. Dufour, O. Despres, and T. Pebayle, “Visual and auditory facilitation in auditory spatial localization,” *Visual Cognition*, vol. 9, no. 6, pp. 741–753, 2002.

[153] L. Girin, J. Schwartz, and G. Feng, “Audio-visual enhancement of speech in noise,” *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, June 2001.

[154] B. Shinn-Cunningham, “Learning reverberation: Considerations for spatial auditory displays,” in *ICAD 2000*, 2000, pp. 126–134.

[155] A. J. King, “Auditory perception: Does practice make perfect,” *Current Biology*, vol. 9, no. 4, pp. 143–146, February 1999.

[156] A. J King, “Sensory experience and the formation of a computational map of auditory space in the brain,” *BioEssays*, vol. 21, no. 11, pp. 900–911, 1999.

[157] P. P. Lennox, T. Myatt, and J. M. Vaughan, “From surround to true 3-D,” in *AES 16th International Conference on Spatial Sound Reproduction*. 1999, Audio Engineering Society.

[158] G. Dickins, “What is surround sound,” *Acoustics Australia*, vol. 26, no. 3, December 1998.

[159] J. Hull, *Surround Sound: Past Present and Future*, Dolby Laboratories, San Fransisco, 1997.

[160] R. Dressler, *Dolby Prologic: Principles of Operation*, Dolby Laboratories Inc, San Fransisco, 1997.

[161] R. Elen, “Whatever happened to ambisonics,” *Audio Media Magazine*, , no. 11, November 1991.

[162] M. A. Gerzon, “Compatibility of and conversion between multispeaker systems,” in *AES 93rd Convention*, San Fransisco, 1992, Audio Engineering Society, Preprint 3405.

[163] M. A. Gerzon, “Hierachical transmission system for multispeaker stereo,” *Journal of the Audio Engineering Society*, vol. 40, no. 9, pp. 692, September 1992.

[164] M. F. Davis, “The AC-3 multichannel coder,” *Jounal of the Audio Engineering Society*, vol. 95th Convention, 1993.

[165] G. Brockhouse, “2 will get you,” *Stereo Review*, pp. 59–63, August 1998.

[166] R. Pellegrini, “Quality assessment of auditory virtual environments,” in *Proceedings of the 2001 International Conference on Auditory Displays*, Espoo, Finland, August 2001, pp. 161–168, ICAD.

[167] G. Dickins, D. McGrath, A. McKeag, A. Reilly, and L. Layton, “The fat controller - 5 channel headphone virtualizer demonstration,” Tech. Rep., Lake Technology, Sydney, 1998, Exhibited AES 105th San Fransisco.

[168] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1998.

[169] D. B. Ward, R. A. Kennedy, and R. C. Williamson, “On the theory and design of broadband arrays,” *Journal of the Acoustical Society of America*, vol. 97, no. 2, pp. 1023–1034, February 1995.

[170] D. B. Ward, Z. Ding, and R. A. Kennedy, “Broadband DOA estimation using frequency invariant beamforming,” *IEEE Transactions on Signal Processing*, vol. 46, no. 5, pp. 1463–1469, May 1998.

[171] R. A. Kennedy, T. D. Abhayapala, and D. B. Ward, "Broadband nearfield beamforming using a radial beampattern transformation," *IEEE Transactions on Signal Processing*, vol. 46, no. 8, pp. 2147–2156, August 1998.

[172] R. A. Kennedy, T. D. Abhayapala, and T. S. Pollock, "Modeling Multipath Scattering Environments Using Generalized Herglotz Wave Functions," in *Proc. 4th Australian Communications Theory Workshop*, Melbourne, Australia, February 2003, pp. 87–92.

[173] G. R. Baldcock and T. Bridgeman, *The Mathematical Theory of Wave Motion*, Ellis Horwood Ltd, Chichester, England, 1981.

[174] A. J. Berkhout, *Applied Seismic Wave Theory*, Advanced Seismic Theory. Elsevier, Amsterdam, 1987.

[175] A. J. Berkhout, D. de Vries, and J. J. Sonke, "Array technology for acoustic wave field analysis in enclosures," *Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2757–2770, November 1997.

[176] R. Kennedy, T. Abhayapala, and H. Jones, "Bounds on the spatial richness and dimension of multipath," *Submission to IEEE Transactions on Signal Processing*, 2002.

[177] R. A. Kennedy, T. D. Abhayapala, and H. M. Jones, "Bounds on the Spatial Richness of Multipath," in *Proc. 3rd Australian Communications Theory Workshop*, Canberra, Australia, February 2002, pp. 76–80.

[178] N. J. A. Sloane R. H. Hardin and W. D. Smith, "Spherical codes," Published electronically at www.research.att.com/ njas/packings/, 2000, Book in preparation.

[179] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, pp. 145–152, 1988.

[180] M. R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *Journal of the Audio Engineering Society*, vol. 35, pp. 299–305, 1987.

[181] Z. Meng, K. Sakagami, M. M., and Guoan Bi, "Predictability of a room impulse response," in *109th Convention, Los Angeles*. 2000, Audio Engineering Society, Preprint 5233.

[182] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: Robustness results," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 311–319, May 2000.

[183] J. Blauert, *The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, Massachusetts, 1999.

[184] Lord Rayleigh, "On our perception of sound direction," *Philosophical Magazine*, vol. 13, pp. 214–232, 1907.

[185] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, Academic Press, 1999.

[186] P. A. Nelson, F. Orduna-Bustamante, and H. Hamada, “Multichannel signal processing techniques in the reproduction of sound,” *Journal of the Audio Engineering Society*, vol. 44, no. 11, pp. 973–989, November 1996.

[187] E. M. Wenzel, M. Arruda, D. J Kistler, and F. L. Wightman, “Localization using nonindividualized head-related transfer functions,” *Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, July 1993.

[188] R. Drullman and A. W. Bronkhorst, “Multichannel speech intelligibility and talker recognition using monaural, binaural and three dimensional auditory presentation,” *Journal of the Acoustical Society of America*, vol. 107, no. 4, pp. 2224–2235, April 2000.

[189] S. R. Oldfield and S. P. A. Parker, “Acuity of sound localization: A topography of auditory space II. pinna cues absent,” *Perception*, vol. 13, pp. 601–617, 1984.

[190] Lake Technology, “Huron manual - MultiScape and AniScape,” Tech. Rep., Lake Technology, Sydney, Australia, 1997.

[191] B. Scharf and A. J. M. Houtsma, “Loudness, pitch, localization, aural distortion, pathology,” in *Handbook of Perception and Human Performance*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., chapter 15. John Wiley and Sons, New York, 1986.

[192] J. M. Harrison, “Behavioral specializations underlying the discrimination of sound position,” in *Localization of Sound: Theory and Applications*, R. W. Gatehouse, Ed., Connecticut, 1982, pp. 126–135, Amphora Press.

[193] R. W. Gatehouse, “Introduction,” in *Localization of Sound: Theory and Applications*, Connecticut, 1982, pp. 4–12, Amphora Press.

[194] K. B. Bennett, R. Parasuraman, and J. H. Howard, “Auditory induction of discrete tones in signal detection tasks,” *Perception and Psychophysics*, vol. 35, no. 6, pp. 570–578, 1984.

[195] G. A. Miller and J. C. R. Licklider, “The intelligibility of interrupted speech,” *Journal of the Acoustical Society of America*, vol. 22, pp. 167–173, 1950.

[196] J. A. Jr. Bashford and R. M. Warren, “Multiple phonemic restorations follow the rules fo auditory induction.,” *Perception and Psychophysics*, vol. 42, pp. 114–121, 1987.

[197] W. Chung, S. Carlile, and P. Leong, “A performance adequate computational model for auditory localization,” *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 432–445, January 2000.

[198] C. Jin, M. Schenkel, and S. Carlile, “Neural system identification model of human sound localization,” *Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1215–1235, September 2000.

[199] J. M. Groh, A. S. Trause, A. M. Underhill, K. R. Clark, and S. Inati, “Eye position influences auditory responses in primate inferior colliculus,” *Neuron*, vol. 29, pp. 509–518, 2001.

[200] J. Lewald and W. H Ehrenstein, “Effect of gaze direction on sound localization in rear space,” *Neuroscience Research*, vol. 39, pp. 253–257, 2001.

[201] E. E. David and P. B. Denes, *Human Communication: A Unified View*, pp. xiii–xvii, McGraw-Hill, New York, 1972.

[202] H. Hawkins and J. Presson, “Auditory information processing,” in *Handbook of Perception and Human Performance*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., chapter 26. John Wiley and Sons, New York, 1986.

[203] B. G. Shinn-Cunningham, “Models of plasticity in spatial auditory processing,” *Audiology and Neuro-otology*, vol. 6, no. 4, pp. 187–191, 2001.

[204] E. R. Hafter, T. N. Buell, and V. N Richards, “Onset-coding in lateralization: Its form, site and function,” in *Auditory Function: Neurobiological Bases of Hearing*, W. E. Gall G. M. Eldman and M. W. Cowan, Eds. Wiley, New York, 1988.

[205] R. B. Welch, *Perceptual Modification: Adapting to Altered Sensory Environments*, Academic Press, London, 1978.

[206] R. Held, “Shifts in binaural localization after prolonged exposures to atypical combinations of stimuli,” *American Journal of Psychology*, vol. 68, pp. 526–548, 1955.

[207] B. Shinn-Cunningham, N. I. Durlach, and R. M. Held, “Adapting to supernormal auditory localization cues I: Bias and resolution,” *Journal of the Acoustical Society of America*, vol. 103, pp. 2656–2666, 1998.

[208] B. G. Shinn-Cunningham, N. I. Durlach, and R. M. Held, “Adapting to supernormal auditory localization cues II: Constraints on adaption of mean response,” *Journal of the Acoustical Sociery of America*, vol. 103, no. 6, pp. 3667–3676, June 1998.

[209] S. Komiyama, “Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems,” *Journal of the Audio Engineering Society*, vol. 37, no. 4, pp. 210–214, April 1989.

[210] D. S. Brungart and K. R. Scott, “The effects of production and presentation level on the auditory distance perception of speech,” *Journal of the Acoustical Society of America*, vol. 110, no. 1, pp. 425–440, July 2001.

[211] P. D. Coleman, “Failure to localize the source distance of an unfamiliar sound,” *Journal of the Acoustical Society of America*, vol. 34, no. 3, pp. 345–355, March 1962.

[212] D. W. Massaro, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[213] R. B. Welch and D. H. Warren, “Intersensory interactions,” in *Handbook of Perception and Human Performance*, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., chapter 25. John Wiley and Sons, New York, 1986.

[214] S. J. Freedman, L. Wilson, and J. H. Rekosh, “Compensation for auditory rearrangement in hand-ear co-ordination,” *Perceptual and Motor Skills*, vol. 24, pp. 1207–1210, 1967.

[215] M. P. Hollier, A. N. Rimmell, D. S. Hands, and R. M. Voelcker, "Multi-modal perception," *BT Technical Journal*, vol. 17, no. 1, pp. 35–46, January 1999.

[216] K. W. Grant and P. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1197–1208, September 2000.

[217] A. O'Leary and G. Rhodes, "Cross modal effects on visual and auditory object perception," *Perception and Psychophysics*, vol. 35, no. 6, pp. 565–569, 1984.

[218] H. A. Carr, *An Introduction to Space Perception*, Hafner Publishing, New York, 1935.

[219] Stratton, "Some preliminary experiments on vision without inversion of the retinal image," *Psychology Review*, vol. 3, pp. 611–617, 1896.

[220] R. R. Fay and A. N. Popper, "Evolution of hearing in vertebrates: The inner ears and processing," *Hearing Research*, vol. 149, pp. 1–10, 2000.

[221] William Hartmann, "How we localize sound," *Physics Today*, pp. 24–, November 1999.

[222] S. R. Oldfield and S. P. A. Parker, "Acuity of sound localization: A topography of auditory space I: Normal hearing conditions," *Perception*, vol. 13, pp. 581–600, 1984.

[223] H. Wallach, "The role of head movement and vestibular and visual cues in sound localization," *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 339–368, October 1940.

[224] C. Hendrix and W. Barfield, "The sense of presence within auditory virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 5, no. 3, pp. 290–301, 1996.